Discretization strategies for optimal control problems with parabolic partial differential equations

Dipl. Tech.-Math. Thomas Gerhard Flaig

Vollständiger Abdruck der an der Fakultät für Bauingenieurwesen und Umweltwissenschaften der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereichte Dissertation.

Vorsitzender:		Univ.-Prof. Dr.-Ing. Karl-Christian Thienel
Gutachter:	1.	Univ.-Prof. Dr. rer. nat. habil. Thomas Apel
	2.	Univ.-Prof. Dr. sc. nat. Fredi Tröltzsch (TU Berlin)
	3.	Univ.-Prof. Dr. rer. nat. habil. Boris Vexler (TU München)

Die Dissertation wurde am 23.01.2013 bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Bauingenieurwesen und Umweltwissenschaften am 20.04.2013 angenommen. Die mündliche Prüfung fand am 26.04.2013 statt.

Die vorliegende Dissertation ist auch im Verlag Dr. Hut München erschienen und kann online unter http://www.hut-verlag.de/9783843911023.html oder über den Buchhandel unter der ISBN 978-3-8439-1102-3 bestellt werden.

Abstract

The topic of this PhD thesis are estimates for the discretization error of optimal control problems with parabolic partial differential equations. Such problems occur e.g. during the hydration of young concrete. This problem is introduced in this thesis. Subsequently the discretization of optimal control problems with linear parabolic partial differential equations is discussed in detail.

A focus are tailored Crank-Nicolson schemes with convergence order two. The specialty of these tailored schemes is the commutation of optimization and discretization and the discretization of the state and the control at different discrete times. These discretizations are introduced for abstract parabolic optimal control problems with a cost functional which includes the tracking of a desired state over the full space-time cylinder, the tracking of the terminal state and the control costs.

Furthermore the finite element approximation of some semi-elliptic boundary value problems is discussed. Such problems are connected with parabolic optimal control problems and are of different order of differentiation in different dimensions. The corresponding bilinear form is V-elliptic in an appropriately chosen Hilbert space. For these boundary value problems a regularity estimate and a priori error estimates for the energy norm and L^2-norm are proven.

Numerical examples for the Crank-Nicolson discretization and the finite element approximation confirm the expected rates of convergence.

Zusammenfassung

Diese Dissertation behandelt die Abschätzung des Diskretisierungsfehlers bei Optimalsteuerungsproblemen mit parabolischen partiellen Differentialgleichungen. Solche Probleme treten beispielsweise bei der Hydratation von jungem Beton auf. Dieses Problem wird zu Beginn der Arbeit eingeführt. Anschließend wird die Diskretisierung von Optimalsteuerungsproblemen mit linearen parabolischen partiellen Differentialgleichungen im Detail diskutiert.

Ein Schwerpunkt der Arbeit liegt in angepassten Crank-Nicolson Verfahren, die die Konvergenzordnung zwei liefern. Die Verfahren sind so konstruiert, dass Optimierung und Diskretisierung kommutieren und der Zustand und die Steuerung an verschiedenen Zeitpunkten diskretisiert werden. Diese Diskretisierung wird für abstrakte parabolische Optimalsteuerungsprobleme mit einem Kostenfunktional eingeführt, das das Ansteuern eines gewünschten Zustandes über den ganzen Raum-Zeit-Zylinder, das Ansteuern eines gewünschten Endzustands und die Kontrollkosten beinhaltet.

Außerdem wird eine Finite Elemente Methode für gewisse semielliptische Randwertaufgaben eingeführt. Die betrachteten Probleme stehen im Zusammenhang mit Optimalsteuerungsproblemen parabolischer Differentialgleichungen und weisen verschiedene Differentationsordnungen in verschiedenen Raumdimensionen auf. Die zugehörigen Bilinearformen sind in einem geeignet gewählten Hilbertraum V-elliptisch. Für die Lösung dieser Randwertprobleme wird eine Regularitätsabschätzung hergeleitet und eine a priori Fehlerschranke in der Energienorm und der L^2-Norm bewiesen.

Numerische Beispiele für die Crank-Nicolson-Diskretisierungen und die Finite Element Methode bestätigen die erwarteten Konvergenzraten.

Acknowledgements

This PhD thesis was written during my employment at the Institut für Mathematik und Bauinformatik of the Universität der Bundeswehr München. During this time my work was partially supported by the DFG priority program 1253 “Optimization with partial differential equations”, which I acknowledge thankfully.

The research would not have been possible without the support of many people. At first I would like to express my gratitude to my supervisor Prof. Dr. Thomas Apel for the support, confidence, discussions and supervising this thesis about parabolic partial differential equations. Further I want to thank all the colleagues at the institute for the good atmosphere.

I would also like to mention many fruitful discussions about mathematical topics especially with (in alphabetical order) Olaf Benedix, Prof. Dr. Serge Nicaise, Prof. Dr. Arnd Rösch and Prof. Dr. Boris Vexler, as well as all the engineers who shared their favorite use of concrete and even allowed some questions about the basic properties of concrete, in particular the colleagues of the Universität der Bundeswehr München, Prof. Dr.-Ing. Manfred Keuser and Prof. Dr.-Ing. Karl-Christian Thienel. Finally I thank Bernhard Gehrmann for the long term collaboration on the simulation of the hydration of concrete.

Moreover I thank Prof. Dr. Fredi Tröltzsch and Prof. Dr. Boris Vexler for agreeing to be co-referees of this thesis and Prof. Dr.-Ing. Karl-Christian Thienel for beeing available as chairman of the commitee.

Last but not least I thank my family, especially my wife Ruth, for the constant support.

Neubiberg, January 2013 Thomas Flaig

Contents

1. Introduction

Contents

1.1. Motivation

As computer and computational power became available the first challenge was the solution of equations and large systems of equations. These equations were not solved for their own sake but as approximation of the solution of (partial) differential equations. So the next step was the development of advanced methods for the discretization and approximation of (partial) differential equations. In parallel the development of solution algorithms for systems of equations went on.

Now we are at the point that the numerical solution and approximation of partial differential equations and the corresponding systems of equations are well studied and we observe the next step in the requirements of computations. In technical applications the partial differential equations are not solved for their own sake but the solution should fulfill some requirements. So the application asks for an, in some sense, optimal solution of the problem. In engineering the simulation based optimization approach is widely used, i.e. the problem is solved with several input data and the best solution is chosen afterwards. We discuss the model based simulation, where we consider the problem in an appropriate function space and develop algorithms which yield (an approximation of) the optimal solution.

At the beginning of the work on this PhD thesis it was known by the work of Vexler and coworkers [11, 80, 81, 82, 83, 115] that the approaches optimize-then-discretize and discretize-then-optimize lead to the same discrete scheme for Galerkin time discretizations of optimal control problems with parabolic partial differential equations. At this time it was unknown to many researchers whether a time stepping scheme, which is not also a Galerkin scheme, with this property exists. By the author of this PhD thesis a Crank-Nicolson discretization, for which discretization and optimization commute, was developed. This Crank-Nicolson scheme gives a second order approximation due to a tailored approximation, in which the state and the control are discretized in different time points. As the Crank-Nicolson scheme coincides with the Störmer-Verlet discretization it was suddenly obvious that for some symplectic Runge-Kutta schemes discretization and optimization commute as seen in the case of ordinary differential equations by Bonnans and Laurent-Varin [16, 17] and Hager [54, 55].

In parallel to the author Meidner and Vexler developed a Petrov-Galerkin-Crank-Nicolson scheme, which also provides a second order approximation. In the preprint [83] Meidner and Vexler introduce their scheme for an optimal control problem with a finite dimensional control space and with control constraints, whereas the preprint [4] of Apel and Flaig with a continuous, infinite dimensional control space but no control constraints was available a little bit later. In the meantime both preprints have been published [5, 84] and the Crank-Nicolson discretization of [5] has been adopted to more general optimal control problems for this thesis.

This thesis is divided into several parts, which have strong interconnections. The starting point is the hydration of concrete and connected optimal control problems with parabolic partial differential equations. These optimal control problems motivate also the examination of optimal control problems with linear parabolic partial differential equations. As optimal control problems can be solved only in very special circumstances analytically, the use of discretization schemes is essential.

We now give a further short overview of the main results of the different parts of this thesis and relate them to the literature.

1.2. Hydration of concrete

In the following chapter the real world problem of cracks in young concrete is introduced. Concrete is produced from the mixing of cement, aggregate, admixtures and water. After the mixing an exothermic chemical reaction, known as hydration, begins and the rigid body properties of concrete develop. During this process cracks may occur and are, in most cases, not wanted. In Chapter 2 we give an overview of the thermo-mechanical properties of young concrete and formulate related optimal control problems.

An overview of the hydration of young concrete can be found e.g. in the works of Eierle [38], Krauß [71], Rostásy, Krauß and Gutsch [111] and a series of articles by Rostásy, Krauß and Budelmann [105, 106, 107, 108, 109, 110].

For model based optimization only few references for the hydration of concrete are known to the author. Sperber and Tröltzsch [120] discuss optimality conditions for optimal control problems with semilinear parabolic partial differential equations which were inspired by the hydration of concrete. Kalkowski [69] simplifies the optimal control of the hydration of thin structures of concrete to an optimal control problem of ordinary differential equations. Apel and Flaig [3] present the simulation of the hydration of concrete and introduce a family of optimal control problems for the hydration of young concrete with parabolic partial differential equations. Benedix [12] uses adaptive solution algorithms for some optimal control problems for the hydration of young concrete.

The problem of the hydration of concrete is evocative of the Stefan problem, which describes e.g. the phase transition of water and ice. But in contrast to Stefan problems and connected phase field problems where a discrete or smeared interface exists, there is no macroscopic interface in the hydration of concrete as the chemical reaction proceeds everywhere in the concrete structure at the same time. For optimal control of phase field problems see e.g. [14, 63, 121].

Another real world optimal control problem with parabolic partial differential equations is the cooling of glass [62, Chapter 4.2]. In this case a coupled system of a parabolic partial differential equation and an elliptic partial differential equation with time dependent boundary condition and time dependent right hand side must be controlled. Optimal control problems

with tracking type functional for this system are discussed in [29, 62, 75, 97, 126]. Clever and Lang [29] use a cost functional which involves not only the tracking of the state but also the tracking of the gradient of the state.

1.3. Functional and Numerical analysis

In Chapter 3 and Chapter 4 we recapitulate tools from functional and numerical analysis and introduce a new a priori estimate and finite element discretization of a semi-elliptic boundary value problem.

The class of semi-elliptic equations includes all elliptic boundary equations, but also allows problems with different order of differentiation in different dimensions (see e.g. [59, Definition 1]) as in the problem

$$-y_{tt} + AAy + y = f \qquad \text{in } Q = \Omega \times (0,T), \tag{1.1}$$

with an $H_0^1(\Omega)$-elliptic operator A and the boundary conditions

$$\begin{aligned} y(\cdot,0) &= 0, & &\text{in } \Omega, & y_t(\cdot,T) - Ay(\cdot,T) &= 0, & &\text{in } \Omega, \\ y &= 0 & &\text{on } \Sigma_1, & \frac{\partial y}{\partial n_A} &= 0 & &\text{on } \Sigma_2, \\ Ay &= 0 & &\text{on } \Sigma_1, & \frac{\partial Ay}{\partial n_A} &= 0 & &\text{on } \Sigma_2, \end{aligned}$$

where $\Sigma_i = \Gamma_i \times (0,T)$ for $i = 1,2$ and $\overline{\Gamma}_1 \cup \overline{\Gamma}_2 = \partial\Omega$. It is easy to verify that the corresponding bilinear form is V-elliptic in the Sobolev space $V = H^{(2,1)}(Q)$, the Sobolev space of all $L^2(Q)$-functions whose second derivative with respect to x and first derivative with respect to t are square integrable. Equations of the type (1.1) are not only of academic interest but also appear in the context of optimal control of parabolic partial differential equations. To the knowledge of the author in the context of parabolic optimal control problems such equations have been derived first by Büttner [24] and properties, such as ellipticity, have been discussed by Neitzel, Prüfert and Slawig [88, 89]. Gong, Hinze and Zhou [46] provide a priori and a posteriori error estimates for mixed finite element discretizations.

For the analysis of this problem we introduce in Chapter 3 Sobolev spaces with variable order of differentiation in different dimensions, which are also discussed in [15, 74, 77, 79, 92, 127]. A priori regularity estimates in Besov spaces for the equation (1.1) on the unit circle are given by Triebel [127]. Estimates for general semi-elliptic equations can be found e.g. in [7, 8, 93, 59]. Even if these references are given, many results are less known than their isotropic counterparts, often only proven for special cases and the results are widely scattered in the various references. We collect all the results which are needed for the discussion of a finite element approximation of the boundary value problem (1.1). Further, a regularity estimate for semi-elliptic boundary value problem (1.1) is proven.

Beside this in particular classical Sobolev spaces are introduced and regularity estimates for elliptic and parabolic partial differential equations are given.

In the chapter about numerical analysis of partial differential equations, Chapter 4, we include the presence in the approximation of the right hand side to our analysis. So we prove also the error estimate for the Crank-Nicolson time stepping scheme for parabolic equations in the presence of a numerical evaluation of the right hand side.

Although finite element error estimates are well known for second and fourth order elliptic boundary value problems, much less is known, if the differential equation has different orders in different dimensions. So we introduce an anisotropic finite element method and prove a priori error estimates up to the order $h^2+\tau^k$ in the energy norm and up to the order $(\tau^k+h^2)(\tau+h^2)$ in the $L^2(Q)$-norm with $k=1,2,3$ for the semi-elliptic problem (1.1) in one spatial dimension. In the proof of the interpolation error we apply a technique, which was used by Rachowicz [99] for the anisotropic discretization of elliptic equations.

Oganesyan [93] provides for a similar semi-elliptic problem the error estimate

$$\|y-y_{h\tau}\|_{H^{(2,1)}(Q)} \lesssim \left(\frac{\tau^4+h^8}{\tau^2+h^4}\right)^{1/2} \|y\|_{H^{(4,2)}(Q)},$$

but gives neither a $L^2(Q)$-error bound nor numerical examples.

Another example for a semi-elliptic boundary value problem is given by the Onsager equation

$$\begin{aligned}
\left(\mathrm{e}^x\left(e^x u_{xx}\right)_{xx}\right)_{xx} + bu_{yy} &= f(x,y), && \text{in } (0,1)^2,\\
u_x(0,y) = u_{xx}(0,y) &= 0,\ \left(\mathrm{e}^x\left(\mathrm{e}^x u_{xx}\right)_{xx}\right)_x(0,y) = g(y),\\
u(1,y) = u_x(1,y) &= 0,\ \left(e^x u_{xx}\right)_x(1,y) = 0,\\
-bu_y(x,0) &= d\left(\mathrm{e}^{x/2}u_x\right)_x + \gamma_0(x),\\
bu_y(x,1) &= d\left(\mathrm{e}^{x/2}u_x\right)_x + \gamma_1(x)
\end{aligned}$$

First finite element discretizations of this equation are given in [13, 49]. Eastham and Peterson [37] use an isotropic tensor product finite element with B-spline basis functions. In contrast to our technique they only achieve second order of convergence in $L^2(\Omega)$. It is likely that the transfer of our approach leads to better approximation rates and suggestions for different step size for x and y. The technical details in the derivation of error estimates of the Onsager equation would include the definition of global $\mathcal{C}^2$-continuous finite elements and an interpolation operator which preserves the $\mathcal{C}^2$-continuity.

1.4. Parabolic optimal control problems

In Chapter 5 we consider the optimality conditions of the optimal control problem with parabolic partial differential equations. In particular we consider the optimal control problem

$$\left.\begin{aligned}
&\min_{y,u} J(y,u),\\
&\text{s.t. } My_t + Ay = Gu,\\
&\qquad My(0,\cdot) = Mv(\cdot),
\end{aligned}\right\} \tag{1.2}$$

where the cost functional $J(y,u)$ is defined by

$$\begin{aligned}
J(y,u) = \frac{\alpha}{2}\left\|M_D^{1/2}\left(y(\cdot,T)-y_D(\cdot)\right)\right\|_H^2 + \frac{\beta}{2}\int_0^T \left\|M_d^{1/2}\left(y(\cdot,t)-y_d(\cdot,t)\right)\right\|_H^2 \mathrm{d}\,t +\\
+\frac{\nu}{2}\int_0^T \left\|M_u^{1/2}u\right\|_H^2 \mathrm{d}\,t.
\end{aligned}$$

with the control u and the state y. The Hilbert space H is appropriately chosen, the desired states $y_D \in H$, $y_d(\cdot,t) \in H$ and the initial condition $v \in H$ are given. The linear and continuous operator A maps the subspace $V \subseteq H$ into its dual V^*. Further the linear continuous and self-adjoint operator $M : V^* \to V^*$, the linear and continuous operator $G : H \to V^*$ and the linear, self-adjoint, positive semi-definite and continuous operators $M_D,\ M_d,\ M_u : H \to H$ are given. The coefficients in the cost functional $\alpha,\ \beta,\ \nu \in \mathbb{R}$ are greater than or equal to zero and additionally $\nu > 0$ and $\alpha + \beta > 0$ hold. So the equality in (1.2) is in the sense of $\mathcal{C}(0,T;V^*)$.

The optimality conditions of this problem follow from the theory of optimal control with partial differential equations, which can be found e.g.. in [62, 78, 101, 128, 129]. These conditions are given by a system of equations containing the (forward) state equation and the (backward) adjoint equation which are coupled by the gradient equation, which is sometimes also called optimality condition.

We observe that the optimality conditions are a Hamiltonian system. This is less often used in the context of optimal control with partial differential equations but well known in the context of optimal control problems with ordinary differential equations [16, 17, 27]. In our numerical analysis we will use the fact that the optimality conditions are a Hamiltonian system. Finally we see that the optimality system can be reduced to a single $H^{(2,1)}(Q)$-elliptic equation (see also [24, 46, 86, 87]).

Examples for the problem (1.2) contain the following (but are not restricted to these cases):

1. Optimal control problem for parabolic partial differential equations:

$$\begin{aligned}
\min \frac{\alpha}{2} \|y(\cdot,T) - y_D(\cdot)\| + \int_0^T \frac{\beta}{2} \|y - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2} \|u\|^2_{L^2(\Omega)} \ \mathrm{d}\, t, & \\
\text{s.t. } y_t - Ay = u & \quad \text{in } (0,T] \times \Omega, \\
\frac{\partial}{\partial n_A} y = 0 & \quad \text{on } (0,T] \times \Gamma_1, \\
y = 0 & \quad \text{on } (0,T] \times \Gamma_2, \\
y(\cdot,0) = v & \quad \text{in } \{0\} \times \Omega,
\end{aligned}$$

   where the boundary $\partial\Omega$ of the domain Ω is partitioned into a Neumann boundary Γ_1 and a Dirichlet boundary Γ_2, so that $\partial\Omega = \overline{\Gamma_1} \cup \overline{\Gamma_2}$.

   The problem is well posed if we choose $y_d \in L^2((0,T), L^2(\Omega))$ and $y_D,\ v \in L^2(\Omega)$, but later we will need more regularity to show the second order convergence of discretizations.

2. Optimal control problem for a system of ordinary differential equations:

$$\begin{aligned}
\min \frac{\alpha}{2} \|y(\cdot,T) - y_D(\cdot)\|^2_{\mathbb{R}^n} + \int_0^T \frac{\beta}{2} \|y - y_d\|^2_{\mathbb{R}^n} + \frac{\nu}{2} \|u\|^2_{\mathbb{R}^n} \ \mathrm{d}\, t, & \\
\text{s.t. } My_t + Ay = Mu, & \\
y(0) = v, &
\end{aligned}$$

   We can think of the spatial discretization of a parabolic partial differential equation, where M is the mass matrix and A is the stiffness matrix.

1.5. Discretization of parabolic optimal control problems

In Chapter 6 we discretize the optimal control problem (1.2) with a tailored Crank-Nicolson time discretization scheme so that discretization and optimization commute. In this thesis the results of Apel and Flaig [4, 5] are extended to more general optimal control problems. We introduce a family of time discretizations. One of the Crank-Nicolson discretizations coincides with the application of the Störmer-Verlet scheme to the corresponding Hamiltonian system. For this scheme an a priori error estimate of second order is proven.

In Chapter 7 we discuss the finite element approximation of the single $H^{(2,1)}(Q)$-elliptic equation which was derived by eliminating the adjoint state and the control in the optimality conditions. As a conforming approximation of $H^{(2,1)}(Q)$-elliptic problems was discussed in Chapter 4, we give only a short reference to these and focus on a mixed finite element approximation. We see that this mixed finite element approximation is equivalent to the elimination of the control and adjoint state in one of the Crank-Nicolson discretizations of Chapter 6.

For time discretizations of the optimal control problem we mention here also the literature for optimal control problems with ordinary differential equations and parabolic partial differential equations. Bonnans and Laurent-Varin [16, 17] analyzed the application of symplectic partitioned Runge-Kutta schemes (SPRK) to the optimal control problem

$$\begin{aligned} &\min \Phi\left(y(T)\right) \\ &\text{s.t. } y_t = f\left(u(t), y(t)\right), \qquad y(0) = y_0, \end{aligned} \tag{1.3}$$

with ordinary differential equations and terminal observation in the target function. With the aim that both approaches (optimize then discretize and discretize then optimize) result in the same scheme, they obtained order conditions up to order six, but they mention no method, which fulfills the conditions. On the other hand a tracking type cost functional for an ordinary differential equation can be treated as terminal cost functional (1.3) on additional components of y, see [54, Section 1]: The optimal control problem

$$\begin{aligned} &\min_u \int_0^T \frac{1}{2} \left\|y - y_d\right\|_H^2 + \frac{\nu}{2} \left\|u\right\|_H^2 \mathrm{d}\, t \\ &\text{s.t. } y_t = f(u(t), y(t)), \qquad y(0) = v \end{aligned}$$

is equivalent to the following optimal control problem with terminal observation

$$\begin{aligned} &\min\ z(T) \\ &\text{s.t. } y_t = f(u(t), y(t)), && y(0) = v, \\ &\qquad z_t = \frac{1}{2} \left\|y(t) - y_d(t)\right\|_H^2 + \frac{\nu}{2} \left\|u(t)\right\|_H^2, && z(0) = 0. \end{aligned}$$

Nevertheless as we discuss optimal control problems with partial differential equations, we discuss a cost functional, which may contain a term for the tracking over the full space-time cylinder and a term for the tracking of the terminal state, as we use the structure of the solution of the partial differential equation.

Some of the order conditions of Bonnans and Laurent-Varin can also be found in two papers by Hager [54, 55]. Moreover Chyba, Hairer and Vilmart [27] analyze for what kind of optimal

control problems with ordinary differential equations symplectic methods are superior to non-symplectic methods and come to the conclusion that this depends on the optimal control problem under consideration.

For optimal control problems with ordinary differential equations and constraints we mention an article by Dontchev, Hager and Veliov [35]. They develop a second order Runge-Kutta method for control constrained problems and prove an error estimate for the case, when the derivative of the optimal control has bounded variation.

All these articles deal only with ordinary differential equations but not with partial differential equations.

In recent papers about the optimal control of parabolic partial differential equations the mentioned results about the interchangeability of discretization and optimization for certain time stepping schemes seem to be unknown, or at least uncited, see e.g. [11, 33, 60, 61, 80, 81, 82, 83, 84, 86, 87, 115].

For the optimal control of parabolic partial differential equations space-time finite element methods are very common. In several papers Vexler and coworkers have developed such methods, based on a continuous or discontinuous Galerkin method for the time discretization [11, 80, 81, 82, 83, 84, 115], see also [86]. They also achieve the interchangeability of discretization and optimization. Both, Meidner and Vexler [82] and Neitzel, Prüfert and Slawig [87], discuss optimal control problems with parabolic partial differential equations with control constraints. The approach of space-time finite elements is also used by Deckelnick and Hinze [33], who consider state constraints. Almost all of these discretizations are first order in time. Higher order estimates can be found in [4, 5, 81, 83, 84].

Due to the coupling of the forward in time state equation and the backward in time adjoint equation all these discretizations can not be resolved time step by time step but result in a huge system of equations. Multigrid methods on the space-time grid are particularly efficient for their solution, see the fundamental work by Borzì [18] which extends earlier works by Hackbusch, e.g. [52], and the transfer to flow control problems by Hinze, Köster and Turek [60, 61]. The first order implicit Euler scheme is used for time discretization in all these papers. As the $L^2(\Omega)$-approximation error for linear finite elements is of second order this suggests the choice of $\tau = \mathcal{O}(h^2)$ for balancing the errors. With further refinements this leads to an anisotropic mesh and should influence the smoothing and (semi-)coarsening techniques. With the Crank-Nicolson method, space and time discretization have the same order, and the choice $\tau = \mathcal{O}(h)$ is possible for a well-balanced error distribution. We assume that the isotropic elements simplify the solving techniques.

2. Young concrete

Contents

In this chapter we introduce an optimal control problem of practical interest, which was also discussed in [3]. The dealing with practical problems from engineering in mathematics is a challenge as these problems typically model many phenomena, involve nonlinearities and are already numerically solved by engineers. But even if these problems are simulated often questions concerning the existence or regularity of solutions remain open. The confidence of mathematicians discussing such problems is that a new view to a well known problem also yields new ideas or algorithms which are also of interest for the engineers.

Modern civil engineering and architecture would not be possible without concrete. So measurement and simulation of all the different aspects of concrete is a wide field of activity in civil engineering. Our aim is not the overall simulation of all aspects of concrete, but understanding of a mathematical model and the simulation of the temperature distribution and a prediction of thermal cracks during hydration. Such macroscopic thermal cracks are not wanted as they influence the usability or elegance of the final structure.

The simulation of hydration is needed as input for the calculation of material properties, strengths and stresses and therefore also for the calculation of a crack criterion [114]. Some crack criteria consider only the temperature but generally a criterion would be preferred which also accounts for the development of stresses [104, 111].

We summarize the modeling of the hydration, the thermo-mechanical properties in Section 2.1. In Section 2.2 we discuss the mechanical properties of young concrete, crack criteria are discussed in Section 2.3.

If, after the simulation of the hydration, the crack criterion indicates that there will be thermal cracks, some changes in the input data are required. The possibilities of measures are also described in Section 2.3. The variation of the input data can be interpreted as optimization. We discuss simulation based optimization and model based optimization in Section 2.4. Furthermore we introduce a family of abstract optimal control problems for the hydration of concrete in this section.

2.1. Hydration of concrete

2.1.1. General idea

Concrete is produced from the mixing of cement, aggregate, admixtures and water. By varying the ratio and kind of the ingredients a wide range of concretes for different purposes can be composed. After mixing, an exothermic chemical reaction known as hydration starts. During this reaction the liquid or plastic viscous mass develops the solid properties of the concrete. The specific solid properties of the concrete are determined by the ratio of the ingredients.

The hydration can be studied on different scales. Studies on microscopic scales inspect the chemical reaction on a molecular scale. On the other hand studies on the macroscopic scale investigate the development of heat, mechanical properties and cracks. Therefore the molecular processes are hidden in an averaged description. We want to introduce and simulate such a model.

The basis of this description is the degree of hydration which defines the fraction of the reaction which has occurred until a specific point in time. As this ratio can not be measured directly different indirect definitions are in use. We use a common definition (see e.g. Eierle [38]) of the degree of hydration based on the heat development

$$\alpha(t,x) = \frac{Q(t,x)}{Q_\infty}, \tag{2.1}$$

where Q_∞ is the overall heat that is produced by hydration and $Q(t,x)$ is the heat produced until the time t. Note, that the degree of hydration can assume different values in different points of the same concrete structure.

A model for the degree of hydration is one of the two main ingredients of a description of concrete. On the other hand the reaction rate of this reaction depends on the temperature $y = y(t,x)$, measured in °C. Therefore time in the description of the hydration is replaced by the maturity (effective age). A common form for the maturity is $\tau(t, y(\cdot,\cdot)) = \int_0^t g\left(y(\vartheta,x)\right) \mathrm{d}\,\vartheta$ with an appropriate function $g(\cdot)$.

2.1.2. Maturity

In literature different choices for the maturity are known. We present maturities which are widely used in engineering literature. Freiesleben-Hansen introduced a formula which can be motivated by chemical reaction kinetics (see e.g. Eierle [38])

$$\tau(t, y(x,\cdot)) = \int_0^t \exp\left(\frac{A}{R}\left(\frac{1}{293} - \frac{1}{273 + y(x,\vartheta)}\right)\right) \mathrm{d}\,\vartheta, \tag{2.2}$$

where R is the universal gas constant and the activation energy A is a material parameter. In general A may depend on the temperature but according to Krauß [71, (5.22)] an activation energy which is independent of the temperature is applicable to a large class of cements.

A simpler maturity was introduced by Saul as (see e.g. Eierle [38])

$$\tau(t, y(x,\cdot)) = \int_0^t \frac{y(x,\vartheta) + 10}{30} \mathrm{d}\,\vartheta, \tag{2.3}$$

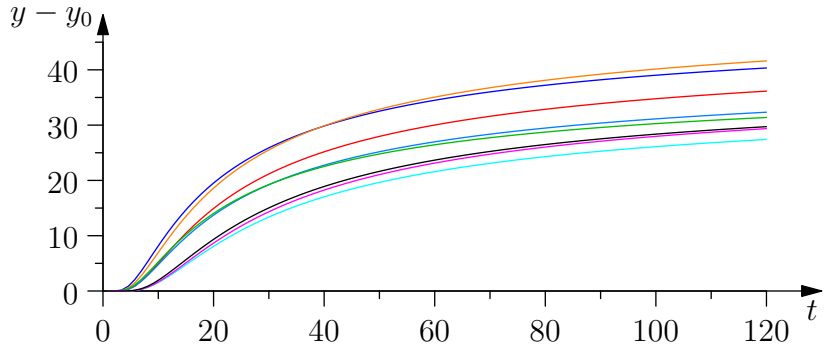


Figure 2.1.: The adiabatic temperature rise for different concrete recipes is plotted according to the model parameter in [34, Figure 8.4 and Table 8-3]. In adiabatic experiments the temperature y is directly proportional to the degree of hydration $\alpha = \frac{Q(t,x)}{Q_\infty}$.The temperature difference of the temperature y and the fresh concrete temperature y_0 is plotted in $^\circ C$ and the time t in hours.

This maturity is independent of the concrete in use. Therefore it seems very logical that the maturity of Freisleben-Hansen (2.2) will produce more realistic simulations. Finally Röhling [102] introduces

$$\tau(t, y(x,\cdot)) = \int_0^t \left(\frac{y(x,\vartheta)+15}{35}\right)^d \mathrm{d}\,\vartheta, \qquad d \approx 2 \tag{2.4}$$

as approximation of the maturity of Freiesleben-Hansen for some specific cements. For $d = 1$ this implies an affine linear relation of the temperature and the derivative of the maturity as the maturity of Saul (2.3).

But nevertheless all the maturities have the abstract form

$$\tau(t, y(x,\cdot)) = \int_0^t g\left(y(x,\vartheta)\right) \mathrm{d}\,\vartheta. \tag{2.5}$$

In our analysis we will use this abstract notation.

2.1.3. Adiabatic heat development

As mentioned in Section 2.1.1 the basis of the description of hydration is the degree of hydration. Even if there is a wide range of applications and compositions of concretes, they all share the basic shape of the development of the degree of hydration (see Figure 2.1). The temperature development in an adiabatic regime is as follows: In the beginning for some time, there is no heat development, then the heat development starts slightly, fastens up and ends in a saturation. Eierle [38] compares different approaches to the modeling of this shape. Wesche [130] proposes

$$\alpha = \alpha(\tau) = \mathrm{e}^{a\tau^b} \qquad a, b < 0. \tag{2.6}$$

Another very common approach is the model of Jonasson [67]

$$\alpha = \alpha(\tau) = \mathrm{e}^{a\left[\log\left(1+\frac{\tau}{\tau_k}\right)\right]^b} \qquad a,\ b < 0,\ \tau_k > 0. \tag{2.7}$$

The parameters $a,\ b$ in the model by Wesche (2.6) and the parameters $a,\ b,\ \tau_k$ in the model by Jonasson (2.7) are obtained by measurement of the adiabatic heat development and a parameter fit. In the model of Jonasson (2.7) the parameter a is often set to -1 (see [51, Section 2.3.2.5] or [94, Formula (4.8) and (4.9) in Section 4.1.2]).

We present the different approaches but we do not want to evaluate them. In our mathematical setting we write the model for the adiabatic heat development as an integral formulation

$$\alpha(\tau) = \int_0^\tau h(\vartheta)\,\mathrm{d}\,\vartheta. \tag{2.8}$$

The development of the mechanical properties of concrete in every point is driven by the hydration, so the simulation of the progress of hydration is an important task. As the degree of hydration is measured by the heat development, the heat distribution must be calculated.

For a new concrete structure Ω the heat distribution is governed by the heat equation with the hydration rate as heat source. Therefore the temperature $y = y(t,x)$ is determined by the heat equation

$$c\rho y_t - \nabla\cdot(\lambda\nabla y) = Q_\infty\dot{\alpha} = Q_\infty\frac{\partial\alpha}{\partial\tau}\frac{\partial\tau}{\partial t} = Q_\infty h(\tau)g(y) \qquad \text{in } \Omega. \tag{2.9}$$

We assume that the material parameters $c,\ \rho$ and λ are independent of space, time, temperature and degree of hydration (see Krauß [71, Chapter 5.1]).

Since the maturity τ is defined in equation (2.5) by an integral over y the heat equation (2.9) is a integrodifferential equation. For the analysis and numerical analysis of this equation we prefer to compute τ as additional function, as proposed by Hairer, Norsett and Wanner [58, Chapter II.17, p.351f]. So we have the additional differential equation

$$\tau_t = g(y) = \frac{\partial\tau}{\partial t}, \tag{2.10}$$

to consider. Therefore, if we apply a method of lines with fixed spatial discretization, we can use for the discretization in time any discretization method for ordinary differential equations and do not need a method specialized for integrodifferential equations as discussed e.g. by Chuanmiao and Tsimi [26].

To complete the system of equations we still need to specify boundary and initial conditions. As initial condition for the equation for the maturity (2.10) we choose zero initial data with respect to the definition of the maturity (2.5). For the temperature y the initial condition is the temperature y_0 of the concrete after mixing and installation. On the boundary we assume Robin conditions

$$\frac{\partial y(t,x)}{\partial\nu} = \sigma(t,x)\left(y_{\mathrm{BND}}(t,x) - y(t,x)\right).$$

This boundary condition describes the heat transfer through the formwork. Typically the coefficient σ is not any $L^2\left((0,T]\times\partial\Omega\right)\cap L^\infty\left((0,T]\times\partial\Omega\right)$ function but a (piecewise) constant in space and has only finitely many jumps between discrete values in time. These jumps occur when the formwork is removed or changed. With these Robin boundary conditions it is also possible to simulate Neumann boundary conditions for symmetry axes ($\sigma = 0$ on Γ_N).

For the ambient temperature we assume the smooth distribution

$$y_{\mathrm{BND}}(t) = y_{\mathrm{med}} + \frac{1}{2} y_{\mathrm{Delta}} \cos\left(\frac{2 \cdot \pi}{24} \cdot t\right),$$

where t is measured in hours and the values y_{med} and y_{Delta} can be chosen according to assumed weather conditions or to the values summarized in the German code DIN 4710 (see [43]).

Altogether the temperature distribution during hydration is described by

$$\left.\begin{aligned} \tau_t &= g(y) && \text{in } (0,T] \times \Omega, \\ c\rho y_t - \lambda \Delta y &= Q_\infty h(\tau) g(y) && \text{in } (0,T] \times \Omega, \\ \frac{\partial y(t,x)}{\partial \nu} &= \sigma(t,x)\left(y_{\mathrm{BND}}(t,x) - y(t,x)\right) && \text{on } (0,T] \times \partial\Omega, \\ \tau(0,x) &= 0 && \text{in } \{0\} \times \Omega, \\ y(0,x) &= y_0(x) && \text{in } \{0\} \times \Omega. \end{aligned}\right\} \tag{2.11}$$

With realistic assumptions on the properties of the functions h and g the existence and uniqueness of the solution y and τ of the problem (2.11) can be proven. The uniqueness follows directly by discussing the difference of two solutions and the choice of an arbitrary test function. For the existence of the solution Schauders fixed point theorem can be used. For the details we refer to Benedix [12, Theorem 7.2 and Theorem 7.10].

2.2. Mechanics of young concrete

2.2.1. Quantities in mechanics of young concrete

The heat distribution influences the development of the mechanical properties of concrete. We consider Young Modulus E, tensile strength f_{ct} and compressive strength f_{cc}. The Poisson ratio ν is assumed to be constant. These quantities develop during hydration, and therefore commonly used models (see e.g. Krauß [71]) describe them depending on the degree of hydration α. We use the following model (see [71, 104, 107, 111])

$$E(\alpha) = E_\infty \left(\frac{\alpha - \alpha_0}{1 - \alpha_0}\right)^{\gamma_1} \tag{2.12}$$

$$f_{ct}(\alpha) = f_{ct,\infty} \left(\frac{\alpha - \alpha_0}{1 - \alpha_0}\right)^{\gamma_2} \tag{2.13}$$

$$f_{cc}(\alpha) = f_{cc,\infty} \left(\frac{\alpha - \alpha_0}{1 - \alpha_0}\right)^{\gamma_3}. \tag{2.14}$$

The final values E_∞, $f_{ct,\infty}$ and $f_{cc,\infty}$ and the exponents γ_1, γ_2 and γ_3 are material parameters, typical values for γ_i are $\gamma_1 = \frac{1}{2}$, $\gamma_2 = 1$ and $\gamma_3 = \frac{3}{2}$ (see e.g. [104, Section 4.3.2.1] or [107, Section 3.2.1]). Further the value α_0 marks the degree of hydration for which solid body properties can be measured the first time. This value also depends on the concrete in use.

With these material properties it is possible to compute the thermoelastic stresses. In thermoelasticity it is assumed that temperature changes induce thermal strains of the form

$$\varepsilon_{\mathrm{therm}}(y) = I_3 \alpha_{\mathrm{therm}} \left(y(\cdot, t_1) - y(\cdot, t_0)\right) \tag{2.15}$$

with

$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

where no shear strains are introduced (see e.g. Barber [9]). The thermal strain ε_{therm} is a tensor of second rank and α_{term} and $[y]_{t_0}^{t_1}$ are scalars. The coefficient of thermal expansion α_{term} can be assumed independent of the temperature, at least for temperatures below $60°C$ [102, Chapter 5.4]. Further these thermal strains are additive to the elastic strains.

For the material model we discuss either a linear elastic material law with time dependent Young modulus or a viscoelastic model according to Röhling [102]. In both cases, as the Young modulus is time dependent, we have a nonlinear material behavior, even if we use linear elastic material law. Due to this time dependent Young modulus it is possible that, after loading and unloading, stresses remain in the structure.

The mechanical properties can be calculated in a post processing step after simulation of the hydration.

2.2.2. Linear elastic material law

The well known linear elastic material law has the form

$$\sigma = \mathbb{C}(E, \nu) : \varepsilon,$$

which can also be written in index-notation as

$$\sigma_{ij} = \mathbb{C}_{ijkl}(E, \nu)\varepsilon_{kl},$$

where the Einstein summation convention is used, i.e.

$$\sigma_{ij} = \sum_{k=1}^{d} \sum_{l=1}^{d} \mathbb{C}_{ijkl}(E, \nu)\varepsilon_{kl}.$$

The fourth order tensor $\mathbb{C}$ possesses the symmetry properties (see e.g. [68])

$$\mathbb{C}_{ijkl} = \mathbb{C}_{jikl} = \mathbb{C}_{ijlk} = \mathbb{C}_{klij},$$

and depends in the usual way on the Young modulus E and the Poisson ratio ν, so that the material law can be written in matrix-vector notation as (see [22, Formula (VI.3.6)])

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{23} \end{bmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & \nu & & & \\ \nu & 1-\nu & \nu & & & \\ \nu & \nu & 1-\nu & & & \\ & & & 1-2\nu & & \\ & & & & 1-2\nu & \\ & & & & & 1-2\nu \end{pmatrix} \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}.$$

Note that some authors use an other scaling for the lower part of the material law (see e.g. [10, Table 4.3 in Section 4.2.3]). They use

$$\begin{bmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{12}\\ \sigma_{13}\\ \sigma_{23}\end{bmatrix} = \frac{E(1-\nu)}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1 & \frac{\nu}{1-\nu} & \frac{\nu}{1-\nu} & & & \\ \frac{\nu}{1-\nu} & 1 & \frac{\nu}{1-\nu} & & & \\ \frac{\nu}{1-\nu} & \frac{\nu}{1-\nu} & 1 & & & \\ & & & \frac{1-2\nu}{2(1-\nu)} & & \\ & & & & \frac{1-2\nu}{2(1-\nu)} & \\ & & & & & \frac{1-2\nu}{2(1-\nu)} \end{pmatrix} \begin{bmatrix}\varepsilon_{11}\\ \varepsilon_{22}\\ \varepsilon_{33}\\ 2\varepsilon_{12}\\ 2\varepsilon_{13}\\ 2\varepsilon_{23}\end{bmatrix},$$

which is an equivalent formulation of the material law.

For constant Young modulus $E(t)$ and some $t^* \in (0, T]$ the thermal stresses can be computed (see [9, 116]) to

$$\sigma(\cdot, t^*) = \mathbb{C}(E, \nu) : \varepsilon_{\text{therm}} = \mathbb{C}(E, \nu) : I_3 \alpha_{\text{therm}} (y(\cdot, T) - y(\cdot, 0)).$$

But as the Young modulus $E(t)$ is not constant in time, we use a stepwise approach with $t_i = \frac{t^*}{N} i$ to compute the thermal stresses as

$$\sigma(\cdot, t^*) = \sum_{i=0}^{N-1} \mathbb{C}(E(t_{i+1}), \nu) : I_3 \alpha_{\text{therm}} (y(\cdot, t_{i+1}) - y(\cdot, t_i)).$$

For the identification of the underlying continuous (differential) equation for the thermal stresses, we subtract the thermal stresses for the times t^* and $t^* - \tau$ with $\tau = \frac{T}{N}$ and divide by the time step size τ, which yields

$$\frac{\sigma(\cdot, t^*) - \sigma(\cdot, t_{N-1})}{\tau} = \mathbb{C}(E(t^*), \nu) : I_3 \alpha_{\text{therm}} \frac{y(\cdot, t^*) - y(\cdot, t_{N-1})}{\tau}.$$

Passing to the limit with $N \to \infty$, which is equivalent to $\tau \to 0$, gives the underlying differential equation

$$\frac{\partial}{\partial t} \sigma(\cdot, t^*) = \mathbb{C}(E(t^*), \nu) : I_3 \alpha_{\text{therm}} \frac{\partial}{\partial t} y(\cdot, t^*).$$

So a linear material law with time dependent Young modulus for thermoelasticity leads to a differential equation with a time dependent coefficient. Together with the condition that there are no thermal stresses initially, the initial value problem for the thermal stresses is

$$\left.\begin{aligned} \frac{\partial}{\partial t} \sigma(\cdot, t) &= \mathbb{C}(E(t), \nu) : I_3 \alpha_{\text{therm}} \frac{\partial}{\partial t} y(\cdot, t), \\ \sigma(\cdot, 0) &= 0. \end{aligned}\right\} \tag{2.16}$$

A linear elastic material law is the simplest possibility of the description of materials. And even this material law implies a non-linear material behavior, as after one cycle of heating up from a certain temperature and cooling down to this temperature stresses may remain in the material due the time-dependent Young modulus.

As seen in the work by Onken and Rostásy [94, Section 2.2.3] a linear elastic material law tends to overestimate the stresses, so a linear elastic material law seems to be on the safe side. A more realistic choice would be a viscoelastic material law. We do not evaluate the choice which material law is used for computations and present a viscoelastic material law in the next section.

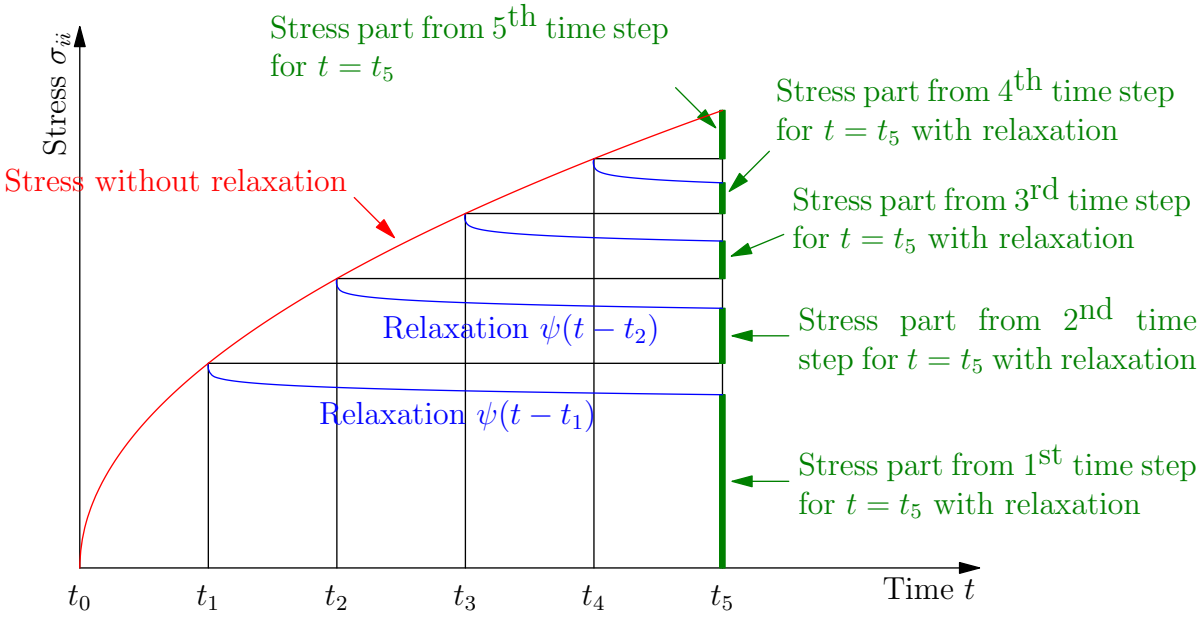


Figure 2.2.: Schematic comparison of the stress development for a viscoelastic material law a linear elastic material law. For the stress σ_{ii} in a point under the assumption of a viscoelastic material law at $t = t_5$ one has to add the green parts, whereas the stress without relaxation is given by the red line.

2.2.3. Viscoelastic material law

For a more involved material we discuss a viscoelastic material law, as introduced by Röhling [102]. The viscoelastic material law considers the relaxation of stresses, i.e. a stress increment, once introduced, will reduce over time. This behavior is sketched in Figure 2.2. At some time t^* and with equidistant time steps $t_i = \frac{t^*}{N} i$ the viscoelastic stresses, which is introduced in [102, formula (6.6)], are given by

$$\sigma(\cdot, t^*) = \sum_{i=0}^{N-1} \mathbb{C}\left(E(t_{i+1}), \nu\right) : I_3 \alpha_{\text{therm}} \left(y(\cdot, t_{i+1}) - y(\cdot, t_i)\right) \psi(t^* - t_{i+1}), \tag{2.17}$$

where the monotone decreasing function ψ with $\psi(0) = 1$ describes the relaxation. As the material law is introduced in a time discrete version, we want to reestablish the underlying differential equation. As in the linear case we subtract the stresses for the times t^* and t_{N-1}, but due to the relaxation described by ψ, there is no cancellation of terms. The difference, after dividing by τ, is given by

$$\begin{aligned}\frac{\sigma(\cdot, t^*) - \sigma(\cdot, t_{N-1})}{\tau} &= \mathbb{C}\left(E(t^*), \nu\right) : I_3 \alpha_{\text{therm}} \frac{\left(y(\cdot, t^*) - y(\cdot, t_{N-1})\right)}{\tau} \cdot \psi(t^* - t_{N-1}) \\ &\quad + \sum_{i=0}^{N-2} \mathbb{C}\left(E(t_{i+1}), \nu\right) : I_3 \alpha_{\text{therm}} \left(y(\cdot, t_{i+1}) - y(\cdot, t_i)\right) \\ &\qquad \cdot \frac{\psi(t^* - t_{i+1}) - \psi(t_{N-1} - t_{i+1})}{\tau}.\end{aligned}$$

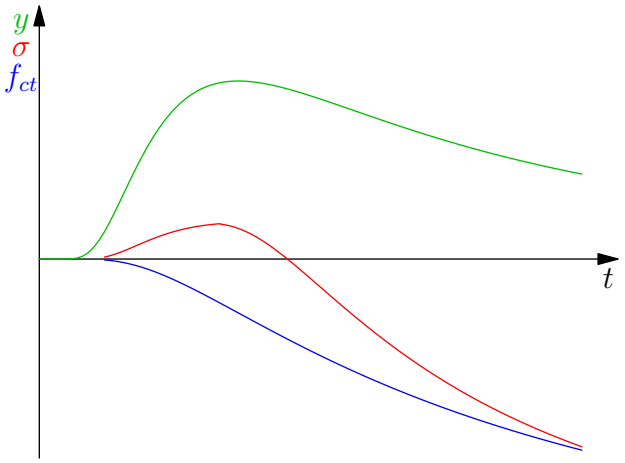


Figure 2.3.: Schematic development of the temperature y (in green), the tensile strength f_{ct} (in blue) and the thermal stresses σ (in red) at some point in the wall. If the absolute value stresses are greater then the absolute value of the tensile strength, cracks may occur.

The finite sum on the right hand side is a sum, which lies between the Riemann-Stieltjes lower and upper sum. The Riemann-Stieltjes lower and upper sum and the Riemann-Stieltjes integral are introduced in Appendix A or e.g. in the text books by Rudin [112, Chapter 6] and Smirnow [119, Chapter 1]. By passing to the limit $\tau \to 0$ we end up with the Riemann-Stieltjes integrodifferential equation

$$\begin{aligned}\frac{\partial}{\partial t}\sigma(\cdot,t^*) &= \mathbb{C}\left(E(t^*),\nu\right) : I_3\alpha_{\text{therm}}\frac{\partial}{\partial t}y(\cdot,t^*) \\ &\quad + \int_0^{t^*} \mathbb{C}\left(E(t),\nu\right) : I_3\alpha_{\text{therm}}\psi'(t^*-\tau)\,\mathrm{d}\,y(\cdot,t).\end{aligned}$$

With the Theorem A.3 (or [112, Theorem 6.17]) about the connection of Riemann and Riemann-Stieltjes integrals, the integrodifferential equation can be written with a Riemann integral. Due to the fact, that initially there are no stresses, the continuous initial value problem for the stresses with the viscoelastic material law (2.17) is given by

$$\left.\begin{aligned}\frac{\partial}{\partial t}\sigma(\cdot,t^*) &= \mathbb{C}\left(E(t^*),\nu\right) : I_3\alpha_{\text{therm}}\frac{\partial}{\partial t}y(\cdot,t^*) \\ &\quad + \int_0^{t^*} \mathbb{C}\left(E(t),\nu\right) : I_3\alpha_{\text{therm}}\psi'(t^*-\tau)\frac{\partial}{\partial t}y(\cdot,\tau)\,\mathrm{d}\,\tau, \\ \sigma(\cdot,0) &= 0.\end{aligned}\right\} \tag{2.18}$$

2.3. Crack criteria

The computation of the mechanical properties is the base of the decision whether thermal cracks in the concrete occur or not. The qualitative development of temperature, tensile strength and thermal stresses is drafted in Figure 2.3. In the first phase of hydration the concrete is still liquid and has no rigid body properties. At some point in time, when the degree of hydration reaches

α_0, rigid body properties are measurable for the first time, then the temperature still rises and therefore small compressive stresses establish as the Young modulus is still small. After the temperature maximum is reached larger tensile stresses can be observed as the Young modulus is already larger. Cracks occur when the tensile stress is larger than the tensile strength.

After this short discussion it is clear that no cracks will occur because of thermal stresses if

$$|\min(0, \sigma(t,x))| \leq |f_{\mathrm{ct}}(t,x)| \qquad \forall (t,x) \in (0,T] \times \Omega. \tag{2.19}$$

In engineering practice one would even add some safety factor so that

$$\frac{|\min(0, \sigma(t,x))|}{|f_{\mathrm{ct}}(t,x)|} \leq k \qquad \forall (t,x) \in (0,T] \times \Omega \tag{2.20}$$

with some $k < 1$ must hold [104, 110, 111].

Crack criteria of this kind need the computation of mechanical properties which is more expensive than the computation of the heat distribution. As the prediction of cracks in young concrete has a long history in civil engineering there are temperature criteria which predict cracks only with the information on the heat distribution. These criteria are computable with less computational effort and ignore solid mechanical properties of the young concrete so they must be less accurate. But often they are exact enough for the assessment of the crack risk (see [110, Section 6.4.1]). If we consider a new concrete wall on a old bottom plate as in Figure 2.4, a temperature criterion would read as

$$|y(t,x_1) - y(t,x_2)| \leq 15K \qquad \forall t \in (0,T], \tag{2.21}$$

where x_1 and x_2 denote the midpoints of the new wall and the bottom plate as denoted in Figure 2.4 (see e.g. [110, Section 6.3] and [102, Sections 7.4 and 7.5 with Tables 7.3–7.5]).

If a crack criterion indicates that cracks are likely, one has to consider counteractive measures. Different measures are proposed in literature and include, but are not restricted, to the following examples. Nietner [90] gives an overview about possibilities how to influence the thermomechanical behavior of concrete. He mentions the fresh concrete temperature, formwork, cooling pipes, the influence of the concrete recipe and also the influence of the size of the casting segments. As the choice of casting segments influences the geometry of the computational domain, we do not discuss this possibility here, but for the other measures further references are given.

Nietner and Schmidt [91] and Braasch [21] propose to reduce the fresh concrete temperature by replacing the water by crashed ice. Nietner and Schmidt present a diagram in [91, Figure 9] for the choice of the largest possible fresh concrete temperature. This approach is based on simulations with different values of the input parameter and therefore it can be understood as simulation based optimization approach. Pree [98] also gives lower and upper bounds of the fresh concrete temperature. Lower bounds of fresh concrete temperature are of interest if low ambient temperatures influence the progress of the hydration. He also gives experienced data for the costs for cooling and warming.

Braasch [21] discusses the influence of different formwork and the influence on the heat transfer coefficient in the boundary condition.

For the cooling during the hydration we mention the use of cooling pipes. Huckfeldt [65] discusses simplified models for the cooling with water in the pipes. Another possibility of the

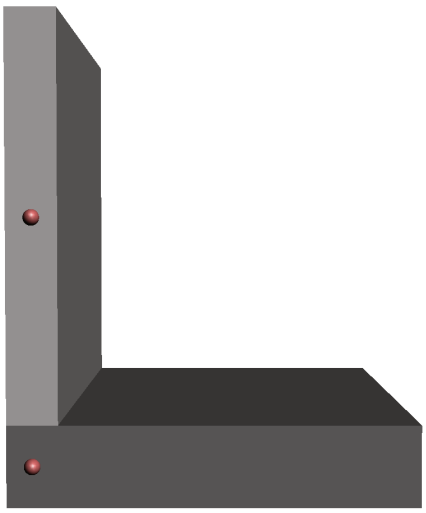


Figure 2.4.: Example for a new wall on an existing basement. For the common temperature criterion the temperature difference between the two red points x_1 and x_2 is measured.

modeling of the cooling is described in the master's thesis of Gehrmann [44], which models directly the physics and is discussed in the Appendix B. Staffa [123] also uses cooling pipes for new structures and heating pipes for warming the existing structures to reduce temperature differences. The Danish Road Directorate's research program HETEK discusses in its technical report [96] the duration of cooling and warming of concrete structures.

Benedix [12] proposes a simple model for the influence of the concrete recipe on the heat development.

For the planing of measures one would not allow any of the measures mentioned above but select the measures which are possible and realizable.

2.4. Towards optimal control of the hydration of concrete

The introduction of the measurements for minimizing the risk for cracks motivates, that this problem can be seen as an optimization problem or optimal control problem. So we introduce the control u, which describes the influence of the measures on the system. The problem can be written as abstract optimal control problem as

$$\left.\begin{aligned} \min J(y,\tau,u) \\ \text{s.t. } y,\ \tau \text{ solve the system (2.11)}, \\ \text{where the control } u \text{ influences some input data of (2.11)} \\ \text{and a crack criterion is fulfilled,} \end{aligned}\right\} \tag{2.22}$$

where J includes realistic costs for the control u and additionally a model for cooling pipes can be included. The costs for the control may appear as natural costs for change of formework, cooling and crack repair, in the sense of a penalty method. An example of such a problem is the

following, where the fresh concrete temperature is used as control and a temperature criterion is used,

$$\left.\begin{aligned}
\min \|u\|^2_{L^2(\Omega_1)} & & \\
\text{s.t. } c\rho y_{1,t} - \lambda \Delta y_1 &= Q_\infty h(\tau) g(y) && \text{in } (0,T] \times \Omega_1, \\
c\rho y_{2,t} - \lambda \Delta y_2 &= 0 && \text{in } (0,T] \times \Omega_2, \\
\tau_t &= g(y_1) && \text{in } (0,T] \times \Omega_1, \\
\frac{\partial y_1(t,x_1)}{\partial \nu} &= \sigma(t,x)\left(y_{\mathrm{BND}}(t,x) - y_1(t,x)\right) && \text{on } (0,T] \times \partial\Omega_1 \setminus \Omega_2, \\
\frac{\partial y_2(t,x)}{\partial \nu} &= \sigma(t,x)\left(y_{\mathrm{BND}}(t,x) - y_2(t,x)\right) && \text{on } (0,T] \times \partial\Omega_2 \setminus \Omega_1, \\
y_1(t,x) &= y_2(t,x) && \text{on } (0,T] \times \partial\Omega_1 \cap \partial\Omega_2, \\
\tau(0,x) &= 0 && \text{in } \{0\} \times \Omega_1, \\
y_1(0,x) &= y_{0,1}(x) - u(x) && \text{in } \{0\} \times \Omega_1, \\
y_2(0,x) &= y_{0,2}(x) && \text{in } \{0\} \times \Omega_2, \\
|y_1(t,x_1) - y_2(t,x_2)| &\leq 15K && t \in [0,T].
\end{aligned}\right\} \tag{2.23}$$

The initial temperature $y_{0,1}$ is cooled by the control u in the fresh concrete in Ω_1. The temperature in the bottom plate Ω_2 can not be controlled. For practical reasons $y_{0,1}$ would be the temperature at which the concrete can be produced without any additional costs for cooling. The cooling of the fresh concrete can be performed by the use of cold water or even crashed ice.

As there are different ways to cool the fresh concrete down, the costs for the cooling are clearly nonlinear. We approximate these costs with a quadratic cost functional which can be motivated by the interpolation of the specific costs for the different cooling possibilities.

For the norm in the cost functional we can use the $L^2(\Omega_1)$-norm if we assume that the fresh concrete can be produced and embedded with different temperatures in different places. Under the assumption that the temperature should be constant in the concrete the functional can be replaced by the square of the temperature reduction, i.e. u^2. After discretization, in space with finite elements and in time with a time discretization scheme, as the Crank-Nicolson scheme, the problem can be solved using solvers for nonlinear optimization problems.

This approach can be seen as the discretize-then-optimize of an optimal control problem of the form

$$\begin{aligned}
\min \quad & \int_0^T j_1(y) + j_2(\tau) + j_3(u) \,\mathrm{d}\, t \\
\text{s.t. } & y_t + A(u)y = f(y,\tau,u), \\
& \tau_t = g(y), \\
& \frac{\partial y}{\partial n} = \sigma(t,x,u)(y_{\mathrm{BND}} - y), \\
& \tau(0,x) = 0, \\
& y(0,x) = y_0(x), \\
& \text{Crack criterion fulfilled},
\end{aligned}$$

where cost functional consists of the functionals j_1, j_2 and j_3. With these functionals it is possible to describe the control costs, tracking of a desired state or penalization of a crack

criterion. This kind of optimal control problem was introduced in [3]. Some instances of this abstract optimal control problem have been discussed by Benedix [12], where adaptive numerical methods have been used for the solution.

In particular Benedix discusses the following three optimal control problems, where he uses cost functionals, which only include the control costs, i.e. the difference to an initial control is penalized. First the initial temperature and heat transfer coefficient in the Robin boundary condition are the controls and the state constraint $y \leq 72°C$ is obtained. As second example Benedix considers a simple model for the influence of the concrete recipe to the heat development. So the concrete recipe is the control and as state constraint the crack criterion $|y(x_1,t) - y(x_2,t)| \leq 15K$ is used. As last example he applies the model of Huckfeldt [65] for cooling pipes. In this example the flow rate is the control and the state constraint $y \leq 57°C$ is considered. For the three examples adaptive numerical calculations are given. They prove the suitability of his approach.

For our discussion of optimal control problems with parabolic partial differential equations we focus in the following chapters on the simpler and still challenging problem

$$\begin{aligned} \min_{y,u} \frac{\alpha}{2} \left\| M_D^{1/2} \left(y(\cdot,T) - y_D(\cdot) \right) \right\|_H^2 + \frac{\beta}{2} \int_0^T \left\| M_d^{1/2} \left(y(\cdot,t) - y_d(\cdot,t) \right) \right\|_H^2 \mathrm{d}\, t \\ + \frac{\nu}{2} \int_0^T \left\| M_u^{1/2} u \right\|_H^2 \mathrm{d}\, t, \\ \text{s.t. } My_t + Ay = Gu, \\ My(0,\cdot) = Mv(\cdot), \end{aligned}$$

with a linear differential operator A and $H = L^2(\Omega)$.

In preparation of this we introduce the tools from functional analysis and numerical analysis in the next two chapters.

3. Functional analysis and partial differential equations

Contents

In this Chapter we repeat some basic facts from functional analysis and discuss the regularity of solutions of partial differential equations.

First we introduce our notation for domains and introduce very few facts from functional analysis, for an overview about this topic we refer to text books about functional analysis, e.g. [72]. Then the well known Sobolev spaces are introduced. After the isotropic Sobolev spaces we discuss also anisotropic Sobolev spaces with mixed order of differentiation and transfer well known results from isotropic Sobolev spaces to the less studied anisotropic spaces. Isotropic Sobolev spaces are introduced in any (modern) book about partial differential equations, e.g. [40, 53, 131], for anisotropic Sobolev spaces only few references exist, e.g. the books [15, 92]. Finally we discuss the regularity of elliptic, semi-elliptic and parabolic partial differential equations. An a priori regularity estimate for semi-elliptic partial differential equations is also proven in this section.

For more details about the analysis of elliptic and parabolic partial differential equations we refer to the text books [40, 53, 131]. For semi-elliptic boundary value problems, there are not so many references.

3.1. Domains

Definition 3.1 (Domain)**.** *[40, Definition in Appendix C.1]. Let $d \in \mathbb{N}$ be a finite number. An open and connected set $\Omega \subset \mathbb{R}^d$ is called domain. The boundary $\Gamma = \partial\Omega$ of a domain Ω is defined as $\overline{\Omega} \setminus \Omega$. The boundary is called $C^{0,1}$-boundary or Lipschitz boundary if for any point $x_0 \in \Gamma$ there exists a ball $U_\varepsilon(x_0)$ around x_0 and a $C^{0,1}$-function $\gamma : \mathbb{R}^{n-1} \to \mathbb{R}$ such that (after*

relabeling the coordinates and an affine transformation if necessary) we have

$$\Omega \cap U_\varepsilon(x_0) = \{x \in U_\varepsilon(x_0) : \ x_n > \gamma(x_1, \ldots, x_{n-1})\}$$

Definition 3.2 (Basic Notation)**.** *For the rest of this thesis let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ be a bounded Lipschitz domain with boundary $\Gamma = \partial\Omega$, which may be partitioned into Γ_1 and Γ_2 with $\Gamma_1 \cap \Gamma_2 = \emptyset$ and $\overline{\Gamma_1} \cup \overline{\Gamma_2} = \Gamma$. Further let T be a finite number, so that $(0, T)$ is finite time interval and we define the space time domain $Q = \Omega \times (0, T)$ with lateral boundary $\Sigma = \Gamma \times (0, T) = \partial\Omega \times (0, T)$, which may be partitioned into $\Sigma_1 = \Gamma_1 \times (0, T)$ and $\Sigma_2 = \Gamma_2 \times (0, T)$. This notation is very common in the analysis of parabolic partial differential equations.*

3.2. Basic results from functional analysis

We will deal with norms and operators on Hilbert spaces, therefore we repeat some important fact about these.

Definition 3.3 (Equivalent norms)**.** *[72, Definition 2.4-4] Let H be a real valued vector space. Two norms $\|\cdot\|_1$, $\|\cdot\|_2$ are called equivalent norms for the normed space H if there are two constants $c, C > 0$ so that*

$$c \|x\|_1 \leq \|x\|_2 \leq C \|x\|_1$$

holds for any $x \in H$.

Theorem 3.4 (Continuous and bounded linear operator)**.** *[72, Theorem 2.7-9] A linear operator $A : X \to Y$ is continuous iff the operator is bounded, i.e.*

$$\|Ax\|_Y \lesssim \|x\|_X , \qquad \forall x \in X.$$

Furthermore if the linear operator A is continuous in a single point, it is continuous everywhere.

Theorem 3.5 (Operator norm and dual space)**.** *[53, Exercise 6.1.8. and Section 6.3.1] The operator norm of the operator $A : X \to Y$ is defined by*

$$\|A\|_{L(X,Y)} = \sup_{\|x\|_X = 1} \|Ax\|_Y = \sup_{\|x\|_X \neq 0} \frac{\|Ax\|_Y}{\|x\|_X}.$$

The space $L(X, Y)$ of all continuous linear operators mapping from the normed space X to the Banach space Y with the operator norm is a Banach space.

The space $X^* = L(X, \mathbb{R})$ of all bounded, linear mappings onto $\mathbb{R}$ is called dual space to X and is Banach space with the operator norm.

Theorem 3.6 (Continuation of operators)**.** *[53, Theorem 6.1.11.]. Let X_0 be a dense subspace of the normed space X and let Y be a Banach space. For any continuous and linear operator $T_0 : X_0 \to Y$ there is a unique continuation T with the same operator norm, i.e.*

- *$T_0 x = Tx$ for all $x \in X_0$,*
- *For any sequence $x_n \to x$ with $x_n \in X_0$ and $x \in X$ holds $Tx = \lim_{n\to\infty} T_0 x_n$,*

- $\|T_0\|_{L(X_0,Y)} = \|T\|_{L(X,Y)}$.

Definition 3.7 (Adjoint or dual operator and positive definite operator)**.** *[72, Definition 9.4-1] We call the operator A^* the adjoint or dual operator of the operator $A : H \to H$ iff*

$$\langle Ax, y\rangle_{H\times H} = \langle x, A^*y\rangle_{H\times H}, \qquad \forall x, y \in H.$$

The operator A is called self adjoint if $A^* = A$.

A self adjoint operator $T : H \to H$ is called positive definite iff

$$\langle Tx, x\rangle_{H\times H} \geq 0, \qquad \forall x \in H.$$

For a positive definite, self adjoint Operator $T : H \to H$ we define the square root A of T by $A^2 = AA = T$. If A is a positive definite operator it is called positive square root and we write $A = T^{1/2}$.

Theorem 3.8 (Existence and uniqueness of square root of positive definite operators)**.** *[72, Theorem 9.4-2] Every positive definite bounded self adjoint linear operator $T : H \to H$ on a complex Hilbert space H has a unique positive square root A.*

Corollary 3.9. *If the operator A is the positive definite square root of a positive definite self adjoint linear operator $T : H \to H$ on a real Hilbert space H, then the operator A is self adjoint.*

Proof. By simple computation we have

$$\begin{aligned} \langle Tx, y\rangle_{H\times H} &= \langle AAx, y\rangle_{H\times H} = \langle Ax, A^*y\rangle_{H\times H} = \langle x, A^*A^*y\rangle_{H\times H} \\ \text{and } \langle Tx, y\rangle_{H\times H} &= \langle x, Ty\rangle_{H\times H} = \langle x, AAy\rangle_{H\times H}. \end{aligned}$$

Thus we have $AA = A^*A^* = T$. If we prove that the operator A^* is positive we have finished the proof as the positive square root of T is unique by the previous theorem.

But it is easy to see that the operator A^* is positive definite as $\langle A^*x, x\rangle_{H\times H} = \langle x, Ax\rangle_{H\times H} \geq 0$, where we have used that A is positive definite. □

3.3. Sobolev spaces

For the analysis of partial differential equations we will use function spaces. First we introduce the classical Sobolev spaces for functions defined on a spatial domain Ω. These spaces are defined as subspaces of the Lebesgue space $L^2(\Omega)$ by using a multi-index notation.

3.3.1. Classic Sobolev spaces $H^k(\Omega)$

Definition 3.10 (L^p-Spaces)**.** *[39, Definition B.4] The Lebesgue space $L^p(\Omega)$ for $p \in [1, \infty)$ is defined by*

$$L^p(\Omega) := \left\{ u : \Omega \to \mathbb{R} : \ u \text{ is Lebesgue measurable and } \|u\|^p_{L^p(\Omega)} = \int_\Omega |u|^p \,\mathrm{d}\,\omega < \infty \right\}$$

and for $p = \infty$ by

$$L^p(\Omega) := \left\{ u : \Omega \to \mathbb{R} : \ u \text{ is Lebesgue measurable and } \|u\|_{L^p(\Omega)} = \operatorname{esssup}_\Omega |u| < \infty \right\}.$$

Definition 3.11 (Multi-index)**.** *[36, Section 2] We call a vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^d$ multi-index and we define for the multi-index $\alpha \in \mathbb{N}_0^d$ the following operations*

$$\begin{aligned} |\alpha| &= \alpha_1 + \alpha_2 + \cdots + \alpha_d, \\ x^\alpha &= x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_d^{\alpha_d}, \\ D^\alpha &= \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial}{\partial x_2}\right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_d}\right)^{\alpha_d}. \end{aligned}$$

Theorem 3.12 (Sobolev spaces $H^k(\Omega)$)**.** *Let $\Omega \in \mathbb{R}^d$ be a Lipschitz domain. We recall the following results about weak derivatives and Sobolev spaces.*

- *[40, Section 5.2.1.] Given the multi-index $\alpha \in \mathbb{N}_0^d$. The weak derivative $D^\alpha u$ is defined as*

$$\int_\Omega D^\alpha u \varphi \,\mathrm{d}\,\omega = (-1)^{|\alpha|} \int_\Omega u D^\alpha \varphi \,\mathrm{d}\,\omega, \qquad \forall \varphi \in C_0^\infty(\Omega).$$

  For a function u the weak derivative $D^\alpha u$ is at most unique.

- *[53, Theorem 6.2.6.]. For $k \in \mathbb{N}$ the classical Sobolev space $H^k(\Omega)$ is defined by*

$$H^k(\Omega) := \left\{ u \in L^2(\Omega) : \ D^\alpha u \in L^2(\Omega) \ \forall \alpha \in \mathbb{N}_0^d : |\alpha| \leq k \right\}.$$

  The Sobolev space $H^k(\Omega)$ is a Hilbert space with scalar product

$$(u, v)_{H^k(\Omega)} = \sum_{\alpha \leq k} \int_\Omega D^\alpha u \, D^\alpha v \,\mathrm{d}\,\omega,$$

  and norm $|||u|||_{H^k(\Omega)} = \sqrt{(u,u)_{H^k(\Omega)}}$.

- *[119, Chapter IV.116, IV.118] If the domain Ω can be decomposed into finitely many subdomains which are star shaped with respect to a ball, then the norm*

$$\|u\|_{H^k(\Omega)}^2 = \sum_{\alpha=k} \int_\Omega |D^\alpha u|^2 \,\mathrm{d}\,\omega + \langle u, u \rangle_{L^2(\Omega) \times L^2(\Omega)},$$

  is an equivalent norm for the Hilbert space $H^k(\Omega)$.

- *[85, Theorem 4.5.1], [119, Theorem IV.115.3] If the domain Ω is star shaped with respect to a ball and the linear and continuous functionals $l_i(\cdot)$, $i = 1, \ldots, N$ do not vanish simultaneously for any non-zero polynomial of degree at most $k-1$, then the norm*

$$\|u\|_*^2 = \sum_{\alpha=k} \int_\Omega D^\alpha u \, D^\alpha u \,\mathrm{d}\,\omega + \sum_{i=1}^N |l_i(u)|,$$

  is an equivalent norm for the Hilbert space $H^k(\Omega)$.

- *[28, (1.2.3)] If the finite domain Ω has a Lipschitz boundary Γ, then there exists a unique bounded linear operator $T : H^1(\Omega) \to L^2(\Gamma)$ with*

$$Tu = u|_\Gamma \qquad \forall u \in C^\infty(\bar{\Omega}).$$

Tu is called the trace of u on $\partial\Omega$. We identify Tu with u. The boundedness of the operator results in the inequality

$$\|u\|_{L^2(\Gamma)} \lesssim \|u\|_{H^1(\Omega)}.$$

- *[53, Theorem 6.2.42.]. We introduce the space $H_0^k(\Omega)$ as completion of all smooth functions with compact support in Ω and its dual space as*

$$H_0^k(\Omega) = \overline{C_0^\infty(\Omega)}, \qquad H^{-k}(\Omega) = (H_0^k(\Omega))^*.$$

The space $H_0^k(\Omega)$ is a Hilbert space with the scalar product of $H^k(\Omega)$. For Lipschitz domains Ω we have the characterization

$$H_0^1(\Omega) = \left\{ v \in H^1(\Omega) : v|_{\partial\Omega} = 0 \right\}.$$

Remark 3.13. *If the domain Ω is finite we have $H_0^k(\Omega) \subset H^k(\Omega)$ and $H^k(\Omega) \neq H_0^k(\Omega)$ and therefore*

$$H^{-k}(\Omega) \neq \left(H^k(\Omega) \right)^*.$$

Definition 3.14. *[53, (6.4.1)] Given two Hilbert spaces V and H with $V \subseteq H$ and V is dense in H. We call the inclusion $V \subseteq H \subseteq V^*$ Gelfand triplet.*

Remark 3.15. *For parabolic partial differential equations the two Gelfand triplets*

$$H^k(\Omega) \subseteq L^2(\Omega) \cong (L^2(\Omega))^* \subseteq (H^k(\Omega))^*,$$
$$H_0^k(\Omega) \subseteq L^2(\Omega) \cong (L^2(\Omega))^* \subseteq H^{-k}(\Omega)$$

are of interest. The Gelfand triplet property is well known, see e.g. [53, (6.3.8a), (6.3.8b)]. We do not identify $H^k(\Omega)$ with its dual because for this identification we would need to use the $H^k(\Omega)$ inner product instead of the duality pairing, which coincides with the $L^2(\Omega)$ inner product for $L^2(\Omega)$ functions (see also [53, Paragraph with the superscription "Attention" between the Proof of Corollary 6.3.10. and Exercise 6.3.11]).

3.3.2. Sobolev spaces involving time

For time dependent partial differential equations it is usual to discuss functions with values in some Hilbert space.

Definition 3.16. *[40, Chapter 5.9.2.] The space $L^2(0,T;X)$ consists of all strongly measurable function $u : [0,T] \to X$ with*

$$\|u\|^2_{L^2(0,T;\Omega)} = \int_0^T \|u\|_X^2 \,\mathrm{d}\, t < \infty,$$

and the space $\mathcal{C}(0,T;X)$ consists of all continuous functions $u : [0,T] \to X$ with

$$\|u\|_{\mathcal{C}(0,T;X)} = \max_{0 \le t \le T} \|u(t)\|_X < \infty.$$

Just as we have defined Sobolev spaces based on Lebesgue spaces with values in $\mathbb{R}$, we now define Sobolev spaces based on Lebesgue spaces with values in X.

Definition 3.17. *[40, Chapter 5.9.2.] We say $v \in L^2(0,T;X)$ is the weak derivative of $u \in L^2(0,T;X)$, iff*

$$\int_0^T \phi'(t)u(t)\,\mathrm{d}\,t = -\int_0^T \phi(t)v(t)\,\mathrm{d}\,t \qquad \forall \phi \in \mathcal{C}_0^\infty(0,T;\mathbb{R}).$$

The Sobolev space $H^l(0,T;X)$ is defined as the subspace of all $L^2(0,T;X)$-functions with first derivative in the space $L^2(0,T;X)$. Its norm is

$$\|u\|^2_{H^l(0,T;X)} = \|u\|^2_{L^2(0,T;X)} + \sum_{k=1}^{l} \left\| \frac{\mathrm{d}^k}{\mathrm{d}\,t^k} u \right\|^2_{L^2(0,T;X)}.$$

Next we discuss in which function space the functions can be identified with a continuous function

Theorem 3.18. *[40, Theorems 5.9.2. and 5.9.3] Assume, that $u \in H^1(0,T;X)$ holds. Then we have also $u \in \mathcal{C}([0,T];X)$ and the integral representation*

$$u(t) = u(s) + \int_s^t u'(\tau)\,\mathrm{d}\,\tau.$$

Assume that $u \in L^2(0,T;H_0^1(\Omega))$ with $u' \in L^2(0,T;H^{-1}(\Omega))$, then we have also

$$u \in \mathcal{C}([0,T];L^2(\Omega)).$$

For the time derivative of the norm holds

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\|u\|^2_{L^2(\Omega)} = 2\langle u', u\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}.$$

Remark 3.19. *For the Gelfand triplet $V \subseteq H \subseteq V^*$ with $H = L^2(\Omega)$ and $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$ we introduce*

$$\mathcal{Y} = H^1(0,T;V^*) \cap L^2(0,T;V), \tag{3.1}$$
$$\mathcal{P} = L^2(0,T;V). \tag{3.2}$$

We choose $V = \{v \in H^1(\Omega) : \ v|_{\Gamma_1} = 0\}$ and $W = H^2(\Omega) \cap V$.

Remark 3.20. *[131] and [62]. In literature the function space $\mathcal{Y}$ is sometimes denoted by $W(0,T)$ or $W(0,T;H,V)$.*

3.3.3. Sobolev spaces with mixed order of differentiation

Another view to the Sobolev spaces involving time, is the discussion of spaces with functions, which are differentiable with different order in different direction.

Theorem 3.21 (Anisotropic Sobolev spaces)**.** *[79, Chapter 2.1] The Sobolev-spaces $H^{\alpha}(Q)$ with respect to the pair $\alpha = (r, s) \in \mathbb{N}^2$ defined as*

$$H^{(r,s)}(Q) = L^2(0, T; H^r(\Omega)) \cap H^s(0, T; L^2(\Omega))$$

is an Hilbert space with the norm

$$|||y|||^2_{H^{(r,s)}(Q)} = \|y(t)\|^2_{L^2(0,T;H^r(\Omega))} + \|y\|^2_{H^s(0,T;L^2(\Omega))}\,.$$

Remark 3.22. *The Sobolev space $H^{(2,1)}(Q)$ was introduced by several authors. The definition above can be found in Lions and Magenes [79] or Ladyzhenskaya, Solonnikov and Ural'ceva [74]. They assume that all lower order derivatives have to be in $L^2(Q)$. The definition of Nikol'skiĭ [92] and Triebel [127] is slightly different. They assume only that the function itself and its highest order derivative in every direction are $L^2(Q)$-functions. But Nikol'skiĭ also shows which lower order derivatives can be estimated by the norm of u and the norm of the highest order derivatives. We repeat a special case of this theorem and collect some equivalent norms of $H^{(2,1)}(Q)$.*

Theorem 3.23. *Assume that $Q = \Omega \times (0, 1)$ is a $d+1$ dimensional rectangle. Let $r^\star, s \in \mathbb{N}$. If*

$$y,\ D^{(r,0)}y,\ D^{(0,s)}y \in L^2(Q)$$

for all $r \in \mathbb{N}_0^d$ with $|r| = r^\star$ and the multi-index $l = (l_1, l_2) = (l_{1,i}, \ldots, l_{1,i}, l_2) \in \mathbb{N}_0^{d+1}$ fulfills

$$1 - \sum_{i=1}^{d} \frac{l_{1,i}}{r^\star} - \frac{l_2}{s} \geq 0,$$

then the derivative $D^l y$ is a $L^2(Q)$ function and can be estimated by

$$\|D^l y\|_{L^2(Q)} \lesssim \|y\|_{H^{(r^\star,s)}(Q)}.$$

Proof. This is the anisotropic version of a similar result for the isotropic case which can be found in the book of Smirnow [119, Comment in Section IV.112.] (see also Theorem 3.12). The isotropic case is proven for domains that are star shaped with respect to a non-empty ball in [119, Section IV.116.]. In [119, Section IV.118.] this is generalized in for domains which can be decomposed in finitely many subdomains which are star shaped with respect to a non-empty sphere.

For $Q = \mathbb{R}^n$ the anisotropic version of the result of Theorem 3.23 can be found in the book of Nikol'skiĭ [92, Theorem 9.2.2.].

For a class of special finite domains (including rectangular domains) a more general result, which includes the result of Theorem 3.23 as special case, is proven in [15, Theorem 13.6.1.]. □

So we are prepared to discuss equivalent norms of the Sobolev space $H^{(2,1)}(Q)$.

Theorem 3.24. *For a one-dimensional domain $\Omega = (a,b)$ and $Q = \Omega \times (0,T)$ the norms*

$$\begin{aligned} |||y|||^2_{H^{(2,1)}(Q)} &= |y|^2_{H^{(2,1)}(Q)} + 2\,\|y\|^2_{L^2(Q)} + \|y_x\|^2_{L^2(Q)}\,, \\ \|y\|^2_{H^{(2,1)}(Q)} &= |y|^2_{H^{(2,1)}(Q)} + \|y\|^2_{L^2(Q)}\,, \end{aligned}$$

where the semi-norms are defined by

$$\begin{aligned} |y|^2_{H^{(r,s)}(Q)} &= |y|^2_{H^{(r,0)}(Q)} + |y|^2_{H^{(0,s)}(Q)}\,, \\ |y|^2_{H^{(r,0)}(Q)} &= \iint\limits_Q \left|D^{(r,0)}y\right|^2 \mathrm{d}\,x\,\mathrm{d}\,t, \qquad |y|^2_{H^{(0,s)}(Q)} = \iint\limits_Q \left|D^{(0,s)}y\right|^2 \mathrm{d}\,x\,\mathrm{d}\,t \end{aligned}$$

are equivalent norms for the space $H^{(2,1)}(Q)$.

Proof. The inequality $\|y\|^2_{H^{(2,1)}(Q)} \le |||y|||^2_{H^{(2,1)}(Q)}$ is clear. On the other hand if $y, y_{xx} \in L^2(Q)$ it follows by Theorem 3.23 that

$$\|y_x\|_{L^2(Q)} \lesssim \|y\|_{L^2(Q)} + \|y_{xx}\|_{L^2(Q)}$$

which proves $|||y|||_{H^{(2,1)}(Q)} \lesssim \|y\|_{H^{(2,1)}(Q)}$. □

Theorem 3.25. *Assume that there exist D linear and bounded functionals l_i, $i = 1,\dots,D$ in $H^{(r,s)}(Q)$ with $D = s \cdot \binom{d+r-1}{r-1}$ which do not vanish at the same time for any non-zero polynomial of degree at most $r-1$ in the spatial dimensions and at most $s-1$ in t. Then the norm*

$$\|y\|_* = |y|_{H^{(r,s)}(Q)} + \sum_{i=1}^{D} |l_i(y)|$$

is an equivalent norm for $H^{(r,s)}(Q)$.

Remark 3.26. *The dimension of the space of all polynomials of degree at most $r-1$ in the spatial variables and degree at most $s-1$ in the temporal variable is given by $D = s\cdot\binom{d+r-1}{r-1}$. According to Kunz [73, Example A.12a)] the dimension of all monomials in d variables of degree k is given by $\binom{d+k-1}{d-1}$. The definition of the binomial coefficient gives $\binom{d+k-1}{d-1} = \binom{d+k-1}{k}$. For the dimension of all polynomials of degree at most $r-1$ summing up yields $\sum_{k=0}^{r-1}\binom{d+k-1}{k} = \binom{d+r-1}{r-1}$. The identity $\sum_{k=0}^{m}\binom{n+k}{n} = \binom{n+m+1}{n+1}$ can be easily shown with induction, using the well known recursion formula $\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$.*

Proof of Theorem 3.25. This is the anisotropic version of a norm equivalence presented in Theorem 3.12 and was proven in [6] for $d = 1$. For convenience of the reader we repeat the proof here.

As recommended in [93, Proof of Theorem 1] we follow the ideas of the proof of the isotropic case, which can be found e.g. in [85, Theorem 4.5.1] and [119, Theorem IV.114.3.]. The inequality

$$\|y\|_* \lesssim |||y|||_{H^{(r,s)}(Q)}$$

is clear by the definition of $\|\cdot\|_*$ as the functionals l_i are bounded in $H^{(r,s)}(Q)$.

We prove the inequality

$$|||y|||_{H^{(r,s)}(Q)} \lesssim \|y\|_* \tag{3.3}$$

by contradiction. Therefore we assume, that there is a sequence $\{v_n\}_{n=0}^{\infty} \in H^{(r,s)}(Q)$ with

$$|||v_n|||_{H^{(r,s)}(Q)} > n \|v_n\|_* . \tag{3.4}$$

Obviously we have $v_n \neq 0$ and without loss of generality we can assume that the members of this sequence are normed, i.e. $|||v_n|||_{H^{(r,s)}(Q)} = 1$. The sequence v_n is bounded in $H^{(r,s)}(Q)$ and therefore compact in $L^2(Q)$, as $H^{(r,s)}(Q) \hookrightarrow H^1(Q) \hookrightarrow L^2(Q)$ and the embedding $H^1(Q) \hookrightarrow L^2(Q)$ is compact. So there is a convergent subsequence with limit v. We denote this subsequence again by v_n. Therefore we have

$$\|v_n - v\|_{L^2(Q)} \to 0.$$

With the assumption (3.4) we have

$$\|v_n\|_* < \frac{1}{n} |||v_n|||_{H^{(r,s)}(Q)}$$

and therefore $\|v_n\|_* \to 0$. By the definition of the norm $\|\cdot\|_*$ this also implies the convergence $|v_n|_{H^{(r,s)}(Q)} \to 0$. Therefore we have $D^{(r,0)}v = 0$ and $D^{(0,s)}v = 0$ in the sense of $L^2(Q)$. This implies also $D^{(r+n,0)}v = 0$ and $D^{(0,s+n)}v = 0$ for all $n \in \mathbb{N}$. By choosing n large enough and the Sobolev embedding theorem the function v is a (r,s)-times continuously differentiable function.

This implies that the limit v is a polynomial of degree $s-1$ in t and degree $r-1$ in x. As the function v is the limit of the sequence $\{v_n\}$ it follows that $\|v\|_* = 0$ and therefore

$$\sum_{i=1}^{D} |l_i(v)| = 0.$$

As the functionals l_i do not vanish for any non-zero polynomial of degree $s-1$ in t and degree $r-1$ in x, this implies $v \equiv 0$ which is a contradiction to the assumption $\|v_n\|_{H^{(2,1)}(Q)} = 1$. □

Furthermore we introduce Sobolev spaces with respect to a multi-index set.

Definition 3.27. *Let the set $\mathcal{A}$ be a finite set of multi-indices, then we define the Sobolev space*

$$H^{\mathcal{A}}(Q) = \left\{ u \in L^2(Q) : \ D^{\alpha}u \in L^2(Q), \ \forall \alpha \in \mathcal{A} \right\}.$$

The connection between the Sobolev space $H^{(r,s)}(Q)$ and the space $H^{\mathcal{A}}(Q)$ is given (implicitly) by the Theorem 3.23.

3.4. Partial differential equations

3.4.1. Elliptic equations: Boundary value problems

We consider the boundary value problem

$$\left.\begin{aligned} Ay &= u \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \Gamma_1, \\ \frac{\partial y}{\partial n_A} &= 0 \quad \text{on } \Gamma_2, \end{aligned}\right\} \tag{3.5}$$

where the operator A is a second order elliptic operator in divergence form, i.e.

$$Ay = -\nabla \cdot \tilde{A}(x)\nabla y + b(x) \cdot \nabla y + c(x)y, \tag{3.6}$$

with a symmetric positive definite matrix $\tilde{A}(x) \in L^\infty(\Omega; \mathbb{R}^{d\times d})$, $b \in L^\infty(\Omega; \mathbb{R}^d)$ with $\nabla \cdot b \in L^\infty(\Omega)$ and $c \in L^\infty(\Omega)$. The conormal derivative is defined by

$$\frac{\partial y}{\partial n_A} = \tilde{A}\nabla y \cdot \vec{n} \tag{3.7}$$

with the outer normal vector $\vec{n}$ on $\partial\Omega$. Instead of the solution of the boundary value problem (3.5) we discuss the corresponding weak or variational formulation given by

$$a(y, \varphi) = \langle u, \varphi \rangle_{V^* \times V} \qquad \forall \varphi \in V, \tag{3.8}$$

with

$$a(y, \varphi) = \int_\Omega \tilde{A}\nabla y \nabla \varphi + b \cdot \nabla y \varphi + cy\varphi \,\mathrm{d}\,\omega, \tag{3.9}$$

$$V = \left\{ v \in H^1(\Omega) : \ y|_{\Gamma_1} = 0 \right\}. \tag{3.10}$$

It is well known that a unique solution of the weak formulation exists under mild assumptions.

Theorem 3.28 (Lax-Milgram Lemma)**.** *[28, Theorem 1.1.3.] Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a continuous and V-elliptic bilinear form, i.e.*

$$\|y\|_V^2 \lesssim a(y, y) \qquad \forall y \in V, \tag{3.11}$$

$$|a(y, \varphi)| \lesssim \|y\|_V \, \|\varphi\|_V \qquad \forall y, \varphi \in V, \tag{3.12}$$

and let $u \in V^*$. Then the variational problem of finding a function $y \in V$ with

$$a(y, \varphi) = \langle u, \varphi \rangle_{V^* \times V} \qquad \forall \varphi \in V$$

has one and only one solution.

Remark 3.29. *1. The condition* (3.11) *is called V-ellipticity of the bilinear form and the condition* (3.12) *is the definition of the continuity of a bilinear form.*

2. *The V-ellipticity for the bilinear form $a(\cdot,\cdot)$ of model problem* (3.8) *with the Hilbert space V as in* (3.10) *can be shown with the standard assumptions that the matrix $\tilde{A}(x)$ is uniformly elliptic*

$$z^T \tilde{A}(x) z \gtrsim \|z\|^2_{\mathbb{R}^d}, \qquad \forall z \in \mathbb{R}^d, \text{ almost everywhere in } \Omega,$$

and the coefficients c and b fulfill the condition

$$c - \frac{1}{2}\nabla \cdot b \geq 0 \qquad \text{almost everywhere in } \Omega.$$

These conditions on the coefficients are standard assumptions (see e.g. [39, 48]).

The Lax-Milgram Lemma, Theorem 3.28, yields the existence of a unique solution $y \in V$.

But we will also be interested in more regular solutions, which are contained in the subspace $W \subset V$ with $W = H^2(\Omega) \cap V$. So we recall the classic regularity result for the Poisson equation.

Theorem 3.30 (Regularity)**.** *[47, Theorems 3.2.1.2 and 3.2.1.3] Let $u \in L^2(\Omega)$ and the domain Ω be a convex polygonal ($d = 2$) or convex polyhedral ($d = 3$) bounded domain. Let for the coefficients additionally $\tilde{A} \in \mathcal{C}^{0,1}(\bar{\Omega}; R^{d\times d})$, with*

$$z^T \tilde{A}(x) z \gtrsim \|z\|^2_{\mathbb{R}^d}, \qquad \forall z \in \mathbb{R}^d, \forall x \in \Omega,$$

and $b = 0$ hold. For the cases

1. *$c = 0$ and $V = H^1_0(\Omega)$*
2. *$c \in \mathbb{R}$ with $c > 0$ and $V = H^1(\Omega)$*

the unique solution $y \in V$ of

$$a(y, \varphi) = \int_\Omega u\varphi \,\mathrm{d}\,\omega \qquad \forall \varphi \in V$$

fulfills $y \in W = H^2(\Omega)$.

3.4.2. Semi-elliptic equations: Boundary value problems

After the discussion of the basic solution properties of elliptic equations we discuss a certain class of semi-elliptic partial differential equations with elliptic bilinear form. For a self adjoint second order $H^1(\Omega)$-elliptic operator A we discuss the model problem

$$\left.\begin{aligned} -y_{tt} + A^2 y + \frac{1}{\nu} y &= f \quad \text{in } Q, \\ y &= 0 \quad \text{on } \Sigma_1, \\ Ay &= 0 \quad \text{on } \Sigma_1, \\ \frac{\partial}{\partial n_A} y &= 0 \quad \text{on } \Sigma_2, \\ \frac{\partial}{\partial n_A} Ay &= 0 \quad \text{on } \Sigma_2, \\ y(x,0) &= 0 \quad \text{in } \Omega \times \{0\}, \\ y_t(x,T) + Ay(x,T) &= 0 \quad \text{in } \Omega \times \{T\}, \end{aligned}\right\} \tag{3.13}$$

with the constant $\nu \in \mathbb{R}$, $\nu > 0$ and where the derivative $\frac{\partial}{\partial n_A} y$ is the conormal derivative is defined as in (3.7).

Theorem 3.31. *For the variational formulation of the model equation* (3.13) *and a self-adjoint second order $H^1(\Omega)$-elliptic operator A given by*

$$\left.\begin{aligned} a(y,\varphi) &= (f,\varphi) \qquad \forall \varphi \in V, \\ a(y,\varphi) &= \iint_Q y_t\varphi_t + AyA\varphi + \frac{1}{\nu} y\varphi \,\mathrm{d}\, x \,\mathrm{d}\, t + \int_\Omega y_x(x,T)\varphi_x(x,T) \,\mathrm{d}\, x, \\ (f,\varphi) &= \iint_Q \frac{1}{\nu} y_d \varphi \,\mathrm{d}\, x \,\mathrm{d}\, t, \\ V &= \left\{ v \in H^{(2,1)}(Q) : \ v(x,0) = 0, \ v = 0 \text{ on } \Sigma_1, \ \frac{\partial}{\partial n_A} v = 0 \text{ on } \Sigma_2 \right\}, \end{aligned}\right\} \tag{3.14}$$

there exists a unique solution $y \in V$ for $y_d \in V^*$.

Proof. We prove this Theorem for the case $\nu = 1$. The modifications for arbitrary $\nu \in \mathbb{R}^+$ are obvious. The existence of a unique solution follows with the Lax-Milgram Lemma, if we can prove the V-ellipticity and continuity of the bilinear form $a(\cdot,\cdot)$. The V-ellipticity follows directly as

$$\|y\|_{H^{(2,1)}(Q)} \lesssim a(y,y) - \int_\Omega (y_x(x,T))^2 \,\mathrm{d}\, x \le a(y,y).$$

For the continuity we use the Cauchy-Schwarz inequality

$$a(y,\varphi) \le \|y\|_{H^{(2,1)}(Q)} \|\varphi\|_{H^{(2,1)}(Q)} + \|y_x(x,T)\|_{L^2(\Omega)} \|\varphi_x(x,T)\|_{L^2(\Omega)}.$$

As $H^{(2,1)}(Q) \hookrightarrow C([0,T]; H^1(\Omega))$ (see e.g. [30, (XVIII.1.61.iii)]) we have

$$\|y_x\|_{L^2(\Omega)} \le \|y\|_{H^1(\Omega)} \le \|y\|_{C([0,T],H^1(\Omega))} \lesssim \|y\|_{H^{(2,1)}(Q)}.$$

With this estimate we have proven the continuity of the bilinear form $a(\cdot,\cdot)$, and therefore the existence of the unique solution y follows by the Lax-Milgram Lemma. □

Now we provide an a priori estimate for semi-elliptic equations, which is needed for the proof of a $L^2(Q)$-error estimate with the Aubin-Nitsche trick.

Theorem 3.32. *If $f \in L^2(Q)$ and A is a self adjoint operator, then the solution y of the boundary value problem* (3.13) *fulfills the estimate*

$$\|y\|^2_{L^2(D(A^2))} + \|y\|^2_{H^1(D(A))} + \|y\|^2_{H^2(L^2(\Omega))} \lesssim \|f\|^2_{L^2(Q)}.$$

Remark 3.33. *If the domain is smooth, we have $D(A) = H^2(\Omega)$ and $D(A^2) = H^4(\Omega)$ and therefore in this case the estimate of the theorem is*

$$\|y\|_{H^{(4,2)(Q)}} \lesssim \|f\|_{L^2(Q)}.$$

This is an analogue of Theorem 3.30 for semi elliptic boundary value problems.

Proof of Theorem 3.32. We introduce the set $\{\varphi_k\}_{k=1}^{\infty}$ of orthonormal eigenfunctions of the operator A with the corresponding eigenvalues λ_k^2, which also fulfill the boundary conditions

$$\varphi_k = 0, \qquad \text{on } \Gamma_1, \qquad \frac{\partial}{\partial n}\varphi_k = 0, \qquad \text{on } \Gamma_2.$$

It is well known that the orthonormal eigenfunctions of a self-adjoint elliptic operator form an orthonormal basis of $L^2(\Omega)$ [40, Theorem 6.5.1.]. By the definition of the eigenfunctions we have $A\varphi_k = \lambda_k^2\varphi_k$ and therefore the boundary conditions

$$A\varphi_k = 0, \qquad \text{on } \Gamma_1, \qquad \frac{\partial}{\partial n}A\varphi_k = 0, \qquad \text{on } \Gamma_2.$$

are also fulfilled. So we write the solution of the equation as eigenfunction expansion

$$y = \sum_{k=1}^{\infty} y_k(t)\varphi_k$$

with time-dependent coefficients $y_k(t)$. When we insert this representation into the differential equation, this yields

$$-y_{k,tt} + \left(\lambda_k^4 + \frac{1}{\nu}\right) y_k = f_k \tag{3.15}$$

for every k with the (time-dependent) Fourier coefficients $f_k = \int_\Omega f\varphi_k \,\mathrm{d}\,\omega$ of the right hand side and initial and terminal conditions

$$\begin{aligned} y_k(0) &= 0, \\ y_{k,t}(T) + \lambda_k^2 y_k(T) &= 0. \end{aligned}$$

The weak form of this problem for every y_k is

$$\begin{aligned} \int_0^T f_k z \,\mathrm{d}\,t = \int_0^T y_{kt} z_t + \left(\lambda_k^4 + \frac{1}{\nu}\right) y_k z \,\mathrm{d}\,t + \lambda_k^2 y_k(T) z(T) =: a_k(y_k, z), \\ \forall z \in H^1(0,T) : z(0) = 0. \end{aligned}$$

If we use y_k as test function and the Cauchy-Schwarz inequality we have the estimate

$$\begin{aligned} a_k(y_k, y_k) &= \|y_{k,t}\|_{L^2(Q)}^2 + \left(\lambda_k^4 + \frac{1}{\nu}\right)\|y_k\|_{L^2(Q)}^2 + \lambda_k^2 y_k^2(T) = \int_0^T f_k y_k \,\mathrm{d}\,t \\ &\le \|f_k\|_{L^2(Q)} \|y_k\|_{L^2(Q)} . \end{aligned} \tag{3.16}$$

This yields directly

$$\left(\lambda_k^4 + \frac{1}{\nu}\right)\|y_k\|_{L^2(Q)} \lesssim \|f_k\|_{L^2(\Omega)} . \tag{3.17}$$

With (3.16) and (3.17) we can also estimate

$$\|y_{k,t}\|_{L^2(Q)}^2 \le \|f_k\|_{L^2(Q)} \|y_k\|_{L^2(Q)} \le \|f_k\|_{L^2(Q)}^2 \frac{1}{\lambda_k^4 + \frac{1}{\nu}}.$$

Taking the square root gives

$$\lambda_k^2 \|y_{k,t}\|_{L^2(Q)} \lesssim \|f_k\|_{L^2(Q)} .$$

Further we have an estimate for $y_{k,tt}$ with (3.15), the triangle inequality and (3.17)

$$\begin{aligned} \|y_{k,tt}\|_{L^2(Q)} &\leq \|f_k\|_{L^2(Q)} + \left(\lambda_k^4 + \frac{1}{\nu}\right) \|y_k\|_{L^2(Q)} \\ &\lesssim \|f_k\|_{L^2(Q)} . \end{aligned}$$

Altogether the estimate

$$\|y_{k,tt}\|_{L^2(Q)}^2 + \lambda_k^4 \|y_{k,t}\|_{L^2(Q)}^2 + \lambda_k^8 \|y_k\|_{L^2(Q)}^2 \lesssim \|f_k\|_{L^2(Q)}^2$$

is established.

Summing up over k implies $y \in H^2(0,T;L^2(\Omega)) \cap H^1(0,T;D(A)) \cap L^2(0,T;D(A^2))$ and the bound

$$\|y_{tt}\|_{L^2(Q)}^2 + \|Ay_t\|_{L^2(Q)}^2 + \left\|A^2 y\right\|_{L^2(Q)}^2 \lesssim \|f\|_{L^2(Q)}^2 ,$$

which is the desired estimate. □

3.4.3. Parabolic equations: Initial boundary value problems

After the discussion of partial differential equations with V-elliptic bilinear form, we now consider the parabolic initial boundary value problem

$$\left.\begin{aligned} y_t + Ay &= u \quad \text{in } Q, \\ y &= v \quad \text{on } \Omega \times \{0\}, \\ y &= 0 \quad \text{on } \Sigma_1, \\ \frac{\partial y}{\partial n_A} &= 0 \quad \text{on } \Sigma_2, \end{aligned}\right\} \tag{3.18}$$

where the operator A is a partial differential operator of second order. We assume that the spatial operator A has a representation in the form of equation (3.6) and is elliptic in the sense that all eigenvalues of coefficient matrix $\tilde{A}$ of the leading part are positive but we do not need that the operator A is V-elliptic. As indicated above, we discuss this initial boundary value problem in the space $\mathcal{Y}$ with the Gelfand triplet $V \subset H \subset V^*$, where we choose $V = \{v \in H^1(\Omega) : \ v|_{\Gamma_1} = 0\}$ and $H = L^2(\Omega)$.

As in Section 3.4.1 we associate with the differential operator A the bilinear form $a(\cdot,\cdot)$. We assume that the bilinear form is continuous and fulfills Gårding's inequality, i.e.

$$a(u,\varphi) \leq C \|u\|_V \|\varphi\|_V , \qquad \forall u, \varphi \in V, \tag{3.19}$$

$$c\|u\|_V^2 \leq a(u,u) + k\|u\|_H^2 , \qquad \text{with some } 0 \leq k < \infty \text{ and } \forall u \in V. \tag{3.20}$$

In contrast to the elliptic problems, we need only to assume Gårding's inequality with some $k \in \mathbb{R}$ but not the stronger V-ellipticity, which is Gårding's inequality with $k = 0$. Due to the temporal component of parabolic problems there are properties, which are fulfilled without loss of generality.

Theorem 3.34. *Consider the parabolic initial boundary value problem* (3.18)*. We can assume without loss of generality, that*

1. *if Gårding's inequality* (3.20) *holds for a parabolic initial boundary value problem for some* $k > 0$*, Gårding's inequality* (3.20) *holds also for* $k = 0$*,*
2. *if the operator* A *is given as in* (3.6) *then the condition* $c - \frac{1}{2}\nabla \cdot b \geq 0$ *is fulfilled and*
3. *we have homogeneous initial conditions* $y(\cdot, 0) = 0$*.*

Proof. The proof of the first point of this Theorem can be found in the proof of [131, Theorem 26.1]. For convenience of the reader we repeat these arguments here. Let y be the solution of the problem (3.18). The function $w = y \cdot \mathrm{e}^{-\lambda t}$ with $\lambda \in \mathbb{R}$ and with the derivatives

$$\begin{aligned} \frac{\partial w}{\partial t} &= y_t \cdot \mathrm{e}^{-\lambda t} - \lambda y \cdot \mathrm{e}^{-\lambda t}, \\ Aw &= \mathrm{e}^{-\lambda t} Ay, \end{aligned}$$

fulfills the initial boundary value problem

$$\left.\begin{aligned} w_t + Aw + \lambda w = \mathrm{e}^{-\lambda t} u \quad & \text{in } Q, \\ w = v \quad & \text{on } \Omega \times \{0\}, \\ w = 0 \quad & \text{on } \Sigma_1, \\ \frac{\partial w}{\partial n_A} = 0 \quad & \text{on } \Sigma_2. \end{aligned}\right\} \tag{3.21}$$

The bilinear form $a(u, v) + \lambda \langle u, v \rangle_{H \times H}$ associated with the initial boundary value problem (3.21) is continuous and, for λ large enough, fulfills Gårding's inequality with $k = 0$. The second point follows just in the same way by choosing λ large enough.

For the proof of the last property we split the solution $y(x, t)$ into

$$y(x, t) = v(x) + \tilde{y}(x, t).$$

The function $\tilde{y}(x, t)$ solves the problem

$$\begin{aligned} \tilde{y}_t + A\tilde{y} = u - Av \quad & \text{in } Q, \\ \tilde{y} = 0 \quad & \text{on } \Omega \times \{0\}, \\ \tilde{y} = v \quad & \text{on } \Sigma_1, \\ \frac{\partial \tilde{y}}{\partial n_A} = \frac{\partial \tilde{v}}{\partial n_A} \quad & \text{on } \Sigma_2, \end{aligned}$$

so it is clear, that we can assume $y(\cdot, 0)v = 0$ in the original problem if we solve for $\tilde{y}$ and allow an inhomogeneous right hand side of the equation. □

Next we define the weak solution of the problem (3.18).

Definition 3.35 (Weak solution for parabolic partial differential equations I)**.** *[62, Definition 1.27] A weak solution $y \in \mathcal{Y}$ of* (3.18) *is the solution of the variational problem*

$$\left.\begin{aligned} \langle y_t, \varphi\rangle_{V^*\times V} + a(y,\varphi) &= \langle u, \varphi\rangle_{H\times H} \quad \forall \varphi \in V, \\ \langle y(\cdot,0), \varphi\rangle_{H\times H} &= \langle v, \varphi\rangle_{H\times H} \quad \forall \varphi \in H. \end{aligned}\right\} \tag{3.22}$$

This is not the only possibility of the definition of weak solutions of parabolic initial boundary value problems. An alternative definition is the following.

Definition 3.36 (Weak solution for parabolic partial differential equations II)**.** *[62, Definition 1.28] A weak solution $y \in \mathcal{Y}$ of* (3.18) *is the solution of the variational problem*

$$\left.\begin{aligned} B(y,\varphi) &= \int_0^T \langle u, \varphi\rangle_{H\times H}\,\mathrm{d}\,t && \forall \varphi \in \mathcal{P}, \\ \langle y(\cdot,0), \varphi\rangle_{H\times H} &= \langle v, \varphi\rangle_{H\times H} && \forall \varphi \in H, \\ B(y,\varphi) &= \int_0^T \langle y_t, \varphi\rangle_{V^*\times V} + a(y,\varphi)\,\mathrm{d}\,t. \end{aligned}\right\} \tag{3.23}$$

with space $\mathcal{P}$ defined in (3.2)

Next we see, that the both definitions are equivalent and that there exists a unique weak solution to the problem (3.18).

Theorem 3.37 (Existence and uniqueness for parabolic partial differential equations)**.** *Let T finite and Ω a domain with Lipschitz boundary and let the inequalities* (3.19) *and* (3.20) *hold. then*

- *[62, Theorem 1.33] the definitions* (3.22) *and* (3.23) *of a weak solution of* (3.18) *are equivalent and*
- *[131, Theorem 26.1.] the problem* (3.18) *has a unique weak solution $y \in \mathcal{Y}$.*

Finally we give the improved regularity results for parabolic problems. For such results we need compatibility conditions which we introduce in the following assumption.

Assumption 3.38. *We define the regularity assumption up to order m as*

$$\left.\begin{aligned} g_0 &= v && \in V, \\ g_1 &= u(\cdot,0) - Ag_0 && \in V, \\ &\vdots && \vdots \\ g_m &= \frac{\mathrm{d}^{m-1}}{\mathrm{d}\,t^{m-1}} u(\cdot,0) - Ag_{m-1} && \in V. \end{aligned}\right\} \tag{CA_m}$$

With this conditions we give two regularity results, one for only the temporal regularity and one which also gives higher spatial regularity.

Theorem 3.39 (Abstract regularity result)**.** *[131, Theorem 27.2 and Theorem 27.3] Let for the initial boundary value problem* (3.18) *hold the conditions of Gårding's inequality and continuity* (3.19) *and* (3.20)*. Let further the right hand side*

$$u \in H^k(0,T;V^*),$$

the compatibility assumptions (CA_m) *up to order* m *be fulfilled and additionally*

$$\frac{\mathrm{d}^m}{\mathrm{d}\,t^m}u(\cdot,0) - Ag_m \in H$$

hold. Then the solution y *has the improved regularity with respect to* t*, namely*

$$y \in H^{m+1}(0,T;V), \qquad \frac{\mathrm{d}^{m+2}}{\mathrm{d}\,t^{m+2}}y \in L^2(0,T;V^*).$$

Theorem 3.40 (Spatial regularity)**.** *[40, Theorem 7.5 and Theorem 7.6]. Consider a smooth domain* Ω *with* C^∞*-boundary. Let the operator* A *in the initial boundary value problem* (3.18) *be a symmetric operator, let Gårding's inequality* (3.20) *and the condition of continuity* (3.19) *hold and let the Neumann boundary* Σ_2 *be empty. Let the regularity assumptions*

$$\begin{aligned} v &\in H^{2m+1}(\Omega), \\ \frac{\mathrm{d}^k\,u}{\mathrm{d}\,t^k} &\in L^2(0,T;H^{2m-2k}(\Omega)) \qquad \text{for } k = 0,\dots,m \end{aligned}$$

and the compatibility assumptions (CA_m) *up to order* m *hold. Then the solution* y *of the initial boundary value problem has the following improved regularity with respect to* t *and* x

$$\frac{\mathrm{d}^k\,y}{\mathrm{d}\,t^k} \in L^2(0,T;H^{2m+2-2k}(\Omega)) \qquad \text{for } k = 0,\dots m+1,$$

Remark 3.41. *For homogeneous Dirichlet boundary conditions the proof of Theorem 3.40 in [40] gives additionally* $\frac{\mathrm{d}^m}{\mathrm{d}\,t^m}y \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$*.*

4. Numerical analysis for differential equations

Contents

In this chapter we deal with the numerical analysis of partial differential equations and Hamiltonian systems. These results are of general purpose and not restricted to the numerical analysis for optimal control problems, therefore we discuss them in a self-contained chapter.

First we discuss the finite element discretization of partial differential equations with elliptic bilinear form. In particular we prove an a priori estimate for the finite element error for some $H^{(2,1)}(Q)$-elliptic equations. As finite elements we use Hermite-Lagrange tensor product elements.

For parabolic equations we discuss the Crank-Nicolson discretization for the case that the right hand side is an approximation of the exact right hand side. This is a focus in our analysis that we consider the case that the right hand side of the equation is also approximated and not evaluated exactly.

Finally we recall some results of the numerical analysis of Hamiltonian systems.

For the numerical analysis we need the space of polynomials which we define now.

Definition 4.1 (Polynomial spaces)**.** *The space of all real valued polynomials of degree less or equal n is defined as*

$$\mathbb{P}_n = span\left\{t^0,\ t^1, \ldots,\ t^n\right\},$$

the space of all polynomials of degree less or equal n on the interval (a,b) with values in some Banach or Hilbert space V is given as

$$\mathbb{P}_n\left((a,b),V\right) = \left\{\sum_{i=0}^{n}(t-a)^i\varphi_i,\ \text{with } t\in(a,b) \text{ and } \varphi_i\in V \text{ for } i=1,\ldots,n\right\},$$

and in the same way we we define the space

$$\mathbb{P}_n\left(\Omega,V\right),$$

of all polynomials of degree less or equal n on the domain Ω with values in V.

4.1. Partial differential equations with V-elliptic bilinear form

4.1.1. General results

We start with the numerical analysis for V-elliptic bilinear forms and introduce the Galerkin approximation of this equation.

Definition 4.2 ((Conforming) Galerkin method)**.** *The solution $y \in V$ of the variational problem fulfills*

$$a(y, \varphi) = \langle f, \varphi \rangle_{H \times H} \qquad \forall \varphi \in V. \tag{4.1}$$

For the (conforming) Galerkin approximation of this problem we search a solution $y_h \in V_h \subset V$ such that

$$a(y_h, \varphi) = \langle f, \varphi \rangle_{H \times H} \qquad \forall \varphi \in V_h, \tag{4.2}$$

where the subspace V_h is finite dimensional.

Theorem 4.3 (Céa's Lemma)**.** *[28, Theorem 2.4.1. and Remark 2.4.1.] Let $a(\cdot, \cdot)$ be a V-elliptic continuous bilinear form and let the functions y and y_h be the solutions of the variational problem* (4.1) *and* (4.2)*. Then the estimate*

$$\|y - y_h\|_V \leq C \inf_{v_h \in V_h} \|y - v_h\|_V ,$$

holds with a constant C which is independent of the subspace V_h.

Remark 4.4. *1. So we have bounded the error of the Galerkin approximation by the error of the best approximation in the space V_h.*

2. In the numerical realization of the variational problem (4.2) *we need to evaluate the integrals $\langle f, \varphi \rangle_{H \times H}$ for a basis of the space V_h exactly. As this is not possible in general, a common remedy is the use of a quadrature rule for the evaluation of this integral. This is equivalent to replace the function f by an appropriately chosen interpolant f_h. Therefore one solves the problem*

$$a(y_h, \varphi) = \langle f_h, \varphi \rangle_{H \times H} \qquad \forall \varphi \in V_h. \tag{4.3}$$

As the test and the ansatz functions on the left hand side are known, one can choose a numerical integration scheme, for which the application of the numerical integration and exact integration of the terms of the bilinear form yields the same result, at least for constant coefficients. For variable coefficients the application of a quadrature rule may lead to an approximating bilinear form, so that one solves the problem

$$a_h(y_h, \varphi) = \langle f_h, \varphi \rangle_{H \times H} \qquad \forall \varphi \in V_h. \tag{4.4}$$

The additional errors which are introduced due to the numerical integration is discussed in the following Theorem.

Theorem 4.5 (First Strang Lemma)**.** *[28, Theorem 4.1.1] Let y_h be the solution of* (4.4) *and let the function y be the solution of the equation* (4.1)*. Assume that the bilinear form of* (4.4) *is uniformly elliptic, i.e. the constant in the ellipticity inequality* (3.8) *can be chosen independent of the discretization parameter h of the family of subspaces V_h. Then the estimate*

$$\|y - y_h\|_V \lesssim \inf_{v_h \in V_h} \left\{ \|y - v_h\|_V + \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_V} \right\} + \sup_{w_h \in V_h} \frac{|\langle f - f_h, w_h \rangle_{H \times H}|}{\|w_h\|_V}$$

holds.

Remark 4.6.

1. *We will focus on the case for which $a_h(v_h, w_h) = a(v_h, w_h)$ for $v_h, w_h \in V_h$ is fulfilled. This condition can easily established in the case of constant coefficients.*
2. *In the estimate are two expressions with a supremum. These expressions are operator norms and therefore the norm of the corresponding operators in V_h^*.*
3. *The first term of the estimate is the best approximation error as in the Céa Lemma. Therefore we see that the approximation f_h of the right hand side f should have at least the same order as the best approximation error.*

Until now we have discussed error estimates in the norm of the Hilbert space V. Now we establish an error estimate in the norm of the pivot space H of the Gelfand triplet.

Theorem 4.7 (Aubin Nitsche Trick)**.** *Let the functions y and y_h be the solutions of the variational problem* (4.1) *and* (4.3)*. Then we have for the H-norm the estimate*

$$\|y - y_h\|_H \lesssim \sup_{\|g\|_H = 1} \inf_{z_h \in V_h} \left\{ \|y - y_h\|_V \, \|z_g - z_h\|_V + |\langle f - f_h, z_h \rangle_{H \times H}| \right\} \tag{4.5}$$

where the function z_g is defined as the solution of the variational problem

$$a(\varphi, z_g) = \langle g, \varphi \rangle_{H \times H}, \qquad \forall \varphi \in V. \tag{4.6}$$

Proof. We transfer the ideas of the proof of [28, Theorem 3.2.4], where the Aubin-Nitsche trick is proved for the case that one uses exact integration of the right hand side $\langle f, \varphi_h \rangle_{H \times H}$, to the case that the right hand side is approximated by $\langle f_h, \varphi_h \rangle_{H \times H}$. This is also part of [28, Exercise 4.1.3.].

With the variational problem (4.1) and its numerical Galerkin approximation (4.3) we have

$$a(y - y_h, \varphi_h) = \langle f - f_h, \varphi_h \rangle_{H \times H} \qquad \forall \varphi_h \in V. \tag{4.7}$$

If the numerical approximation f_h coincides with f and the integration is performed exact, this relation is called Galerkin orthogonality.

Using this relation and the difference $y - y_h$ as test function for the dual problem (4.6) yields for any function $z_h \in V_h$

$$\begin{aligned} \langle g, y - y_h \rangle_{H \times H} &= a(y - y_h, z_g) = a(y - y_h, z_g - z_h) + \langle f - f_h, z_h \rangle_{H \times H} \\ &\lesssim \|y - y_h\|_V \, \|z_g - z_h\|_V + \langle f - f_h, z_h \rangle_{H \times H}. \end{aligned}$$

As this inequality holds for any z_h in V_h we can take the infimum on the right hand side. Due to the identification $H \cong H^*$ we can estimate the norm of H by the operator norm

$$\|y - y_h\|_H \lesssim \|y - y_h\|_{H^*} = \sup_{\|g\|_H=1} \langle g, y - y_h\rangle_{H\times H^*}$$

and the proof of this theorem is finished. □

The goal of the abstract error estimate (4.5) is to prove that the error in the norm of the space H has a better convergence rate then the error in the norm of the space V. This is done under general assumptions in the following theorem.

Theorem 4.8 (Error Estimate based on the Aubin Nitsche Trick)**.** *For a general error estimate we need the following three assumptions.*

1. *There is a subspace $W \subseteq H$, such that for any $f \in H$ the solution y of the variational problem* (4.1) *is in the space $V \cap W$ and the a priori estimate*

$$\|y\|_W \lesssim \|f\|_H \tag{4.8}$$

   holds with a constant C which does not depend on the choice of the right hand side f.

2. *The a priori estimate carries over to the solution of the dual problem* (4.6)*, so that*

$$\|z_g\|_W \lesssim \|g\|_H . \tag{4.9}$$

3. *For any function $z_g \in W$ exists an approximation $z_h \in V_h$ with*

$$\|z_g - z_h\|_V \lesssim h^s \|z_g\|_W . \tag{4.10}$$

If these assumptions are satisfied, the error between the solution y of the problem (4.1) *and the solution y_h of the problem* (4.3) *is bounded by*

$$\|y - y_h\|_H \lesssim h^s (\|y - y_h\|_V + \|f - f_h\|_H) + \|f - f_h\|_{W^*} . \tag{4.11}$$

Proof. For the proof of this Theorem we bound the terms of the estimate (4.5) of the abstract Aubin-Nitsche-Theorem 4.7.

For the first term we use the approximation result in W

$$\|y - y_h\|_V \|z_g - z_h\|_V \lesssim \|y - y_h\|_V h^s \|z_g\|_W ,$$

the a priori estimate for the dual problem and the norm of the right hand side of the dual problem $\|g\|_H = 1$

$$\|y - y_h\|_V h^s \|z_g\|_W \lesssim h^s \|y - y_h\|_V \|g\|_H \lesssim h^s \|y - y_h\|_V .$$

For the second term we add a zero and use the fact, that the scalar product of the space H and the primal-dual pairing $W^* \times W$ coincide to get

$$\begin{aligned}\langle f - f_h, z_h\rangle_{H\times H} &= \langle f - f_h, z_h - z_g\rangle_{H\times H} + \langle f - f_h, z_g\rangle_{H\times H} \\ &\lesssim \|f - f_h\|_H \|z_g - z_h\|_H + \|f - f_h\|_{W^*} \|z_g\|_W .\end{aligned}$$

For the first term of this expression we repeat the application of the estimate for the approximation, the a apriori estimate for the dual problem and the norm of the right hand side of the dual problem

$$\|f - f_h\|_H \, \|z_g - z_h\|_H \lesssim \|f - f_h\|_H \, h^s \, \|z_g\|_W \lesssim h^s \, \|f - f_h\|_H \, \|g\|_H \lesssim h^s \, \|f - f_h\|_H \, .$$

Finally we apply the stability estimate for the dual problem

$$\|f - f_h\|_{W^*} \, \|z_g\|_W \lesssim \|f - f_h\|_{W^*} \, \|g\|_H \lesssim \|f - f_h\|_{W^*} \, .$$

Altogether we have proven the estimate of the theorem. □

Finite element error estimates for second order elliptic partial differential equations are well known (see e.g. [28]) and shortly sketched in Appendix C. In the following section we focus on finite element error estimates for semi-elliptic partial differential equations with V-elliptic bilinear forms.

4.1.2. Semi-elliptic partial differential equations

Discretization and error estimates

After the recapitulation of general results, we apply the ideas now to the semi-elliptic model problem (3.13), which was given by

$$\left.\begin{aligned} -y_{tt} + A^2 y + \frac{1}{\nu} y &= f \quad \text{in } Q, \\ y &= 0 \quad \text{on } \Sigma_1, \\ Ay &= 0 \quad \text{on } \Sigma_1, \\ \frac{\partial}{\partial n} y &= 0 \quad \text{on } \Sigma_2, \\ \frac{\partial}{\partial n} Ay &= 0 \quad \text{on } \Sigma_2, \\ y(x,0) &= 0 \quad \text{in } \Omega \times \{0\}, \\ y_t(x,T) + Ay(x,T) &= 0 \quad \text{in } \Omega \times \{T\}, \end{aligned}\right\} \tag{4.12}$$

and the corresponding variational problem

$$\left.\begin{aligned} a(y,\varphi) &= \langle f, \varphi \rangle_{L^2(Q) \times L^2(Q)} \qquad \forall \varphi \in V, \\ a(y,\varphi) &= \int_0^T \int_\Omega y_t \varphi_t + AyA\varphi + \frac{1}{\nu} y\varphi \, \mathrm{d}\,\omega \, \mathrm{d}\,t - \int_\Omega Ay(x,T)\varphi(x,T) \, \mathrm{d}\,t, \end{aligned}\right\} \tag{4.13}$$

with the space $V = \left\{ v \in H^{(2,1)}(Q) : \; v|_{\Sigma_1} = 0, \; \frac{\partial}{\partial n} v\big|_{\Sigma_2} = 0 \right\}$. As seen in Theorem 3.31 the bilinear form is V-elliptic and continuous.

For the discretization we restrict us to the case, where the variable x is one dimensional. We use a finite element method with a finite element mesh $\mathfrak{T}_{h,\tau}$ with rectangular elements θ and a tensor product ansatz with a linear, quadratic or cubic Lagrange ansatz in the temporal dimension and a continuously differentiable cubic Hermite ansatz in the spatial dimension

for which the discretization parameters h and τ can be chosen independently. We define the interpolation operator on the finite element mesh with the nodes (x_i, t_j) by

$$\begin{aligned} I_{h\tau}^k : H^{(3,2)}(Q) &\to \mathcal{C}^1(0,X) \otimes \mathcal{C}^0(0,T), \\ I_{h\tau}^k w\Big|_\theta &\in \mathbb{P}^3 \otimes \mathbb{P}^k, \\ I_{h\tau}^k w\left(x_i, t_j + \frac{m}{k}\tau\right) &= w\left(x_i, t_j + \frac{m}{k}\tau\right) && \text{for } m = 0, \cdots, k, \\ D^{(1,0)} I_{h\tau}^k w\left(x_i, t_j + \frac{m}{k}\tau\right) &= D^{(1,0)} w\left(x_i, t_j + \frac{m}{k}\tau\right) && \text{for } m = 0, \cdots, k. \end{aligned}$$

The finite element approximation $y_{h\tau} \in V_{h,\tau}$ is the solution of

$$a(y_{h\tau}, \varphi) = \langle f_{h\tau}, \varphi \rangle_{L^2(Q) \times L^2(Q)} \qquad \forall \varphi \in V_{h\tau}, \tag{4.14}$$

where the approximation of the right hand side is given by $f_{h\tau} = I_{h\tau} f$ and the function space $V_{h\tau}$ is defined as

$$V_{h\tau} = \left\{ v \in V : \ v \in \mathcal{C}^1(0,X) \otimes \mathcal{C}^0(0,T), \ v|_\theta \in \mathbb{P}^3 \otimes \mathbb{P}^k, \ \forall \theta \in \mathcal{T}_{h\tau} \right\}.$$

The interpolation error is estimated in the following Theorem.

Theorem 4.9 (Interpolation error estimate)**.** *For a function $y \in H^A(Q) \cap H^{(3,2)}(Q)$ with the multi-index set $A = \{(0,0), (0,k+1), (i,1), (4,0), (2,j)\}$ with $i \in \{1, \ldots, 4\}$ and $j \in \{1, \ldots, k\}$ the interpolation error can be estimated by*

$$\begin{aligned} \left\| y - I_{h\tau}^k y \right\|_{H^{(2,1)}(Q)} &\lesssim h^i \left\| D^{(i,1)} y \right\|_{L^2(Q)} + \tau^k \left\| D^{(0,k+1)} y \right\|_{L^2(Q)} \\ &\quad + \tau^j \left\| D^{(2,j)} y \right\|_{L^2(Q)} + h^2 \left\| D^{(4,0)} y \right\|_{L^2(Q)}, \\ \left\| y - I_{h\tau}^k y \right\|_{L^2(Q)} &\lesssim \tau^{k+1} \left\| D^{(0,k+1)} y \right\|_{L^2(Q)} + h^4 \left\| D^{(4,0)} y \right\|_{L^2(Q)}. \end{aligned}$$

We use this interpolation error estimate for the estimates of the finite element error. As the proof is technical and rather long, we give the proof later.

Before, we apply it to the discretization error estimates for the finite element solution $y_{h\tau}$.

Theorem 4.10 (Error estimate in the energy norm)**.** *If for the exact solution of the variational problem $y \in H^A(Q) \cap H^{(3,2)}(Q)$ with the multi-index set $A = \{(0,0), (0,k+1), (i,1), (4,0), (2,j)\}$ with $i \in \{1, \ldots, 4\}$ and $j \in \{1, \ldots, k\}$ holds, the approximation error for finite element solution $y_{h\tau}$ with an ansatz of polynomial degree k in time can be bounded by*

$$\begin{aligned} \| y - y_{h\tau} \|_{H^{(2,1)}(Q)} &\lesssim h^i \left\| D^{(i,1)} y \right\|_{L^2(Q)} + \tau^k \left\| D^{(0,k+1)} y \right\|_{L^2(Q)} \\ &\quad + \tau^j \left\| D^{(2,j)} y \right\|_{L^2(Q)} + h^2 \left\| D^{(4,0)} y \right\|_{L^2(Q)} \\ &\quad + \sup_{w_h \in V_{h\tau}} \frac{\left| \langle f - f_{h\tau}, w_h \rangle_{L^2(Q) \times L^2(Q)} \right|}{\| w_h \|_{H^{(2,1)}(Q)}}. \end{aligned}$$

Proof. As the bilinear form $a(\cdot,\cdot)$ is V-elliptic, $V_{h\tau} \subseteq V$ and the functions y and $y_{h\tau}$ are the solutions of the variational problems (4.13) and (4.14), we can apply the first Strang Lemma Theorem 4.5 to get

$$\|y - y_{h\tau}\|_{H^{(2,1)}(Q)} \lesssim \inf_{v_h \in V_h} \|y - v_h\|_{H^{(2,1)}(Q)} + \sup_{w_h \in V_{h\tau}} \frac{\left|\langle f - f_{h\tau}, w_h\rangle_{L^2(Q)\times L^2(Q)}\right|}{\|w_h\|_{H^{(2,1)}(Q)}}.$$

Therefore, for the error in the energy norm, we need to bound the best approximation errors in the first Strang Lemma. To bound the best approximation errors we can use the interpolation error estimate of the Theorem 4.9 and the proof is done.

Note that for the approximation of the second term we need only to estimate the interpolation error in the $L^2(Q)$-norm. □

Remark 4.11. *The last term in the error estimate of Theorem 4.10 can be estimated with the $L^2(Q)$-interpolation error estimate of Theorem 4.9 as*

$$\begin{aligned}\sup_{w_h \in V_{h\tau}} \frac{\left|\langle f - f_{h\tau}, w_h\rangle_{L^2(Q)\times L^2(Q)}\right|}{\|w_h\|_{H^{(2,1)}(Q)}} &\leq \sup_{w \in V} \frac{\left|\langle f - f_{h\tau}, w\rangle_{L^2(Q)\times L^2(Q)}\right|}{\|w\|_{H^{(2,1)}(Q)}} \\ &\leq \|f - f_{h\tau}\|_{L^2(Q)} \lesssim \tau^{k+1} \left\|D^{(0,k+1)} y\right\|_{L^2(Q)} + h^4 \left\|D^{(4,0)} y\right\|_{L^2(Q)}\end{aligned}$$

for $f \in H^{(4,k+1)}(Q)$. If the right hand side f is less regular but the products of the right hand side with test functions can be integrated exactly, the use of Céa's Lemma, Theorem 4.3, instead of the first Strang Lemma is a remedy.

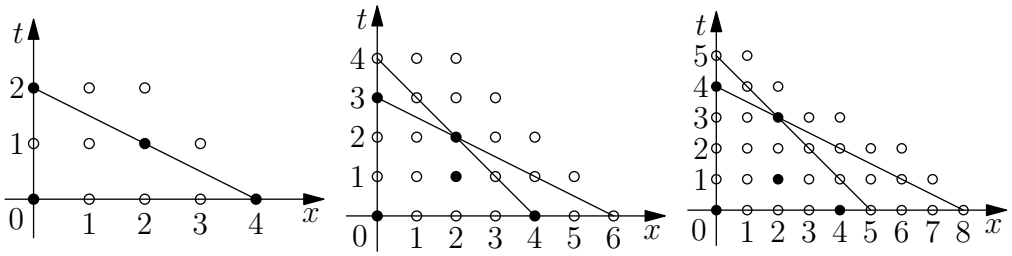
Remark 4.12. *In Theorem 4.10 the regularity assumption is given in terms of Sobolev spaces with respect to a multi-index set. We discuss now, for which Sobolev spaces $H^{(r,s)}(Q)$ the regularity assumptions are fulfilled in the most interesting case $i = 2$ and $j = k$, in which the Theorem provides an error estimate of order 2 with respect to the spatial discretization and of order k with respect to the temporal discretization. Our tool for this discussion is Theorem 3.23. In Figure 4.1 we have illustrated, which mixed derivatives are bounded for certain Sobolev spaces $H^{(r,s)}(Q)$:*

1. *For $k = 1$ the multi-index set is $A = \{(0,0),(0,2),(4,0),(2,1)\}$. These derivatives exists for functions in the space $H^{(4,2)}(Q)$.*
2. *For $k = 2$ the multi-index set is $A = \{0,0),(0,3),(4,0),(2,1),(2,2)\}$. These derivatives exist for functions in the spaces $H^{(4,4)}(Q)$ or $H^{(6,3)}(Q)$.*
3. *For $k = 3$ the multi-index set is $A = \{0,0),(0,4),(4,0),(2,1),(2,3)\}$. These derivatives exist for functions in the spaces $H^{(5,5)}(Q)$ or $H^{(8,4)}(Q)$.*

Before we prepare the proof of Theorem 4.10 with some Lemmas, we also give a $L^2(Q)$-error estimate.

Theorem 4.13 ($L^2(Q)$-error estimate with the Aubin-Nitsche trick)**.** *For a solution y, which fulfills the regularity assumptions of the previous Theorem 4.10 for $i = 2$ and $j = k$ and a right hand side $f \in H^A(Q)$, the error in the $L^2(Q)$-norm can be estimated by*

$$\|y - y_{h\tau}\|_{L^2(Q)} \lesssim (h^2 + \tau^k)(h^2 + \tau) \|y\|_{H^A(Q)} + \|f - f_{h\tau}\|_{H^{(4,2)^*}(Q)}. \tag{4.15}$$



(a) Multi-indices for the temporal discretization with polynomials of degree $k = 1$.

(b) Multi-indices for the temporal discretization with polynomials of degree $k = 2$.

(c) Multi-indices for the temporal discretization with polynomials of degree $k = 3$.

Figure 4.1.: With these figures we illustrate the multi-indices which are needed for the estimates in Theorem 4.10 for the case $i = 2$ and $j = k$ in black. For better overview additional multi-indices are added as circles. For the Sobolev space $H^{(r,s)}(Q)$ all the derivatives corresponding to the multi-indices below the line, which connects $(0, s)$ and $(r, 0)$ are $L^2(Q)$ functions, according to Theorem 3.23.

Proof. For the proof of this Lemma the proof of Lemma 4.8 can directly be transferred. For the spaces we set $H = L^2(Q)$, $V = H^{(2,1)}$ and $W = H^A$. The a priori estimates (4.8) and (4.9) were proven in Theorem 3.32. The approximation estimate (4.10) in the form

$$\|z_g - z_h\|_{H^{(2,1)}(Q)} \lesssim \left(h^2 + \tau\right) \|z_g\|_{H^A(Q)}$$

is the result of Theorem 4.10.

Therefore the Aubin Nitsche Theorem 4.8 yields

$$\|y - y_{h\tau}\|_{L^2(Q)} \lesssim \left(h^2 + \tau\right) \left(\|y - y_{h\tau}\|_{H^{(2,1)}(Q)} + \|f - f_{h\tau}\|_{L^2(Q)}\right) + \|f - f_{h\tau}\|_{H^{(4,2)^*}(Q)}.$$

The application of the energy error and interpolation error estimates given in Theorem 4.9 and Theorem 4.10 yield the result. □

Remark 4.14. *The term*

$$\|f - f_{h\tau}\|_{H^{(4,2)^*}(Q)}$$

in the $L^2(Q)$-error estimate can be treated as in Remark 4.11.

Remark 4.15. *The error estimates in Theorems 4.10 and 4.13 imply the following choice of discretization parameters:*

1. *For the linear ansatz in time the error estimates of Theorems 4.10 and 4.13 imply a choice of $\tau \sim h^2$ for balancing the discretization error in the energy-norm and the $L^2(Q)$-norm.*

   With this choice of discretization parameters the error in the energy norm behaves asymptotically like $\mathcal{O}(h^2)$ and the error in the $L^2(Q)$-norm behaves asymptotically like $\mathcal{O}(h^4)$. With respect to the number of unknowns N the errors are of order $N^{-2/3}$ and $N^{-4/3}$, respectively.

2. *For the quadratic ansatz in time the error estimate of Theorem 4.10 implies a choice of $\tau \sim h$ for balancing the discretization error in the energy-norm.*

   This choice of the discretization parameters leads to an asymptotic error behavior of $\mathcal{O}(h^2)$ in the energy norm and $\mathcal{O}(\tau h^2 + \tau^3 + \tau^2 h^2 + h^4) \sim \mathcal{O}(h^3)$ in the $L^2(Q)$-norm. With respect to the number of unknowns N the errors are of order N^{-1} and $N^{-3/2}$.

3. *For the quadratic ansatz in time the error estimate of Theorem 4.13 implies at least a choice of $\tau \sim h^2$ to get an error estimate of order h^4 in the $L^2(Q)$-norm. So the asymptotic error is like $\mathcal{O}(h^2)$ in the energy norm and $\mathcal{O}(h^4)$ in the $L^2(Q)$-norm. With respect to the number of unknowns N the errors are of order $N^{-2/3}$ and $N^{-4/3}$, i.e. worse in comparison with the choice $\tau \sim h$.*

4. *For the cubic ansatz in time the error estimates of Theorem 4.10 implies a choice of $\tau \sim h^{2/3}$ for second order convergence in the energy norm. This choice of the discretization parameters leads to an asymptotic error behavior of $\mathcal{O}(h^2)$ in the energy norm and $\mathcal{O}(h^{8/3})$ in the $L^2(Q)$-norm. With respect to the number of unknowns N the errors are of order $N^{-6/5}$ and $N^{-24/15}$.*

Interpolation error estimate

We split the proof of the Theorem 4.9 into three lemmas. We will prove an estimate on the reference element $R = (0,1)^2$ and get the convergence order by transformation to the world element.

Lemma 4.16. *Let $y \in H^A(Q) \cap H^{(3,2)}(Q)$ with the multi-index set $A = \{(0, k+1),\ (i,1)\}$ with $i \in \{1, \dots, 4\}$. Then the time derivative of the interpolation error on one element θ can be bounded by*

$$\left\| D^{(0,1)} \left(y - I^k_{h\tau} y \right) \right\|_{L^2(\theta)} \lesssim \tau^k \left\| D^{(0,k+1)} y \right\|_{L^2(\theta)} + h^i \left\| D^{(i,1)} y \right\|_{L^2(\theta)}.$$

Proof. For the proof we use the standard transfer to the reference element $R = (0,1)^2$ and follow the ideas of [99, Section 2.1] On R, we denote all quantities by $\hat{\cdot}$. We start with

$$\left\| D^{(0,1)} \left(y - I^k_{h\tau} y \right) \right\|^2_{L^2(\theta)} = \int_R \frac{\tau h}{\tau^2} \left(\hat{D}^{(0,1)} \left(\hat{y} - \hat{I}^k_{h\tau} \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega}.$$

Next we introduce the temporal interpolation

$$\begin{aligned} \hat{I}^k_\tau : H^{(3,2)}(R) &\to H^3\left((0,1)\right) \otimes \mathcal{C}^0(0,T), \\ \hat{I}^k_\tau \hat{y} &\in H^3\left((0,1)\right) \otimes \mathbb{P}^k, \\ \hat{I}^k_\tau \hat{y} \left(\hat{x}, \frac{m}{k} \right) &= \hat{y} \left(\hat{x}, \frac{m}{k} \right), \qquad \text{for } m = 0, \dots, k, \end{aligned}$$

that is well-defined for almost all $\hat{x} \in (0,1)$. By adding and subtracting this function and the triangle inequality we have to estimate

$$\begin{aligned} \left\| D^{(0,1)} \left(y - I^k_{h\tau} y \right) \right\|_{L^2(\theta)} \leq{}& \sqrt{\frac{h}{\tau}} \left(\int_R \left(\hat{D}^{(0,1)} \left(\hat{y} - \hat{I}^k_\tau \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega} \right)^{1/2} \\ &+ \sqrt{\frac{h}{\tau}} \left(\int_R \left(\hat{D}^{(0,1)} \left(\hat{I}^k_\tau \hat{y} - \hat{I}^k_{h\tau} \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega} \right)^{1/2} \end{aligned} \tag{4.16}$$

For some fixed $\hat{x}^* \in (0,1)$ we can use the standard one dimensional interpolation result

$$\int_0^1 \left(\hat{D}^{(0,1)}\left(\hat{y} - \hat{I}_\tau^k \hat{y}\right)(\hat{x}^*, \hat{t})\right)^2 \mathrm{d}\,\hat{t} \lesssim \int_0^1 \left(\hat{D}^{(0,k+1)} \hat{y}\,(\hat{x}^*, \hat{t})\right)^2 \mathrm{d}\,\hat{t},$$

which yields

$$\int_R \left(\hat{D}^{(0,1)}\left(\hat{y} - \hat{I}_\tau^k \hat{y}\right)\right)^2 \mathrm{d}\,\hat{\omega} \lesssim \int_R \left(\hat{D}^{(0,k+1)} \hat{y}\right)^2 \mathrm{d}\,\hat{\omega}.$$

The other integral in the estimate (4.16) is also an one-dimensional interpolation error as $\hat{I}_{h\tau}^k \hat{y}$ is an interpolant of $\hat{I}_\tau^k \hat{y}$. The application of the standard one dimensional interpolation result yields for $i = 1, \ldots, 4$ the estimate

$$\int_R \left(\hat{D}^{(0,1)}\left(\hat{I}_\tau^k \hat{y} - \hat{I}_{h\tau}^k \hat{y}\right)\right)^2 \mathrm{d}\,\hat{\omega} \lesssim \int_R \left(\hat{D}^{(i,1)} \hat{I}_\tau^k \hat{y}\right)^2 \mathrm{d}\,\hat{\omega}.$$

To end the proof of this lemma we need finally to prove the estimate

$$\int_R \left(\hat{D}^{(i,1)} \hat{I}_\tau^k \hat{y}\right)^2 \mathrm{d}\,\hat{\omega} \lesssim \int_R \left(\hat{D}^{(i,1)} \hat{y}\right)^2 \mathrm{d}\,\hat{\omega}. \tag{4.17}$$

With the nodal Lagrangian interpolation basis $\varphi_i(t) \in \mathbb{P}^k$, $i = 0, \ldots, k$ with

$$\varphi_i\left(\frac{i}{k}\right) = \delta_{ik}$$

the action of the temporal interpolation operator $\hat{I}_\tau^k$ can be described by

$$\hat{I}_\tau^k \hat{w}(\hat{x}, \hat{t}) = \sum_{i=0}^{k} \hat{w}(\hat{x}, \hat{t}_i) \varphi_i(\hat{t}).$$

With the basis $\chi_i = \sum_{j=0}^{i} \varphi_j$ (see also [2, Section 5]) the interpolation can be written as

$$\begin{aligned}
\hat{I}_\tau^k \hat{w} &= \sum_{i=0}^{k-1} \left(\hat{w}(\hat{x}, \hat{t}_i) - \hat{w}(\hat{x}, \hat{t}_{i+1})\right) \chi_i(\hat{t}) + \hat{w}(\hat{x}, \hat{t}_k) \\
&= -\sum_{i=0}^{k-1} \left(\int_{\hat{t}_i}^{\hat{t}_{i+1}} \hat{D}^{(0,1)} \hat{w}(\hat{x}, \hat{s})\,\mathrm{d}\,\hat{s}\right) \chi_i(\hat{t}) + \hat{w}(\hat{x}, \hat{t}_k).
\end{aligned}$$

Therefore the first derivative of the interpolant is given by

$$\hat{D}^{(0,1)} \hat{I}_\tau^k \hat{w} = -\sum_{i=0}^{k-1} \left(\int_{\hat{t}_i}^{\hat{t}_{i+1}} \hat{D}^{(0,1)} \hat{w}\,\mathrm{d}\,\hat{s}\right) \chi_i'(\hat{t})$$

and the $L^2(R)$-norm of this derivative can be estimated as

$$\begin{aligned}
\left\|\hat{D}^{(0,1)} \hat{I}_\tau^k \hat{w}\right\|_{L^2(R)} &\leq \sum_{i=0}^{k-1} \left\|\hat{D}^{(0,1)} \hat{w}\right\|_{L^1(\hat{t}_i, \hat{t}_{i+1}; L^2(0,1))} \left\|\chi_i'(\hat{t})\right\|_{L^2((0,1))} \\
&\lesssim \left\|\hat{D}^{(0,1)} \hat{w}\right\|_{L^1(0,1; L^2(0,1))} \lesssim \left\|\hat{D}^{(0,1)} \hat{w}\right\|_{L^2(R)}.
\end{aligned}$$

Choosing $\hat{w} = \hat{D}^{(i,0)}\hat{y}$ yields the estimate (4.17). Altogether we have proven the estimate

$$\left\| D^{(0,1)} \left(y - I_{h\tau}^k y \right) \right\|_{L^2(\theta)}^2 \leq \frac{\tau h}{\tau^2} \int_R \left(\hat{D}^{(0,k+1)} \hat{y} \right)^2 \mathrm{d}\,\hat{\omega} + \frac{\tau h}{\tau^2} \int_R \left(\hat{D}^{(i,1)} \hat{y} \right)^2 \mathrm{d}\,\hat{\omega}.$$

Transferring the integrals back on the element θ yields the result. □

Lemma 4.17. *Assume that $y \in H^A(Q) \cap H^{(3,2)}(Q)$ with the multi-index set $A = \{(4,0),(2,j)\}$ with $j \in \{1,\dots,k\}$. Then the second spatial derivative of the interpolation error on the element θ can be bounded by*

$$\left\| D^{(2,0)} \left(y - I_{h\tau}^k y \right) \right\|_{L^2(\theta)} \lesssim \tau^j \left\| D^{(2,j)} y \right\|_{L^2(\theta)} + h^2 \left\| D^{(4,0)} y \right\|_{L^2(\theta)}.$$

Proof. As in the proof of the previous lemma we follow the ideas of [99, Section 2.1] and transfer the integral onto the reference element, where we denote quantities on the reference element by $\hat{\cdot}$. This yields

$$\left\| D^{(2,0)} \left(y - I_{h\tau}^k y \right) \right\|_{L^2(\theta)}^2 = \int_R \frac{\tau h}{h^4} \left(\hat{D}^{(2,0)} \left(\hat{y} - \hat{I}_{h\tau}^k \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega}.$$

Next we introduce the spatial interpolation

$$\begin{aligned} \hat{I}_h : H^{(3,2)}(R) &\to \mathcal{C}^1\left((0,1)\right) \otimes H^2(0,T), \\ \hat{I}_h \hat{y} &\in \mathbb{P}^3 \otimes H^2\left((0,T)\right), \\ D^{(i,0)} \hat{I}_h \hat{y}\left(m,\hat{t}\right) &= D^{(i,0)} \hat{y}\left(m,\hat{t}\right), \qquad \text{for } i = 0,1 \text{ and } m = 0,1. \end{aligned}$$

By adding and subtracting this interpolant and the triangle inequality we split the integral into

$$\begin{aligned} \left\| \hat{D}^{(2,0)} \left(y - I_{h\tau}^k y \right) \right\|_{L^2(\theta)}^2 \lesssim{}& \frac{\tau h}{h^4} \int_R \left(\hat{D}^{(2,0)} \left(\hat{y} - \hat{I}_h \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega} \\ &+ \frac{\tau h}{h^4} \int_R \left(\hat{D}^{(2,0)} \left(\hat{I}_h \hat{y} - \hat{I}_{h\tau}^k \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega}. \end{aligned} \tag{4.18}$$

As in the previous lemma the first integral can be estimated as an one dimensional interpolation error, which yields

$$\int_R \left(\hat{D}^{(2,0)} \left(\hat{y} - \hat{I}_h \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega} \lesssim \int_R \left(\hat{D}^{(4,0)} \hat{y} \right)^2 \mathrm{d}\,\hat{\omega}.$$

Again the other integral in the estimate (4.18) is also an one-dimensional interpolation error as $\hat{I}_{h\tau}^k \hat{y}$ is an interpolant of $\hat{I}_h \hat{y}$. The application of the standard one dimensional interpolation result yields with $j = 1,\dots,k$ the estimate

$$\int_R \left(\hat{D}^{(2,0)} \left(\hat{I}_h \hat{y} - \hat{I}_{h\tau}^k \hat{y} \right) \right)^2 \mathrm{d}\,\hat{\omega} \lesssim \int_R \left(\hat{D}^{(2,j)} \hat{I}_h \hat{y} \right)^2 \mathrm{d}\,\hat{\omega}.$$

To end the proof of this lemma we need finally to prove the estimate

$$\int_R \left(\hat{D}^{(2,j)} \hat{I}_h \hat{y} \right)^2 \mathrm{d}\,\hat{\omega} \lesssim \int_R \left(\hat{D}^{(2,j)} \hat{y} \right)^2 \mathrm{d}\,\hat{\omega}.$$

To this end let

$$f(\hat{x}) = \hat{D}^{(0,j)}\hat{y}\Big|_{\hat{t}=\hat{t}^*}, \qquad g(\hat{x}) = \hat{D}^{(0,j)}\hat{I}_h\hat{y}\Big|_{\hat{t}=\hat{t}^*} = \hat{I}_h\hat{D}^{(0,j)}\hat{y}\Big|_{\hat{t}=\hat{t}^*},$$

for some fixed $\hat{t}^*$.

As the solution of the variational problem

$$\min_{p\in H^2(0,1)} \int_0^1 \left(\frac{\mathrm{d}^2}{\mathrm{d}\,\hat{x}^2}p(\hat{x})\right)^2 \mathrm{d}\,\hat{x}$$
$$\text{s. th. } p(0)=a,\ p(1)=b,\ \frac{\mathrm{d}}{\mathrm{d}\,x}p(0)=c,\ \frac{\mathrm{d}}{\mathrm{d}\,x}p(1)=d,$$

is given by the Hermite interpolant (using that the corresponding Euler-Lagrange equation is $p_{xxxx}=0$) we have

$$\int_0^1 \left(\frac{\mathrm{d}^2}{\mathrm{d}\,\hat{x}^2}g(\hat{x})\right)^2 \mathrm{d}\,\hat{x} \le \int_0^1 \left(\frac{\mathrm{d}^2}{\mathrm{d}\,\hat{x}^2}f(\hat{x})\right)^2 \mathrm{d}\,\hat{x}$$

Returning to the definition of the functions f and g and recalling that $\hat{t}^*$ was chosen arbitrarily, the estimate holds for (almost) all $\hat{t}\in(0,1)$ and therefore we have

$$\int_R \left(\hat{D}^{(2,j)}\hat{I}_h\hat{y}\right)^2 \mathrm{d}\,\hat{\omega} \lesssim \int_R \left(\hat{D}^{(2,j)}\hat{y}\right)^2 \mathrm{d}\,\hat{\omega}.$$

Altogether we have proven the estimate

$$\left\|D^{(2,0)}\left(y - I^k_{h\tau}y\right)\right\|^2_{L^2(\theta)} \lesssim \frac{\tau h}{h^4}\int_R \left(\hat{D}^{(4,0)}\hat{y}\right)^2 \mathrm{d}\,\hat{\omega} + \frac{\tau h}{h^4}\int_R \left(\hat{D}^{(2,j)}\hat{y}\right)^2 \mathrm{d}\,\hat{\omega}.$$

Transferring the integrals back on the element θ yields the result. □

Lemma 4.18. *Assume that $y\in H^A(Q)$ with the multi-index set*

$$A=\{(0,k+1),\ (j,1),\ (4,0),\ (2,i)\} \qquad \text{with } j\in\{1,\dots,4\} \text{ and } i\in\{1,\dots,k\}.$$

Then the interpolation error on an element can be bounded by

$$\left\|y - I^k_{h\tau}y\right\|_{L^2(\theta)} \lesssim \tau^{k+1}\left\|D^{(0,k+1)}y\right\|_{L^2(\theta)} + h^4\left\|D^{(4,0)}y\right\|_{L^2(\theta)}.$$

Proof. As in the Lemmas 4.16 and 4.17 we transfer the error to the reference element. On the reference element we can estimate the $L^2(R)$-norm by the stronger $H^{(4,k+1)}(R)$-norm, and by using Theorem 3.25, we get

$$\begin{aligned}\left\|\hat{y} - \hat{I}^k_{h\tau}\hat{y}\right\|_{L^2(R)} &\le \left\|\hat{y} - \hat{I}^k_{h\tau}\hat{y}\right\|_{H^{(4,k+1)}(R)}\\ &\lesssim \left|\hat{y} - \hat{I}^k_{h\tau}\hat{y}\right|_{H^{(4,k+1)}(R)} + \sum_{i=1}^{4\cdot(k+1)}\left|l_i(\hat{y} - \hat{I}^k_{h\tau}\hat{y})\right|\\ &= |\hat{y}|_{H^{(4,k+1)}(R)} + \sum_{i=1}^{4\cdot(k+1)}\left|l_i(\hat{y} - \hat{I}^k_{h\tau}\hat{y})\right|.\end{aligned}$$

For the linear functionals l_i we choose

$$\begin{aligned}
l_i(y) &= y\left(0, \frac{i-1}{k}\right), && \text{for } i = 1, \dots, k+1,\\
l_i(y) &= y\left(1, \frac{i-(k+2)}{k}\right), && \text{for } i = k+2, \dots, 2(k+1),\\
l_i(y) &= D^{(1,0)}y\left(0, \frac{i-2(k+1)-1}{k}\right), && \text{for } i = 2(k+1)+1, \dots, 3(k+1),\\
l_i(y) &= D^{(1,0)}y\left(1, \frac{i-3(k+1)-1}{k}\right), && \text{for } i = 3(k+1)+1, \dots, 4(k+1).
\end{aligned}$$

By the uniqueness of the polynomial interpolation it is clear, that the condition on the functionals of Theorem 3.25 is fulfilled. With the interpolation properties of the interpolation operator $I^k_{h\tau}$ we see that

$$\sum_{i=1}^{4\cdot(k+1)} \left| l_i(\hat{y} - \hat{I}^k_{h\tau}\hat{y}) \right| = 0.$$

By transferring back to the element θ the proof is finished. □

So we have proven all results which we need to prove the interpolation error estimate of Theorem 4.9.

Proof of Theorem 4.9. The interpolation error on every element is bounded with Lemma 4.16, Lemma 4.17 and Lemma 4.18. For an interpolation error estimate on the whole domain we split the integration over the domain to the integration over the elements and sum up. □

Numerical example

Example 4.19. *For a numerical example we consider the problem*

$$\begin{aligned}
-y_{tt} + y_{xxxx} + y &= f && \text{in } (0,1)^2,\\
y &= 0 && \text{on } \{0\} \times (0,1),\\
y_{xx} &= 0 && \text{on } \{0\} \times (0,1),\\
y_x &= 0 && \text{on } \{1\} \times (0,1),\\
y_{xxx} &= 0 && \text{on } \{1\} \times (0,1),\\
y &= 0 && \text{in } (0,1) \times \{0\},\\
y_t - y_{xx} &= 0 && \text{in } (0,1) \times \{T\},
\end{aligned}$$

where the right hand side is chosen, so that the function

$$y = (t-1)^2\, tx^3\, (x-1)^4 \tag{4.19}$$

is the exact solution. For the computation we use finite element meshes with $\tau^2 = h$ and observe the predicted convergence rates in Figure 4.2.

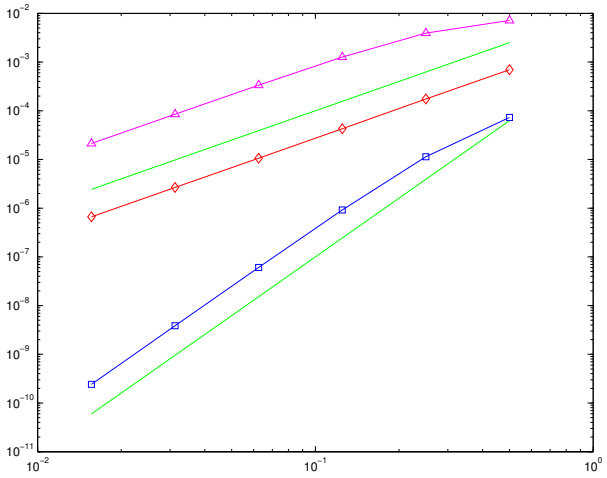


Figure 4.2.: Observed convergence rates for the Hermite-Lagrange tensor product finite element method for the numerical Example 4.19. The $L^2(Q)$-norm of the error is plotted in blue with squares, the $H^{(0,1)}(Q)$-semi norm in red with diamonds and the $H^{(2,0)}(Q)$-semi norm in magenta with triangles. The lines in green without any markers indicate h^2 and h^4.

4.2. Parabolic partial differential equations

In this Section we discuss the discretization of parabolic initial boundary value problems

$$\left.\begin{aligned} \langle y_t, \varphi\rangle_{V^*\times V} + a(y,\varphi) &= \langle u, \varphi\rangle_{H\times H} \\ y(\cdot,0) &= v, \end{aligned}\right\} \tag{4.20}$$

where the bilinear form $a(\cdot,\cdot)$ fulfills Gårding's inequality. We will focus on Crank-Nicolson schemes as time stepping schemes for the discretization of this problem.

Assumption 4.20 (Time Discretization)**.** *For the time discretization we introduce the time grid $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ with time step size $\tau_i = t_i - t_{i-1}$ and time intervals $I_i = (t_{i-1}, t_i)$. We denote the midpoints of the time intervals with $t_{i+\frac{1}{2}} = \frac{t_{i+1}+t_i}{2}$. We introduce $\underline{\tau} = \min_i \tau_i$ and $\overline{\tau} = \max_i \tau_i$ as the minimal and the maximal time step size in the discretization. Further we assume that there is a constant $\gamma > 0$, independent of the discretization level such that*

$$\overline{\tau} \leq \gamma \underline{\tau}.$$

We denote an approximation of a function y at the time $t = t_i$ by y_i and of a function u at the time $t = t_{i+\frac{1}{2}}$ by $u_{i+\frac{1}{2}}$.

As discretization of the parabolic initial boundary value problem 4.20 we discuss the Crank-Nicolson scheme

$$\left.\begin{aligned} \langle y_{h,0}, \varphi\rangle_{H\times H} &= \langle v, \varphi\rangle_{H\times H} \qquad \forall \varphi \in V_h, \\ \left\langle \frac{y_{h,i+1} - y_{h,i}}{\tau}, \varphi \right\rangle_{H\times H} + a\left(\frac{y_{h,i+1} + y_{h,i}}{2}, \varphi\right) &= \left\langle u_{h,i+\frac{1}{2}}, \varphi \right\rangle_{H\times H} \\ &\text{for } i = 0,\ldots,N, \forall \varphi \in V_h, \end{aligned}\right\} \tag{4.21}$$

where the finite dimensional subspace $V_h \subset V$ is chosen as finite element space $V_h \subset V$.

Remark 4.21. *In literature there are different interpretation of the term Crank-Nicolson scheme. In the book of Thomée [125] the scheme* (4.21) *is called Crank-Nicolson scheme. There the midpoint rule is used for the time discretization of the right hand side. In the book of Johnson [66, (8.24)] the scheme*

$$\left\langle \frac{y_{h,i+1} - y_{h,i}}{\tau}, \varphi \right\rangle_{H\times H} + a\left(\frac{y_{h,i+1} + y_{h,i}}{2}, \varphi \right) = \left\langle \frac{u_{h,i} + u_{h,i+1}}{2}, \varphi \right\rangle_{H\times H}$$

is considered as Crank-Nicolson scheme, where the trapezoidal rule is used for the time discretization of the right hand side. The only difference is the approximation of the right hand side.

We will call any scheme of the form

$$\left\langle \frac{y_{h,i+1} - y_{h,i}}{\tau}, \varphi \right\rangle_{H\times H} + a\left(\frac{y_{h,i+1} + y_{h,i}}{2}, \varphi \right) = \left\langle \tilde{u}_{h,i+\frac{1}{2}}, \varphi \right\rangle_{H\times H}$$

with

$$\left\langle \tilde{u}_{h,i+\frac{1}{2}}, \varphi \right\rangle_{H\times H} = \left\langle u(\cdot, t_{i+\frac{1}{2}}), \varphi \right\rangle_{H\times H} + C\left(h^2 + \tau^2\right) \qquad \forall \varphi \in V_h \tag{4.22}$$

a Crank-Nicolson scheme. So for the time approximation of the right hand side the midpoint rule or the trapezoidal rule can be used but also Simpsons rule

$$\tilde{u}_{i+\frac{1}{2}} = \frac{1}{6} u_{k-\frac{1}{2}} + \frac{4}{6} u_{k+\frac{1}{2}} + \frac{1}{6} u_{k-\frac{3}{2}}$$

is equally well suited. This is motivated by the next theorem, in which it is shown that the condition (4.22) *guarantees second order convergence of the Crank-Nicolson scheme.*

Let $y \in \mathcal{Y}$ be the solution of the continuous problem and $y_h \in \mathcal{Y}_h = L^2((0,T), V_h)$ the solution of the problem after discretization in space with linear finite elements. Finally let $y_{h,i}$ be the approximation of y_h with the scheme (4.21) at the time t_i.

Theorem 4.22. *Let*

$$C_1(y, v) = \|v\|_W + \int_0^T \|y_t(\cdot, s)\|_W \,\mathrm{d}\, s, \tag{4.23}$$

$$C_2(y, u) = \int_0^T \|y_{ttt}(\cdot, s)\|_H + \|A y_{,tt}(\cdot, s)\|_H \,\mathrm{d}\, s + \int_0^T \|u_{tt}(\cdot, s)\|_H \,\mathrm{d}\, s. \tag{4.24}$$

Further assume, that

- *the symmetric bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ fulfills Gårding's inequality and is continuous, and further*

$$0 \le a(y, y) \qquad \forall y \in V,$$

- *for the initial data $v \in H^2(\Omega)$ holds,*

- *linear finite elements are used for the spatial discretization,*
- *for the exact solution* $y \in H^3((0,T),L^2(\Omega)) \cap H^2((0,T),H^2(\Omega))$ *holds.*

For a given approximation of the right hand side $u_{h,i+\frac{1}{2}}$ *with*

$$\left\| u_{h,i+\frac{1}{2}} - u(\cdot, t_{i+\frac{1}{2}}) \right\|_{L^2(\Omega)} \leq C_1 h^2 + C_2 \tau^2$$

for $i = 0, \ldots, N-1$ *with* C_1 *and* C_2 *specified in* (4.23) *and* (4.24)*, the error between the solution* y *and the solution* $y_{h,i}$ *of the Crank-Nicolson scheme* (4.21) *is bounded by*

$$\|y_{h,i}(\cdot) - y(\cdot, t_i)\|_{L^2(\Omega)} \lesssim C_1 h^2 + C_2 \tau^2.$$

Remark 4.23 (Regularity)**.** *In the analysis of Crank-Nicolson schemes as time stepping schemes we assume* $y \in H^3((0,T),L^2(\Omega)) \cap H^2((0,T),H^2(\Omega))$*. For such a regularity in a problem with parabolic partial differential equations we need a smooth right hand side and further compatibility conditions on initial and boundary conditions. These are discussed in Theorems 3.39 and 3.40.*

In the example of a smooth domain Ω*, e.g. if the domain is one dimensional, one obtains from Theorem 3.40*

$$y \in L^2((0,T),H^2(\Omega)) \cap L^\infty((0,T),H_0^1(\Omega)) \cap H^1((0,T),L^2(\Omega)).$$

If we assume further

$$u \in H^2((0,T),L^2(\Omega)) \cap H^1((0,T),H^2(\Omega)) \cap L^2((0,T),H^4(\Omega)),$$
$$v \in H_0^1(\Omega),\ Av \in H_0^1(\Omega),\ AAv \in H_0^1(\Omega),\ AAAv \in L^2(\Omega)$$

we get the improved regularity

$$y \in H^3((0,T),L^2(\Omega)) \cap H^2((0,T),H^2(\Omega)) \cap \cap H^1((0,T),H^4(\Omega)) \cap L^2((0,T),H^6(\Omega)).$$

Definition 4.24. *For the error splitting we define the projection* $R_h : V \to V_h$ *as*

$$a(R_h y(\cdot,t_i),\varphi) = a(y(\cdot,t_i),\varphi) \qquad \forall \varphi \in V_h, \tag{4.25}$$

for the case that Gårding's inequality is fulfilled with $k = 0$ *and as*

$$a(R_h y(\cdot,t_i),\varphi) = a(y(\cdot,t_i),\varphi) \qquad \forall \varphi \in V_h, \tag{4.26}$$

$$\text{and} \int_\Omega R_h y(\cdot,t_i)\,\mathrm{d}\,\omega = \int_\Omega y(\cdot,t_i)\,\mathrm{d}\,\omega. \tag{4.27}$$

for the case that $k = 0$ *in Gårding's inequality is not possible.*

Remark 4.25. *In Theorem 3.34 we have seen that we can assume* $k = 0$ *in Gårding's inequality without loss of generality. In Definition 4.24 we have nevertheless distinguished the cases* $k = 0$ *and* $k \neq 0$ *in Gårding's inequality as it is not necessary for the numerical realization of the Crank-Nicolson scheme to perform the transformation* $w = \mathrm{e}^{-\lambda t} y$ *and to discretize the function* w *instead of the function* y*. Furthermore for the use of the transformation* $w = \mathrm{e}^{-\lambda t} y$ *in the context of optimal control problem one has also to care that the state is also part of the cost functional and of the right hand side of the adjoint state.*

Lemma 4.26. *The projection $R_h y(\cdot, t_i)$ is well-defined and if the domain Ω is convex we have the estimate*

$$\|R_h y(\cdot, t_i) - y(\cdot, t_i)\|_{L^2(\Omega)} \leq h^2 \|y(\cdot, t_i)\|_{H^2(\Omega)}.$$

Proof. This projection estimate is well known, even for the case $k = 0$ in Gårding's inequality [23, Chapter 5.2, Chapter 5.7 and Theorem 5.7.6].

We discuss first the case that $k = 0$ in Gårding's inequality is not possible. Consider the function

$$\tilde{y}(\cdot) = y(\cdot, t_i) - \frac{1}{\operatorname{meas}(\Omega)} \int_\Omega y(\cdot, t_i) \,\mathrm{d}\,\omega,$$
$$\tilde{y} \in H^\star(\Omega) = \left\{ v \in H^1(\Omega) : \int_\Omega v \,\mathrm{d}\,\omega = 0 \right\}$$

for any t_i and its projection $\tilde{y}_h \in V_h^\star = \left\{ v \in V_h : \int_\Omega v_h \,\mathrm{d}\,\omega = 0 \right\}$ defined by

$$a(\tilde{y}_h, \varphi) = a(\tilde{y}, \varphi) \qquad \forall \varphi \in V_h. \tag{4.28}$$

It is well known that this projection is unique ([23, Chapter 5.2] or [124, Theorem 4.4.]) and it is well known that

$$\|\tilde{y} - \tilde{y}_h\|_{H^1(\Omega)} \lesssim h \,|\tilde{y}|_{H^2(\Omega)}.$$

As the domain Ω is convex we get second order convergence in $L^2(\Omega)$ with the usual duality argument (see [23, Theorem 5.7.6]). We compute the projection $R_h y(\cdot, t_i)$ as

$$R_h y(\cdot, t_i) = \tilde{y}_h + \frac{1}{\operatorname{meas}(\Omega)} \int_\Omega y(\cdot, t_i) \,\mathrm{d}\,\omega. \tag{4.29}$$

It is easy to see that this $R_h y$ fulfills (4.26) and (4.27).

The projection is unique as (4.28) has a unique solution and any function $y \in H^1(\Omega)$ can be written as $y = y_0 + c$ with $y_0 \in H^\star(\Omega)$ and a constant c.

In the case of $k = 0$ the well known estimate

$$\|\tilde{y} - \tilde{y}_h\|_{H^1(\Omega)} \lesssim h \,|\tilde{y}|_{H^2(\Omega)}.$$

can be applied directly. □

We split the errors into the difference between the exact solution and its projection

$$\rho_i^y(\cdot) = R_h y(\cdot, t_i) - y(\cdot, t_i),$$

and the difference between the projection and the numerical approximation

$$\theta_i^y(\cdot) = y_{h,i}(\cdot) - R_h y(\cdot, t_i),$$

Lemma 4.27. *[125, Theorem 1.2] Assume that the assumptions of Theorem 4.22 are fulfilled. The error between the solution and the corresponding projection can be estimated by*

$$\|\rho_i^y\|_{L^2(\Omega)} = \|R_h y(\cdot, t_i) - y(\cdot, t_i)\|_{L^2(\Omega)} \lesssim h^2 \|v\|_{H^2(\Omega)} + h^2 \int_0^{t_i} \|y_{,t}(\cdot, s)\|_{H^2(\Omega)} \,\mathrm{d}\, s.$$

Proof. The projection the estimate

$$\|R_h y(\cdot,t_i) - y(\cdot,t_i)\|_{L^2(\Omega)} \lesssim h^2 \|y(\cdot,t_i)\|_{H^2(\Omega)} \tag{4.30}$$

is well known (see Lemma 4.26). With the fundamental theorem of calculus (see Theorem 3.18) we have

$$\begin{aligned}\|y(\cdot,t_i)\|_{H^2(\Omega)} &= \left\| v(\cdot) + \int_0^{t_i} y_{,t}(\cdot,s)\,\mathrm{d}\,s \right\|_{H^2(\Omega)} \\ &\leq \|v\|_{H^2(\Omega)} + \int_0^{t_i} \|y_{,t}(\cdot,s)\|_{H^2(\Omega)}\,\mathrm{d}\,s,\end{aligned}$$

which is the desired estimate. □

Proof of Theorem 4.22. With the error splitting and Lemma 4.27 it is sufficient to discuss the difference θ_i^y between the projection and the numerical approximation. For this estimate we follow the proof of [125, Theorem 1.6]. Therefore we transform the right hand side of (4.21) for θ^y and obtain by the use of the discretization scheme

$$\begin{aligned}&\left\langle \frac{\theta_i^y - \theta_{i-1}^y}{\tau}, \varphi \right\rangle_{H\times H} + a\left(\frac{\theta_i^y + \theta_{i-1}^y}{2}, \varphi\right) = \\ &= \left\langle u_{h,i-\frac{1}{2}}(\cdot), \varphi \right\rangle_{H\times H} - a\left(\frac{R_h y(\cdot,t_i) + R_h y(\cdot,t_{i-1})}{2}, \varphi\right) - \left\langle \frac{R_h y_i(\cdot) - R_h y_{i-1}(\cdot)}{\tau}, \varphi \right\rangle_{H\times H}.\end{aligned}$$

The definition of R_h and adding and subtracting the differential equation for the exact solution yields

$$\begin{aligned}&\left\langle \frac{\theta_i^y - \theta_{i-1}^y}{\tau}, \varphi \right\rangle_{H\times H} + a\left(\frac{\theta_i^y + \theta_{i-1}^y}{2}, \varphi\right) \\ &= \left\langle y_{,t}(\cdot,t_{i+\frac{1}{2}}), \varphi \right\rangle_{H\times H} + a\left(y(\cdot,t_{i+\frac{1}{2}}), \varphi\right) + \left\langle u_{h,i-\frac{1}{2}}(\cdot) - u(\cdot,t_{i-\frac{1}{2}}), \varphi \right\rangle_{H\times H} \\ &- a\left(\frac{y(\cdot,t_i) + y(\cdot,t_{i-1})}{2}, \varphi\right) - \left\langle \frac{R_h y_i(\cdot) - R_h y_{i-1}(\cdot)}{\tau}, \varphi \right\rangle_{H\times H}\end{aligned}$$

Using the regularity assumption $Ay(\cdot,t) \in H$ and adding another zero ends in

$$\begin{aligned}&\left\langle \frac{\theta_i^y - \theta_{i-1}^y}{\tau}, \varphi \right\rangle_{H\times H} + a\left(\frac{\theta_i^y + \theta_{i-1}^y}{2}, \varphi\right) \\ &= -\left\langle (R_h - I)\frac{y(\cdot,t_i) - y(\cdot,t_{i-1})}{\tau} + \frac{y(\cdot,t_i) - y(\cdot,t_{i-1})}{\tau} - y_{,t}(\cdot,t_{i+\frac{1}{2}}), \varphi \right\rangle_{H\times H} \\ &- \left\langle A\frac{y(\cdot,t_{i+1}) + y(\cdot,t_i)}{2} - Ay(\cdot,t_{i+\frac{1}{2}}), \varphi \right\rangle_{H\times H} + \left\langle u_{h,i-\frac{1}{2}}(\cdot) - u(\cdot,t_{i-\frac{1}{2}}), \varphi \right\rangle_{H\times H} \\ &=: -\langle \omega_{1,i}, \varphi \rangle_{H\times H} - \langle \omega_{2,i}, \varphi \rangle_{H\times H} + \left\langle u_{h,i-\frac{1}{2}}(\cdot) - u(\cdot,t_{i-\frac{1}{2}}), \varphi \right\rangle_{H\times H}.\end{aligned}$$

By simple computation we have the identity

$$\begin{aligned}\left\langle \frac{\theta_i^y - \theta_{i-1}^y}{\tau}, \frac{\theta_i^y + \theta_{i-1}^y}{2} \right\rangle_{H\times H} &= \frac{1}{2\tau}\|\theta_i\|_H^2 - \frac{1}{2\tau}\|\theta_{i-1}\|_H^2 \\ &= \frac{1}{2\tau}\left(\|\theta_i\|_H + \|\theta_{i-1}\|_H^2\right)\left(\|\theta_i\|_H - \|\theta_{i-1}\|_H^2\right).\end{aligned}$$

We recall the the assumption

$$0 \leq a\left(\frac{\theta_i^y + \theta_{i-1}^y}{2}, \frac{\theta_i^y + \theta_{i-1}^y}{2}\right).$$

and use $\frac{\theta_i^y + \theta_{i-1}^y}{2}$ as test function which gives

$$\begin{aligned}
&\frac{1}{2\tau}\left(\|\theta_i\|_H + \|\theta_{i-1}\|_H^2\right)\left(\|\theta_i\|_H - \|\theta_{i-1}\|_H^2\right) \leq \\
&\left\langle \frac{\theta_i^y - \theta_{i-1}^y}{\tau}, \frac{\theta_i^y + \theta_{i-1}^y}{2} \right\rangle_{H\times H} + a\left(\frac{\theta_i^y + \theta_{i-1}^y}{2}, \frac{\theta_i^y + \theta_{i-1}^y}{2}\right) \leq \\
&\left(\|\omega_{1,i}\|_H + \|\omega_{2,i}\|_H + \left\|u_{h,i-\frac{1}{2}} - u(\cdot, t_{i-\frac{1}{2}})\right\|_{L^2(\Omega)}\right)\frac{1}{2}\left(\|\theta_i^y\|_H + \left\|\theta_{i-1}^y\right\|_H\right).
\end{aligned}$$

After cancellation of the common factor $\|\theta_i^y\|_H + \left\|\theta_{i-1}^y\right\|_H$ and repeated application we get

$$\begin{aligned}
\|\theta_i^y\|_H &\leq \left\|\theta_{i-1}^y\right\|_H + \tau\|\omega_{1,i}\|_H + \tau\|\omega_{2,i}\|_H + \tau\left\|u_{h,i-\frac{1}{2}} - u(\cdot, t_{i-\frac{1}{2}})\right\|_H \\
&\leq \|\theta_0^y\|_H\, L^2(\Omega) + \sum_{j=1}^{i}\tau\|\omega_{1,j}\|_H + \sum_{j=1}^{i}\tau\|\omega_{2,j}\|_H + \tau\sum_{j=1}^{i}\left(C_1h^2 + C_2\tau^2\right) \\
&\leq \|\theta_0^y\|_H + \sum_{j=1}^{i}\tau\|\omega_{1,j}\|_H + \sum_{j=1}^{i}\tau\|\omega_{2,j}\|_H + \left(C_1h^2 + C_2\tau^2\right).
\end{aligned}$$

With the projection estimate of Lemma 4.27 we can estimate

$$\|\theta_0^y\|_H = \|v - R_hv\|_H \lesssim h^2\|v\|_W$$

and

$$\tau\left\|(R_h - I)\frac{y(\cdot, t_i) - y(\cdot, t_{i-1})}{\tau}\right\|_H \leq h^2C_1.$$

The second term of ω_j can be estimated as in [125, Theorem 1.6]

$$\begin{aligned}
&\tau\left\|\frac{y(\cdot, t_i) - y(\cdot, t_{i-1})}{\tau} - y_{,t}(\cdot, t_{i+\frac{1}{2}})\right\|_H \\
&= \frac{1}{2}\left\|\int_{t_{j-1}}^{t_{j-\frac{1}{2}}}(t - t_{j-1})^2y_{,ttt}\,\mathrm{d}\,t + \int_{t_{j-\frac{1}{2}}}^{t_j}(t - t_j)^2y_{,ttt}\,\mathrm{d}\,t\right\|_H \lesssim \tau^2\int_{t_{j-1}}^{t_j}\|y_{,ttt}\|_H\,\mathrm{d}\,t.
\end{aligned}$$

The remaining estimate for $\omega_{2,j}$ given by

$$\tau\left\|Ay(\cdot, t_{i+\frac{1}{2}}) - A\left(\frac{y(\cdot, t_{i+1}) + y(\cdot, t_i)}{2}\right)\right\|_H \lesssim \tau^2\int_{t_{j-1}}^{t_j}\|Ay_{,tt}\|_H\,\mathrm{d}\,t$$

follows similarly. So the proof is done. □

4.3. Hamiltonian systems

Finally we consider the discretization of Hamiltonian systems.

Definition 4.28. *A Hamiltonian system is a system of differential equations with*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,t}p &= -H_q(p,q),\\ \frac{\mathrm{d}}{\mathrm{d}\,t}q &= -H_p(p,q),\end{aligned}$$

where the Hamiltonian H is a given function.

Remark 4.29. *In mechanics Hamiltonian systems are common and have a initial condition for both sets of variables p and q. But the initial conditions are not part of the definition of a Hamiltonian system, it is also possible to pose initial conditions for p and terminal conditions for q. This is the case if we consider Hamiltonian systems for optimal control problems.*

A common scheme for Hamiltonian system is the second order Störmer-Verlet scheme. Hairer, Lubich and Wanner propose in [57, Chapter II.2] and [56, (1.24)] an extension of the Störmer-Verlet scheme to general partitioned problems

$$\dot{y} = g(y,p), \qquad \dot{p} = f(y,p). \tag{4.31}$$

As noted in [56, Section 1.8] this scheme goes back to [32]. The scheme is given as

$$\left.\begin{aligned} p_n &= p_{n-1/2} + \frac{\tau}{2} f(y_n, p_{n-1/2})\\ p_{n+1/2} &= p_n + \frac{\tau}{2} f(y_n, p_{n+1/2})\\ y_{n+1} &= y_n + \frac{\tau}{2}\left(g(y_n, p_{n+1/2}) + g(y_{n+1}, p_{n+1/2})\right)\end{aligned}\right\} \tag{SV}$$

Remark 4.30. *With pure algebraic manipulation on the Störmer-Verlet scheme* (SV) *we see:*

1. *By elimination of the first equation of* (SV) *the Störmer-Verlet scheme can be written as equations in the time points t_i for the function y and in the time points $t_{i+1/2}$ for the function p.*
2. *On the other hand by elimination of the second equation of* (SV)*, the scheme can be written in the time points t_i for both functions y and p.*

Theorem 4.31. *[57, Theorem II.2.2 or Theorem III.2.5 or Theorem VI.3.4] The Störmer-Verlet scheme is a scheme of second order, i.e. let y and p the solution of* (4.31) *and y_i, $p_{i+1/2}$ the solution of* (SV)*, then*

$$\|y(t_i) - y_i\|_{\mathbb{R}^n} + \left\|p(t_{i+1/2}) - p_{i+1/2}\right\|_{\mathbb{R}^n} \lesssim \mathcal{O}(\tau^2).$$

As mentioned optimal control problems are also Hamiltonian systems. The Störmer-Verlet scheme is a symplectic partitioned Runge-Kutta scheme. In [16, 17] Bonnans and Laurent-Varin discuss the application of such schemes to optimal control problems with ordinary differential equations. Order conditions for third and higher order symplectic partitioned Runge-Kutta scheme are also given in [16, 17, 54, 55, 57], but we do not discuss such schemes as they need a high regularity of the solution. The Störmer-Verlet scheme (SV) fulfills the order conditions of [16, 17].

5. Parabolic Optimal Control Problems

Contents

In this Chapter we introduce an abstract parabolic optimal control problem. For this problem we derive the optimality conditions and discuss the Hamiltonian nature of these conditions. We discuss also that the optimality conditions for this problem can be reduced to an equation which only involves the optimal control but not the state or the optimal state but not the optimal control.

In the discussion of the discretization in the next chapter we will discuss the discretization of the optimal control problem, the Hamiltonian system of optimality conditions and the equation which does not involve the control.

5.1. Optimality conditions

For the statement of an abstract parabolic optimal control problem we need some very general assumptions.

Assumption 5.1 (General Assumptions)**.** *We assume that the following very general assumptions hold for the rest of this thesis.*

1. *There is an Gelfand triplet* $V \subseteq H \cong H^* \subseteq V^*$.
2. *There are three constants* α, β, $\nu \in \mathbb{R}$ *with* $\alpha, \beta \geq 0$, $\alpha + \beta > 0$ *and* $\nu > 0$.
3. *For the three functions* $v \in H$, $y_D \in H$ *and* $y_d(\cdot, t) \in H$ *holds.*
4. *The operator* $M : V^* \to V^*$ *is linear, positive definite, self-adjoint and continuous.*
5. *The operators* $A : V \to V^*$ *and* $G : H \to V^*$ *are linear and continuous.*
6. *The operators* $M_D : H \to H$, $M_d : H \to H$ *and* $M_u : H \to H$ *are linear, positive semi definite, self-adjoint and continuous.*

With these assumptions we formulate a general parabolic optimal control problem.

Problem 5.2 (Parabolic Optimal Control Problem)**.** *Let the Assumptions 5.1 be fulfilled. Then the abstract parabolic optimal control problem is defined by*

$$\left.\begin{aligned} &\min_{y,u} J(y,u), \\ &s.t.\ My_t + Ay = Gu, \\ &\qquad My(0,\cdot) = Mv(\cdot), \end{aligned}\right\} \tag{5.1}$$

where the equations should be understood in the sense of $L^2(0,T;V^*)$ and the cost functional $J(y,u)$ is defined by

$$J(y,u) = \frac{\alpha}{2}\left\|M_D^{1/2}\left(y(\cdot,T) - y_D(\cdot)\right)\right\|_H^2 + \frac{\beta}{2}\int_0^T \left\|M_d^{1/2}\left(y(\cdot,t) - y_d(\cdot,t)\right)\right\|_H^2 \mathrm{d}\,t + \\ + \frac{\nu}{2}\int_0^T \left\|M_u^{1/2}u\right\|_H^2 \mathrm{d}\,t.$$

Remark 5.3. *To discuss the solution of* (5.1) *in the space $\mathcal{C}(0,T;V^*)$, we need the regularity $y \in \mathcal{C}^1(0,T;V^*)$. This regularity can be established with the regularity theory for the differential equation in* (5.1)*, regularity assumptions on the data and the discussion of the optimality conditions.*

Remark 5.4. *The existence of the positive square roots $M_D^{1/2}$, $M_d^{1/2}$ and $M_u^{1/2}$ of the operators M_D, M_d and M_u is established in Theorem 3.8. For the computation of the cost functional $J(y,u)$ we do not need to know or compute $M^{1/2}$ as $\|M^{1/2}x\|_H^2 = \langle M^{1/2}x, M^{1/2}x\rangle_{H\times H} = \langle Mx, x\rangle_{H\times H}$ as the root of a self adjoint operator is a self adjoint operator itself (see Corollary 3.9).*

Remark 5.5 (More general problems)**.** *In setting of Problem 5.2 we restrict ourselves to the case that the control space and the space of observation of the state y in the cost functional match with the pivot space H of our Gelfand triplet. For the discussion of the optimality conditions for optimal control problems with more general control spaces U see e.g. [62, Chapter 3] and [78, Chapter III.2.]. If one chooses more general control spaces, one has to be careful as the Riesz isomorphism between U and U^* is not the identity on $L^2(\Omega)$ (see [62]). Therefore here the control space $U = H$ is chosen for clarity and shortness of presentation. With the introduction of a more general control space one could also discuss Neumann boundary control if one chooses $U = L^2(\partial\Omega)$ or $U = H^{1/2}(\partial\Omega)$ and G as the extension operator (see [78, 101]).*

In applications of the Problem 5.2 to optimal control problems with parabolic partial differential equations the operators possess typically more regularity as given in Problem 5.2. Namely we assume the following properties.

Assumption 5.6.

1. *Let $W \subseteq V$ be a Hilbert space.*
2. *Let the Assumptions 5.1 hold.*
3. *The linear operator $M: V^* \to V^*$ is the continuation of a positive definite operator $\tilde{M}: W \to W$.*
4. *The operator A induces a continuous bilinear form which fulfills Gårding's inequality* (3.20)*.*

5. *The linear operator $G : H \to V^*$ is the continuation of an operator $\tilde{G} : V \to V$. Therefore we know that $\tilde{G}^* : V \to V$ and so we have $G^* : V \to V$.*

We give now some examples which are covered by this abstract optimal control problem.

Example 5.7. *Distributed control of the heat equation with homogeneous Dirichlet boundary conditions. In this case we choose $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, $A = -\Delta$ and $G = M = M_d = M_D = M_u = I$. In this case the Assumptions 5.1 and 5.6 hold.*

Example 5.8. *Distributed control of the heat equation with homogeneous Neumann boundary conditions. In this case we choose $V = H^1(\Omega)$, $H = L^2(\Omega)$, $A = -\Delta$ and $G = M = M_d = M_D = M_u = I$.*

Example 5.7 and Example 5.8 are special cases of the more general following example.

Example 5.9. *Distributed control and observation of the heat equation with homogeneous Dirichlet boundary conditions on Γ_1 and homogeneous Neumann boundary conditions on Γ_2, where $\Gamma_1 \cup \Gamma_2 = \partial\Omega$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$. In this case we choose $V = \left\{v \in H^1(\Omega) : \ v|_{\Gamma_1} = 0\right\}$, $H = L^2(\Omega)$, $A = -\Delta$ and $G = M = M_d = M_D = M_u = I$.*

Example 5.10. *Distributed control of the heat equation with homogeneous Neumann boundary conditions, where the control only acts on a sub domain $\Omega_U \subset \Omega$ and the desired states are only given on sub domains on $\Omega_D \subset \Omega$ and $\Omega_d \subset \Omega$ respectively. In this case we choose $V = H^1$, $H = L^2(\Omega)$, $A = -\Delta$, $M = I$. The other operators can be defined by $\langle Gu, \varphi\rangle_{H\times H} = \langle M_u u, \varphi\rangle_{H\times H} = \int_\Omega \chi_{\Omega_U} u\varphi \,\mathrm{d}\,\omega$, $\langle M_D v, \varphi\rangle_{H\times H} = \int_\Omega \chi_{\Omega_D} v\varphi \,\mathrm{d}\,\omega$ and $\langle M_d v, \varphi\rangle_{H\times H} = \int_\Omega \chi_{\Omega_d} v\varphi \,\mathrm{d}\,\omega$ with the characteristic functions χ_{Ω_i} of Ω_i for $\Omega_i \in \{\Omega_D, \Omega_D, \Omega_U\}$.*

This example fits into the setting of Problem 5.2. But due to the definition of the spatial operators one would assume that there are singularities near the boundaries of Ω_d, Ω_D and Ω_U. Therefore the stronger regularity assumptions of Assumption 5.6 are not fulfilled, i.e. the control operator G is not a map $H^1(\Omega) \to H^1(\Omega)$. A remedy could be the use of weighted Sobolev spaces and graded meshes or the discussion of a regularized problems, where the characteristic functions are replaced by

$$(\eta_\varepsilon * \chi_{\Omega_i})(x) = \int_\Omega \eta_\varepsilon(\xi)\chi_{\Omega_i}(x-\xi) \,\mathrm{d}\,\omega$$

with a smooth function η_ε.

Example 5.11. *A finite element discretization of distributed control of the heat equation with homogeneous Dirichlet boundary conditions or homogeneous Neumann boundary conditions. In this case we choose $V = H = \mathbb{R}^n$, A as stiffness matrix. M is a mass matrix, where the Dirichlet boundary conditions are incorporated, and $G = M_d = M_D$ are mass matrices, where no boundary conditions have been considered.*

For the deduction of the optimality conditions of the Problem 5.2 we introduce the Lagrangian

$$\left.\begin{aligned}\mathcal{L}(y,u,p) = {} & \frac{\alpha}{2}\left\|M_D^{1/2}\left(y(\cdot,T) - y_D(\cdot)\right)\right\|_H^2 + \frac{\beta}{2}\int_0^T \left\|M_d^{1/2}\left(y(\cdot,t) - y_d(\cdot,t)\right)\right\|_H^2 \,\mathrm{d}\,t + \\ & + \frac{\nu}{2}\int_0^T \left\|M_u^{1/2}u\right\|_H^2 \,\mathrm{d}\,t + \\ & + \int_0^T \langle My_t + Ay - Gu, p\rangle_{V^*\times V} \,\mathrm{d}\,t + \langle M\left(y(\cdot,0) - v\right), p(\cdot,0)\rangle_{H\times H}\end{aligned}\right\} \tag{5.2}$$

and compute the first order condition for a stationary point of the Lagrangian, i.e. we set the first variations to zero. The Lagrange functional is well defined for $y \in \mathcal{Y}$ and $u,\ p \in \mathcal{P}$. For the computation of the first variation we choose admissible variations $\varphi \in \mathcal{P}$ and $\phi \in \mathcal{Y}$ with $\phi(\cdot,0) = 0$. The optimal solution $(\bar{y}, \bar{p}, \bar{u})$ fulfills

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\mathcal{L}(\bar{y},\bar{u},\bar{p}+\varepsilon\varphi)\Big|_{\varepsilon=0} &= \int_0^T \langle M\bar{y}_t + A\bar{y} - G\bar{u}, \varphi\rangle_{V^*\times V}\,\mathrm{d}\,t \\ &\quad + \langle M\left(\bar{y}(\cdot,0) - v\right), \varphi(\cdot,0)\rangle_{H\times H} = 0, \qquad (5.3)\\ \frac{\partial}{\partial\varepsilon}\mathcal{L}(\bar{y},\bar{u}+\varepsilon\varphi,\bar{p})\Big|_{\varepsilon=0} &= \int_0^T \nu\langle M_u\bar{u},\varphi\rangle_{H\times H} - \langle\varphi, G^*\bar{p}\rangle_{V^*\times V}\,\mathrm{d}\,t = 0, \qquad (5.4)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\mathcal{L}(\bar{y}+\varepsilon\phi,\bar{u},\bar{p})\Big|_{\varepsilon=0} &= \alpha\langle M_D\left(\bar{y}(\cdot,T) - y_D(\cdot)\right), \phi\rangle_{H\times H} + \\ &\quad + \beta\int_0^T \langle M_d\left(\bar{y}(\cdot,t) - y_d(\cdot,t)\right), \phi\rangle_{H\times H}\,\mathrm{d}\,t + \\ &\quad + \int_0^T \langle M\phi_t + A\phi, \bar{p}\rangle_{V^*\times V} + \langle M\phi(\cdot,0), \bar{p}(\cdot,0)\rangle_{H\times H}\,\mathrm{d}\,t = 0. \qquad (5.5)\end{aligned}$$

For the last equation we use partial integration in time

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\mathcal{L}(y+\varepsilon\phi,u,p)\Big|_{\varepsilon=0} &= \alpha\langle M_D\left(\bar{y}(\cdot,T) - y_D(\cdot)\right), \phi\rangle_{H\times H} \\ &\quad + \langle M\phi(\cdot,T), \bar{p}(\cdot,T)\rangle_{V^*\times V} - \langle M\phi(\cdot,0), \bar{p}(\cdot,0)\rangle_{V^*\times V} \\ &\quad + \beta\int_0^T \langle M_d\left(\bar{y}(\cdot,t) - y_d(\cdot,t)\right), \phi\rangle_{H\times H}\,\mathrm{d}\,t + \\ &\quad + \int_0^T \langle -M\bar{p}_t, \phi\rangle_{V^*\times V} + \langle A^*\bar{p}, \phi\rangle_{V^*\times V}\,\mathrm{d}\,t \\ &\quad + \langle M\phi(\cdot,0), \bar{p}(\cdot,0)\rangle_{H\times H} = 0. \qquad (5.6)\end{aligned}$$

As we have used admissible variations ϕ with $\phi(\cdot,0) = 0$ the term $\langle M\phi(\cdot,0), \bar{p}(\cdot,0)\rangle_{H\times H}$ vanishes. For the partial integration in time we do not only need $y \in \mathcal{Y}$ and $u,\ p \in \mathcal{P}$ as for the Lagrangian, but we need also $p \in \mathcal{Y}$, so that the time derivative of the adjoint state p is well defined.

As the optimality conditions should be fulfilled for all admissible variations φ and ϕ the optimality conditions are the weak form of the following system.

Problem 5.12 (Optimality Conditions)**.** *Find $\bar{y}, \bar{p} \in L^2(0,T;V) \cap H^1(0,T;V^*)$ and $\bar{u} \in L^2(0,T;V)$ so that*

$$\left.\begin{aligned} M\bar{y}_t + A\bar{y} &= G\bar{u},\\ M\bar{y}(\cdot,0) &= Mv(\cdot),\\ M\bar{p}_t - A^*\bar{p} &= \beta M_d\left(\bar{y} - y_d\right),\\ M\bar{p}(\cdot,T) &= \alpha M_D\left(y_D(\cdot) - \bar{y}(\cdot,T)\right),\\ M_u\bar{u} &= \frac{1}{\nu}G^*\bar{p}\end{aligned}\right\} \qquad (5.7)$$

hold in the sense of V^*.

Remark 5.13. *We have chosen the sign of the Lagrange multiplier so that* $\bar{u} = \frac{1}{\nu} G^* \bar{p}$ *as in [19, 20, 50]. The other choice,* $\bar{u} = -\frac{1}{\nu} G^* \bar{p}$*, is also popular.*

Theorem 5.14. *Let the Assumptions 5.6 hold. Then, for a given control* u *there is a unique solution of the state equation*

$$\begin{aligned} My_t + Ay &= Gu, \\ My(\cdot, 0) &= Mv(\cdot). \end{aligned}$$

Further for a given state y *there is a unique solution of the adjoint equation*

$$\begin{aligned} Mp_t - A^* p &= \beta M_d \left(y - y_d\right), \\ Mp(\cdot, T) &= \alpha M_D \left(y_D(\cdot) - y(\cdot, T)\right). \end{aligned}$$

Proof. The existence and uniqueness of the state equation follows directly with Theorem 3.37 as the norm $\langle My, y \rangle_{H \times H}^{1/2}$ is an equivalent $L^2(\Omega)$-norm.

For the adjoint equation we use the same arguments and the time transformation $\tilde{t} = T - t$, with which the adjoint equation can be transformed to an initial-boundary value problem for a parabolic partial differential equation. □

As the Lagrange method is a formal method we need to establish that the system of the optimality conditions is solvable. This can be done by the introduction of a reduced cost functional $j(u)$, where the state y in the cost functional $J(y, u)$ is replaced by $\mathcal{S}Gu$ with the (linear) solution operator $\mathcal{S}$ of the partial differential equation and we write $y(\cdot, \cdot; u) = \mathcal{S}Gu$, so

$$\begin{aligned} j(u) &= J(\mathcal{S}Gu, u) = \\ &= \frac{\alpha}{2} \left\| M_D^{1/2} \left((\mathcal{S}Gu)(\cdot, T) - y_D(\cdot) \right) \right\|_H^2 + \frac{\beta}{2} \int_0^T \left\| M_d^{1/2} \left((\mathcal{S}Gu)(\cdot, t) - y_d(\cdot, t) \right) \right\|_H^2 \mathrm{d}\, t + \\ &+ \frac{\nu}{2} \int_0^T \left\| M_u^{1/2} u \right\|_H^2 \mathrm{d}\, t. \end{aligned}$$

The existence of the linear operator $\mathcal{S}$ follows by the solvability of the state equation.

Lemma 5.15. *The reduced cost functional* $j(u)$ *has a unique minimum, the optimal control* $\bar{u}$*.*

Proof. The the reduced cost functional is quadratic and therefore a convex functional with

$$\begin{aligned} j(0) &< \infty, \\ \lim_{k \to \infty} j(k\varphi) &= \infty \qquad\qquad \forall \varphi \in H \text{ with } \|\varphi\|_H = 1. \end{aligned}$$

Further the functional is lower semicontinuous. This implies that a minimizer u of the functional exists [78, Remark 1.2]. As the quadratic functional is strictly convex the uniqueness follows by standard arguments. □

For the reduced functional we compute the first directional derivative.

Lemma 5.16. *The directional derivative of the reduced cost functional at the optimal control $\bar{u}$ in the direction u is given by*

$$j'(\bar{u})u = \alpha\langle M_D\mathcal{S}G\bar{u}(\cdot,T) - M_D y_D(\cdot), \mathcal{S}Gu(\cdot,T)\rangle_{H\times H} + \\ + \beta\int_0^T \langle M_d\mathcal{S}G\bar{u} - M_d y_d, \mathcal{S}Gu\rangle_{H\times H}\,\mathrm{d}\,t + \nu\int_0^T \langle M_u\bar{u}, u\rangle_{H\times H}\,\mathrm{d}\,t.$$

Proof. For the computation of the derivative of the reduced cost functional we compute the first variation as Raymond [101]. For the reduced cost functional we have

$$\begin{aligned} j(\bar{u}) = {} & \frac{\alpha}{2}\left\langle M_D\left((\mathcal{S}G\bar{u})(\cdot,T) - y_D(\cdot)\right), (\mathcal{S}G\bar{u})(\cdot,T) - y_D(\cdot)\right\rangle_{H\times H} + \\ & + \frac{\beta}{2}\int_0^T \left\langle M_d\left((\mathcal{S}G\bar{u})(\cdot,t) - y_d(\cdot,t)\right), (\mathcal{S}G\bar{u})(\cdot,t) - y_d(\cdot,t)\right\rangle_{H\times H}\,\mathrm{d}\,t + \\ & + \frac{\nu}{2}\int_0^T \langle M_u\bar{u},\bar{u}\rangle_{H\times H}\,\mathrm{d}\,t \end{aligned}$$

and

$$\begin{aligned} j(\bar{u}+\varepsilon u) = {} & \frac{\alpha}{2}\left\langle M_D\left((\mathcal{S}G(\bar{u}+\varepsilon u))(\cdot,T) - y_D(\cdot)\right), (\mathcal{S}G(\bar{u}+\varepsilon u))(\cdot,T) - y_D(\cdot)\right\rangle_{H\times H} + \\ & + \frac{\beta}{2}\int_0^T \left\langle M_d\left((\mathcal{S}G(\bar{u}+\varepsilon u))(\cdot,t) - y_d(\cdot,t)\right), (\mathcal{S}G(\bar{u}+\varepsilon u))(\cdot,t) - y_d(\cdot,t)\right\rangle_{H\times H}\,\mathrm{d}\,t + \\ & + \frac{\nu}{2}\int_0^T \langle M_u(\bar{u}+\varepsilon u),\bar{u}+\varepsilon u\rangle_{H\times H}\,\mathrm{d}\,t \\ = {} & j(\bar{u}) + \\ & + \varepsilon\alpha\left\langle M_D\left((\mathcal{S}G(\bar{u}))(\cdot,T) - y_D(\cdot)\right), (\mathcal{S}G(u))(\cdot,T) - y_D(\cdot)\right\rangle_{H\times H} + \\ & + \varepsilon\beta\int_0^T \left\langle M_d\left((\mathcal{S}G(\bar{u}))(\cdot,t) - y_d(\cdot,t)\right), (\mathcal{S}G(u))(\cdot,t) - y_d(\cdot,t)\right\rangle_{H\times H}\,\mathrm{d}\,t + \\ & + \varepsilon\nu\int_0^T \langle M_u(\bar{u}), u\rangle_{H\times H}\,\mathrm{d}\,t + \\ & + \varepsilon^2\frac{\alpha}{2}\langle M_D(\mathcal{S}Gu)(\cdot,T), (SGu)(\cdot,T)\rangle_{H\times H} + \varepsilon^2\frac{\beta}{2}\int_0^T \langle M_d SGu, SGu\rangle_{H\times H}\,\mathrm{d}\,t + \\ & + \varepsilon^2\frac{\nu}{2}\int_0^T \langle M_u u, u\rangle_{H\times H}\,\mathrm{d}\,t. \end{aligned}$$

With the definition of the directional derivative $j'(\bar{u})u = \lim_{\varepsilon\to 0}\frac{j(\bar{u}+\varepsilon u)-j(\bar{u})}{\varepsilon}$ the proof is done. □

For the simplification of this derivative we can (re)introduce the adjoint state $\bar{p}$ and use a generalized partial integration which we prove in the following Lemma.

Lemma 5.17. *For the optimal state $\bar{y} = \mathcal{S}G\bar{u}$, defined by the state equation*

$$\begin{aligned} M\bar{y}_t + A\bar{y} &= G\bar{u}, \\ M\bar{y}(\cdot,0) &= Mv \end{aligned}$$

and the adjoint state $\bar{p}$ and y^φ defined by

$$\begin{aligned} M\bar{p}_t - A^*\bar{p} &= \beta M_d\left(\bar{y} - y_d\right), & My_t^\varphi + Ay^\varphi &= G\varphi, \\ M\bar{p}(\cdot,T) &= \alpha M_D\left(y_D(\cdot) - \bar{y}(\cdot,T)\right), & My^\varphi(\cdot,0) &= 0. \end{aligned}$$

the following integration formula holds

$$\begin{aligned} \alpha\langle M_D\bar{y} - M_Dy_D, y^\varphi\rangle_{H\times H} + \beta\int_0^T \langle M_d\bar{y} - M_dy_d, y^\varphi\rangle_{H\times H}\,\mathrm{d}\,t = \\ = -\int_0^T \langle G^*\bar{p},\varphi\rangle_{H\times H} \end{aligned}$$

Proof. We follow the ideas of the proof of Raymond [101, Theorem 5.2.3.]. By integration by parts in time and the use of the initial condition for y^φ and the terminal condition for $\bar{p}$ we get

$$\begin{aligned} \int_0^T \langle My_t^\varphi,\bar{p}\rangle_{V^*\times V}\,\mathrm{d}\,t = &-\int_0^T \langle M\bar{p}_t, y^\varphi\rangle_{V^*\times V}\,\mathrm{d}\,t \\ &+ \langle M\bar{p}(T),\varphi(T)\rangle_{H\times H} - \langle My^\varphi(0),\varphi(0)\rangle_{H\times H} = \\ &= \int_0^T -\langle A^*\bar{p}, y^\varphi\rangle + \beta\langle M_d\left(y_d - \bar{y}\right), y^\varphi\rangle_{H\times H}\,\mathrm{d}\,t \\ &+ \alpha\langle M_D\left(y_D - \bar{y}(T)\right), y^\varphi\rangle_{H\times H} \end{aligned}$$

And therefore we have, as claimed in this lemma

$$\begin{aligned} -\int_0^T \langle G\varphi,\bar{p}\rangle_{H\times H}\,\mathrm{d}\,t = -\int_0^T \langle My_t^\varphi,\bar{p}\rangle_{V^*\times V} + \langle Ay^\varphi,\bar{p}\rangle_{V^*\times V}\,\mathrm{d}\,t = \\ = \int_0^T \beta\langle M_d(\bar{y} - y_d), y^\varphi\rangle_{H\times H}\,\mathrm{d}\,t + \alpha\langle M_D(\bar{y}(T) - y_D), y^\varphi\rangle_{H\times H}. \end{aligned}$$

□

Application of this Lemma to the gradient of the reduced cost functional of Lemma 5.16 yields the following representation of the gradient.

Theorem 5.18. *The gradient of the reduced cost functional has the representation*

$$j'(\bar{u})u = \int_0^T \langle \nu M_u u - G^*\bar{p}, u\rangle_{H\times H}\,\mathrm{d}\,t.$$

With this theorem the optimality condition $j'(\bar{u})u = 0$ is equivalent to the optimality conditions (5.7) and therefore we have justified the use of the formal Lagrange technique.

In the next section we have a closer look to the optimality conditions.

5.2. Connection to Hamiltonian systems

In this section we have a close look on the optimality conditions (5.7),

$$\begin{aligned} M\bar{y}_t + A\bar{y} &= \frac{1}{\nu}GM_u^{-1}G^*\bar{p}, & M\bar{p}_t - A^*\bar{p} &= \beta M_d(\bar{y} - y_d), \\ M\bar{y}(\cdot,0) &= Mv(\cdot) & M\bar{p}(\cdot,T) &= \alpha M_D(y_D(\cdot) - \bar{y}(\cdot,T)), \end{aligned}$$

where we have eliminated the optimal control due to the optimality condition $\nu M_u \bar{u} = G^* \bar{p}$. Our goal is to interpret these conditions as Hamiltonian system.

If we have now a close look at the optimality conditions, we observe that this system has the structure of a Hamiltonian system, i.e.

$$M\bar{y}_t = -H_p = -A\bar{y} + \frac{1}{\nu} G M_u^{-1} G^* \bar{p}, \tag{5.8}$$

$$M\bar{p}_t = H_y = A^*\bar{p} + \beta M_d \bar{y} - \beta M_d y_d \tag{5.9}$$

with the Hamiltonian

$$\begin{aligned} H(y,p) = \frac{\beta}{2}\langle M_d y, y\rangle_{H\times H} - \beta\langle M_d y_d, y\rangle_{H\times H} + \langle Ay, p\rangle_{V^*\times V} \\ - \frac{1}{2\nu}\langle G M_u^{-1} G^* p, p\rangle_{H\times H}. \end{aligned} \tag{5.10}$$

Remark 5.19. *In contrast to classical Hamiltonian systems known in mechanics, where the system has initial conditions for y and p, our system has a initial condition for $\bar{y}$ and a terminal condition for $\bar{p}$. Nevertheless it fulfills the definition of a Hamiltonian System as there is an Hamiltonian with $M\bar{y}_t = -H_p$ and $M\bar{p}_t = H_y$.*

Remark 5.20. *The choice of the Hamiltonian is not unique as*

$$M\bar{p}_t = -\tilde{H}_y = A^*\bar{p} + \beta M_d \bar{y} - \beta M_d y_d \tag{5.11}$$

$$M\bar{y}_t = \tilde{H}_p = -A\bar{y} + \frac{1}{\nu} G M_u^{-1} G^* \bar{p}, \tag{5.12}$$

is also a Hamiltonian system with with the Hamiltonian

$$\tilde{H}(y,p) = -H(y,p). \tag{5.13}$$

In the following two subsections we will discuss conditions with which the optimality conditions are equivalent to a single equations. For Hamiltonian systems in mechanics this is well known, the Hamiltonian mechanics considers first order equations for state and velocity of the system, whereas Lagrangian mechanics considers a system of second order equations for the state.

5.3. Single equations for the state or the adjoint state

In the case that $\beta \neq 0$ and that the operator M_d is invertible, we use the second equation of the Hamiltonian system (5.9) as definition of the optimal state

$$M_d \bar{y} = \frac{1}{\beta} M\bar{p}_t - \frac{1}{\beta} A^*\bar{p} + M_d y_d \tag{5.14}$$

and insert this into the first equation (5.8) to get

$$\begin{aligned} -\frac{1}{\beta} M M_d^{-1} M\bar{p}_{tt} + \frac{1}{\beta} M M_d^{-1} A^*\bar{p}_t - \frac{1}{\beta} A M_d^{-1} M\bar{p}_t + \frac{1}{\beta} A M_d^{-1} A^*\bar{p} + \frac{1}{\nu} G M_u^{-1} G^*\bar{p} \\ = M y_{d,t} + A y_d. \end{aligned}$$

To assure that this is still a valid equation in V^* we need the additional regularity $\bar{p}_{tt}(\cdot,t) \in V^*$, $\bar{p}_t(\cdot,t) \in V$, $A^*\bar{p}(\cdot,t) \in V$, $y_d(\cdot,t) \in V$ and $y_{d,t} \in V^*$ instead of the regularity $\bar{p}(\cdot,t) \in V$, $A^*p(\cdot,t) \in V^*$, $\bar{p}_t(\cdot,t) \in V$ and $y_d(\cdot,t) \in H$ which is implied by the optimality conditions. In the case that the control costs and the observation are measured with the same operator M which is used in the differential equation, i.e. $M_u = M_d = M$, the equation simplifies to

$$-\frac{1}{\beta}M\bar{p}_{tt} + \frac{1}{\beta}A^*\bar{p}_t - \frac{1}{\beta}A\bar{p}_t + \frac{1}{\beta}AM^{-1}A^*\bar{p} + \frac{1}{\nu}GM^{-1}G^*\bar{p} = My_{d,t} + Ay_d.$$

And for a self adjoint operators A this equation is

$$-\frac{1}{\beta}M\bar{p}_{tt} + \frac{1}{\beta}AM^{-1}A\bar{p} + \frac{1}{\nu}GM^{-1}G^*\bar{p} = My_{d,t} + Ay_d. \tag{5.15}$$

On the other hand, if the product of operators $GM_u^{-1}G^*$ is invertible, we can also use the first equation (5.8) as definition of the optimal adjoint state

$$\bar{p} = \nu(GM_u^{-1}G^*)^{-1}M\bar{y}_t + \nu(GM_u^{-1}G^*)^{-1}A\bar{y} \tag{5.16}$$

and insert this into the second equation (5.9). This yields

$$\begin{aligned}-\nu M(GM_u^{-1}G^*)^{-1}M\bar{y}_{tt} - \nu M(GM_u^{-1}G^*)^{-1}A\bar{y}_t \\ +\nu A^*(GM_u^{-1}G^*)^{-1}M\bar{y}_t + \nu A^*(GM_u^{-1}G^*)^{-1}A\bar{y} + \beta M_d\bar{y} = \beta M_d y_d.\end{aligned}$$

For a self adjoint operator A and $G = M_u = M_d = M$ this is

$$-\nu M\bar{y}_{tt} + \nu AM^{-1}A\bar{y} + \beta M\bar{y} = \beta My_d. \tag{5.17}$$

For the solution of the equations (5.15) or (5.17) we still need to specify boundary conditions. The boundary conditions for the temporal boundaries $t = 0$ and $t = T$ for the adjoint state (5.15) are

$$\begin{aligned}M\bar{p}(x,T) &= \alpha M_D\left(\bar{y}(x,T) - y_D(x)\right) && \text{in } \Omega, \\ M\bar{p}_t(x,0) - A\bar{p}(x,0) &= \beta\left(v(x) - y_d(0,x)\right) && \text{in } \Omega.\end{aligned}$$

The first condition is the terminal condition for the adjoint state $\bar{p}$ and the second equation is just the adjoint equation for $t = 0$ together with the initial condition $My(0,x) = Mv(0)$.

For the equation (5.17) we have for the temporal boundary the conditions

$$\begin{aligned}M\bar{y}(x,0) &= Mv(x) && \text{in } \Omega, \\ \nu M\bar{y}_t(x,T) + \nu A\bar{y}(x,T) + \alpha M_D\bar{y}(\cdot,T) &= \alpha M_D y_D(x) && \text{in } \Omega.\end{aligned}$$

The first condition is the initial condition for $\bar{y}$. For the second condition there are two interpretations. On the one hand it is the state equation $M\bar{y}_t + A\bar{y} = \frac{1}{\nu}M\bar{p}$ together with the definition of the adjoint state $\bar{p}$ given by equation (5.14). And on the other hand the condition can be obtained by the terminal condition for the adjoint state

If the operator A is a partial differential operator on the spatial domain Ω we need additional boundary conditions on the spatial boundary for the equations (5.15) or (5.17).

Example 5.21. *Assume that the operator A is a self-adjoint second order elliptic operator with the boundary conditions*

$$\bar{y} = 0 \text{ on } \Sigma_1, \qquad \frac{\partial \bar{y}}{\partial n_A} = 0 \text{ on } \Sigma_2, \tag{5.18}$$

where $\Sigma_1 = \Gamma_1 \times (0,T)$ and $\Sigma_2 = \Gamma_2 \times (0,T)$ with $\overline{\Gamma_1} \cup \overline{\Gamma_2} = \partial\Omega$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$. Then it is well known that the adjoint state has the boundary conditions

$$\bar{p} = 0 \text{ on } \Sigma_1, \qquad \frac{\partial \bar{p}}{\partial n_A} = 0 \text{ on } \Sigma_2. \tag{5.19}$$

In the case that the operator A is a second order elliptic operator the operator product $AM^{-1}A$ in the equation (5.15) *and* (5.17) *is a fourth order operator. Therefore we need to specify two boundary conditions on every spatial boundary.*

For the equation (5.15) *for the adjoint state $\bar{p}$ the first condition is given as the boundary condition* (5.19) *of the original adjoint state. For the second set of boundary conditions we can use the boundary conditions for the state* (5.18) *together with the definition of the state by the adjoint state given by the equation* (5.14)*.*

$$\begin{aligned} 0 = M\bar{y} = \frac{1}{\beta} M\bar{p}_t - \frac{1}{\beta} Ap + My_d \qquad & \text{on } \Sigma_1, \\ 0 = \frac{\partial M\bar{y}}{\partial n_A} = \frac{1}{\beta}\frac{\partial M\bar{p}_t}{\partial n_A} - \frac{1}{\beta}\frac{\partial A\bar{p}}{\partial n_A} + \frac{\partial My_d}{\partial n_A} \qquad & \text{on } \Sigma_2. \end{aligned}$$

Due to the first set of boundary conditions (5.19) *we have $M\bar{p}_t = 0$ on Σ_1 and $\frac{\partial M\bar{p}_t}{\partial n_A} = 0$ on Σ_2. Therefore the second boundary condition simplifies to*

$$A\bar{p} = \beta My_d \qquad \text{on } \Sigma_1, \tag{5.20}$$

$$\frac{\partial A\bar{p}}{\partial n_A} = \frac{\partial My_d}{\partial n_A} \qquad \text{on } \Sigma_2. \tag{5.21}$$

Similarly in this case we need also to specify two boundary conditions on every spatial boundary for the equation (5.17) *for the state $\bar{y}$. Again, the first set of boundary conditions is given by the boundary conditions* (5.18) *for the state y. For the second set of conditions we use the boundary conditions for the adjoint state* (5.19) *together with the definition of the adjoint state by equation* (5.16)*, which implies*

$$\begin{aligned} 0 = \frac{1}{\nu} M\bar{p} = M\bar{y}_t + A\bar{p} \qquad & \text{on } \Sigma_1 \\ 0 = \frac{1}{\nu} M\frac{\partial \bar{p}}{\partial n_A} = M\frac{\partial \bar{y}_t}{\partial n_A} + \frac{\partial \bar{A}\bar{p}}{\partial n_A} \qquad & \text{on } \Sigma_2. \end{aligned}$$

Due to the homogeneous boundary conditions (5.18) *we have $\bar{y}_t = 0$ on Σ_1 and $\frac{\partial \bar{y}_t}{\partial n_A} = 0$ on Σ_2 and therefore the second set of boundary conditions on the spatial boundary is*

$$A\bar{y} = 0 \text{ on } \Sigma_1, \qquad \frac{\partial A\bar{y}}{\partial n_A} = 0 \text{ on } \Sigma_2. \tag{5.22}$$

Altogether we have seen that the solution of the optimal control problem is equivalent to the solution of one of the following two problems.

Problem 5.22 ($H^{(2,1)}(Q)$-elliptic Problem for $\bar{p}$)**.** *Assume that the operator A is a self-adjoint second order elliptic operator with boundary conditions as in Example 5.21 and $M_u = M_D = M_d = G = M$. Then the optimal adjoint state is given as the solution of the following equation together with the boundary conditions*

$$\begin{aligned}
-M\bar{p}_{tt} + AM^{-1}A\bar{p} + \frac{\beta}{\nu}GM^{-1}G^*\bar{p} &= My_{d,t} + Ay_d && \text{in } Q,\\
\frac{\alpha}{\beta}M\bar{p}_t(x,T) + M\bar{p}(x,T) - \frac{\alpha}{\beta}A^*\bar{p}(x,T) &= \alpha My_D(x) - \alpha My_d(x,T) && \text{for } x \in \Omega,\\
M\bar{p}_t(x,0) - A\bar{p}(x,0) &= \beta M(v(x) - y_d(0,x)) && \text{for } x \in \Omega,\\
\bar{p} &= 0 && \text{on } \Sigma_1,\\
\frac{\partial \bar{p}}{\partial n_A} &= 0 && \text{on } \Sigma_2,\\
A\bar{p} &= \beta My_d && \text{on } \Sigma_1,\\
\frac{\partial A\bar{p}}{\partial n_A} &= \frac{\partial My_d}{\partial n_A} && \text{on } \Sigma_2,
\end{aligned}$$

Problem 5.23 ($H^{(2,1)}(Q)$-elliptic Problem for $\bar{y}$)**.** *Assume that the operator A is a self-adjoint second order elliptic operator with boundary conditions as in Example 5.21 and $M_u = M_D = M_d = G = M$. Then the optimal state is given as the solution of the following equation together with the boundary conditions*

$$\begin{aligned}
-M\bar{y}_{tt} + AM^{-1}A\bar{y} + \frac{\beta}{\nu}M\bar{y} &= \frac{\beta}{\nu}My_d && \text{in } Q,\\
M\bar{y}(x,0) &= Mv(x) && \text{for } x \in \Omega,\\
M\bar{y}_t(x,T) + A\bar{y}(x,T) + \frac{\alpha}{\nu}M_D\bar{y}(x,T) &= \frac{\alpha}{\nu}M_D y_D(x) && \text{for } x \in \Omega,\\
\bar{y} &= 0 && \text{on } \Sigma_1,\\
\frac{\partial \bar{y}}{\partial n_A} &= 0 && \text{on } \Sigma_2,\\
A\bar{y} &= 0 && \text{on } \Sigma_1,\\
\frac{\partial A\bar{y}}{\partial n_A} &= 0 && \text{on } \Sigma_2.
\end{aligned}$$

Remark 5.24. *For the case of an self adjoint operator A and the choice of $G = M$ we see that the optimal state $\bar{y}$ and the optimal adjoint state $\bar{p}$ fulfill the same differential equation with different right hand sides and different boundary conditions. Problem 5.23 is well posed for a desired state $y_d \in L^2(Q)$, whereas we need $y_d \in L^2(0,T;H^1(\Omega))$ for Problem 5.22. If the operator A is an elliptic differential operator of order two, this equation is of second order in time and fourth order in space.*

Remark 5.25. *The elimination of the state in the optimality conditions is also discussed in [24] and [86]. They start by taking the time derivative of the state equation.*

Remark 5.26. *Due to the inversion of the operator M in the product of operators $AM^{-1}A$ in the equations for the state* (5.17) *and the adjoint state* (5.15) *it is quite natural to formulate these equations as mixed problems. A three field problem for the state is given by the equations*

$$\left.\begin{aligned} A\bar{y} &= M\bar{z}, \\ M\bar{y} &= M\bar{w}, \\ -\nu M\bar{y}_{tt} + \nu A\bar{z} + \beta M\bar{w} &= \beta M y_d \end{aligned}\right\} \tag{5.23}$$

together with the boundary conditions of Problem 5.23 and a mixed system for the adjoint state is given by

$$\left.\begin{aligned} A\bar{p} &= M\bar{q}, \\ \frac{1}{\nu}G^*\bar{p} &= M\bar{u}, \\ -M\bar{p}_{tt} + A\bar{q} + \beta G\bar{u} &= \beta M y_{d,t} + \beta A y_d \end{aligned}\right\} \tag{5.24}$$

together with the boundary conditions of Problem 5.22.

We consider the systems (5.23) *and* (5.24) *as three field problems, so we can discretize the functions $\bar{y}$ and $\bar{w}$ and the functions $\bar{u}$ and $\bar{p}$ differently.*

Gong, Hinze and Zhou [46] introduce mixed systems similar to (5.24) *and* (5.24)*. But they only consider two field problems, as they do not introduce the functions, which we have called $\bar{w}$ and $\bar{u}$.*

5.4. Summary

In this chapter we discussed the continuous optimality conditions for optimal control problems with parabolic partial differential equations.

But for the solution of the most non-trivial optimal control problem we need to discretize the problem. This discretization can take place at different stages:

1. One can discretize the optimal control Problem 5.2 and solve the resulting optimization problem. This approach is called discretize-then-optimize and is discussed in the next Chapter.

2. One can discretize the continuous optimality conditions of Problem 5.12 and solve the resulting system. This approach is called optimize-then-discretize and is also discussed in the next Chapter.

3. One can discretize the $H^{(2,1)}(Q)$-elliptic Problem 5.23 or 5.22 for a solution of the optimal state $\bar{y}$ or the optimal adjoint state $\bar{p}$ or the optimal control $\bar{u}$. As we have discussed conforming finite element approximations of $H^{(2,1)}(Q)$ semi-elliptic equations in Section 4.1.2, the convergence properties of these discretizations for these problems are also clear.

4. One can discretize the mixed formulation of the $H^{(2,1)}(Q)$-elliptic problem of Remark 5.26. The equivalence of the mixed problem for the state y (5.23) to a discretize-then-optimize and a optimize-then-discretize approach is discussed in Section 7.2.

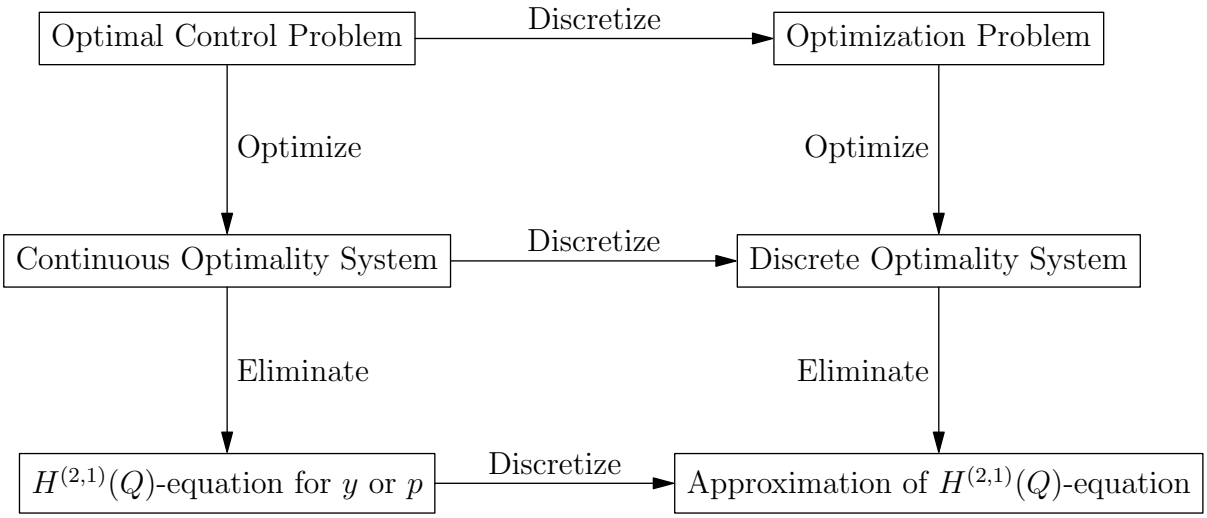


Figure 5.1.: Possibilities of optimization and discretization for optimal control problems with parabolic partial differential equations.

In the remaining part of this thesis we discuss the discretization at different levels and whether the discretization and optimization at the different level result in the same discrete scheme after optimization and discretization, i.e. whether the different paths in Figure 5.1 result in the same discrete scheme or not.

6. Crank-Nicolson and Störmer-Verlet schemes for parabolic optimal control problems

Contents

In the previous chapter we have introduced an abstract parabolic optimal control problem (5.1) and derived the corresponding optimality conditions. In this chapter we want to discretize this problem, compute an approximation of the solution and prove error bounds. We follow the ideas of [4, 5] and generalize these ideas to the optimal control problem (5.1).

For the discretization there are two approaches very common. First we can discretize (5.1) directly and solve a linear-quadratic optimization problem. This is the first-discretize-then-optimize approach. Second we can discretize the optimality conditions (5.7) and discretize the Hamiltonian systems of differential equations. This approach is called first-optimize-then-discretize.

The first and the second approach are very common. Furthermore one would prefer an approach where the resulting discrete system for this two approaches coincides. In first-discretize-then-optimize approaches the discretization of the adjoint equation is given by the choice of the discretization of the state equation and the cost functional. One has no influence if this is an appropriate discretization of the continuous adjoint equation. If one uses the first-optimize-then-discretize approach one can freely choose discretizations for the state and the adjoint state. But this could result in a scheme where the overall solution operator is neither symmetric nor positive definite. Further the gradient equation can contain matrices of dimension $\mathbb{R}^{n_1 \times n_2}$ with $n_1 \neq n_2$.

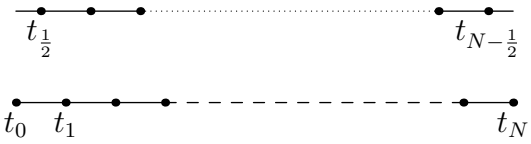


Figure 6.1.: Comparison of the time grids for the discretization of y and p. First line p, second line y.

6.1. Discretize then optimize

We start with the time discretization of the optimal control Problem 5.2, which consists of the equations

$$\begin{aligned} &\min_{y,u} J(y,u), \\ \text{s.t. } &My_t + Ay = Gu, \\ &My(0,\cdot) = Mv(\cdot), \end{aligned}$$

with

$$\begin{aligned} J(y,u) = \frac{\alpha}{2}\left\|M_D^{1/2}\left(y(\cdot,T) - y_D(\cdot)\right)\right\|_H^2 + \frac{\beta}{2}\int_0^T \left\|M_d^{1/2}\left(y(\cdot,t) - y_d(\cdot,t)\right)\right\|_H^2 \mathrm{d}\,t + \\ + \frac{\nu}{2}\int_0^T \left\|M_u^{1/2}u\right\|_H^2 \mathrm{d}\,t. \end{aligned}$$

We focus on the temporal discretization of the equations, for the full discretization one has to replace all the operators A, G, M, M_d, M_D and M_u and the spaces by its discrete counterparts, e.g. finite element matrices and $\mathbb{R}^n$ instead of the space H and V. As introduced in Section 4.2 the Crank-Nicolson scheme for the state equation of the optimal control Problem 5.2 reads

$$\langle M\frac{y_{k+1}-y_k}{\tau}, \varphi\rangle_{H\times H} + \langle A\frac{y_{k+1}+y_k}{2}, \varphi\rangle_{V^*\times V} = \langle G\tilde{u}_{k+\frac{1}{2}}, \varphi\rangle_{H\times H}, \tag{6.1}$$

where we discretize y in the grid points t_k. For the choice of $\tilde{u}_{k+\frac{1}{2}}$ different possibilities exist. With $\tilde{u}_{k+\frac{1}{2}} = u(t_{k+\frac{1}{2}})$ we obtain the midpoint rule.

For the discretization of the optimal control Problem 5.2 not only the differential equation but also the cost functional has to be discretized. In view of (6.1) we discretize the state y again in the grid points t_k and u in the midpoints $t_{k+\frac{1}{2}}$ of the time intervals (see Figure 6.1). A discretization of the cost functional is given by

$$\begin{aligned} &\frac{\beta\tau}{4}\left\|M_d^{1/2}\left(y_0 - y_{d,0}\right)\right\|_H^2 + \frac{\beta\tau}{2}\sum_{k=1}^{N-1}\left\|M_d^{1/2}\left(y_k - y_{d,k}\right)\right\|_H^2 + \frac{\beta\tau}{2}\left\|M_d^{1/2}\left(y_N - y_{d,N}\right)\right\|_H^2 + \\ &+ \frac{\alpha}{2}\left\|M_D^{1/2}(y_N - y_D)\right\|_H^2 + \frac{\tau\nu}{2}\sum_{k=0}^{N-1}\left\|M_u^{1/2}u_{k+\frac{1}{2}}\right\|_H^2, \end{aligned}$$

where the trapezoidal rule is used for the discretization of the first integral with $y_{d,k} = y_d(t_k)$ and the midpoint rule for the second integral. This choice seems quite natural as for the

discretization of the state equation we need only values of y, y_d and u at these points. Together with the discretization of the differential equation

$$M\frac{y_{k+1}-y_k}{\tau}+A\frac{y_{k+1}+y_k}{2}=Gu_{k+\frac{1}{2}},$$
$$My_0=Mv$$

we obtain our first discretization. To solve this linear-quadratic optimization problem we form a Lagrange functional as

$$\begin{aligned}\mathcal{L}(\mathbf{y},\mathbf{u},\mathbf{p}) &= \frac{\beta\tau}{4}\left\|M_d^{1/2}(y_0-y_{d,0})\right\|_H^2+\frac{\beta\tau}{2}\sum_{k=1}^{N-1}\left\|M_d^{1/2}(y_k-y_{d,k})\right\|_H^2+\\ &+\frac{\beta\tau}{4}\left\|M_d^{1/2}(y_N-y_{d,N})\right\|_H^2+\frac{\alpha}{2}\left\|M_D^{1/2}(y_N-y_D)\right\|_H^2+\frac{\tau\nu}{2}\sum_{k=0}^{N-1}\left\|M_u^{1/2}u_{k+\frac{1}{2}}\right\|_H^2+\\ &+\langle M(y_0-v),p_0\rangle_{H\times H}+\tau\sum_{k=0}^{N-1}\langle M\frac{y_{k+1}-y_k}{\tau}+A\frac{y_{k+1}+y_k}{2}-Gu_{k+\frac{1}{2}},p_{k+\frac{1}{2}}\rangle_{H\times H}\\ \text{with } \mathbf{y}&=(y_1,\dots,y_N)^T,\ \mathbf{u}=\left(u_{\frac{1}{2}},\dots,u_{N-\frac{1}{2}}\right)^T \text{ and } \mathbf{p}=\left(p_0,p_{\frac{1}{2}},\dots,p_{N-\frac{1}{2}}\right)^T,\end{aligned}$$

with the Lagrange multipliers $\mathbf{p}$. The choice of the indices $\cdot_{i+\frac{1}{2}}$ for the Lagrange multiplier is motivated by the continuous optimality conditions (5.7). The Lagrange multipliers are only used to determine the optimal control. From this point of view it seems quite natural to discretize the control and adjoint state in the same way. This choice of discretization will be important and is essential if we analyze the discretization of the adjoint state later on. Other authors, as in [11, 103], who discussed Crank-Nicolson or the corresponding continuous Galerkin time stepping schemes did not use this discretization of the adjoint state and were therefore not able to proof second order convergence.

Our discretization was introduced in [4, 5] and also used by other authors [70, 84].

We solve the first order necessary conditions for the optimal solution $(\bar{y},\ \bar{u},\ \bar{p})$

$$\begin{aligned}\frac{\partial\mathcal{L}(\bar{\mathbf{y}},\bar{\mathbf{u}},\bar{\mathbf{p}})}{\partial y_i}&=0 && \text{for } i=0,\dots,N,\\ \frac{\partial\mathcal{L}(\bar{\mathbf{y}},\bar{\mathbf{u}},\bar{\mathbf{p}})}{\partial p_0}&=0 &&\\ \frac{\partial\mathcal{L}(\bar{\mathbf{y}},\bar{\mathbf{u}},\bar{\mathbf{p}})}{\partial p_{i+\frac{1}{2}}}&=0 && \text{for } i=0,\dots,N-1,\\ \frac{\partial\mathcal{L}(\bar{\mathbf{y}},\bar{\mathbf{u}},\bar{\mathbf{p}})}{\partial u_{i+\frac{1}{2}}}&=0 && \text{for } i=0,\dots,N-1.\end{aligned}$$

Note further that we discuss a convex cost functional such that the necessary first order

optimality conditions are sufficient, too. The resulting system is the weak form of

$$\left.\begin{aligned} M\bar{y}_0 &= Mv, \\ M\frac{\bar{y}_{i+1}-\bar{y}_i}{\tau} + A\frac{\bar{y}_{i+1}+\bar{y}_i}{2} &= G\bar{u}_{i+\frac{1}{2}} && \text{for } i=0,\dots,N, \\ \nu M_u \bar{u}_{i+\frac{1}{2}} &= G^*\bar{p}_{i+\frac{1}{2}} && \text{for } i=0,\dots,N-1, \\ M\frac{\bar{p}_{\frac{1}{2}}-\bar{p}_0}{\tau} - A^*\frac{\bar{p}_{\frac{1}{2}}}{2} &= \beta M_d \frac{\bar{y}_0 - y_{d,0}}{2}, \\ M\frac{\bar{p}_{i+\frac{1}{2}}-\bar{p}_{i-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{i+\frac{1}{2}}+\bar{p}_{i-\frac{1}{2}}}{2} &= \beta M_d\left(\bar{y}_i - y_{d,i}\right) && \text{for } i=0,\dots,N-2, \\ -M\frac{\bar{p}_{N-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{N-\frac{1}{2}}}{2} &= \beta M_d \frac{\bar{y}_N - y_{d,N}}{2} \\ &\quad + \alpha M_D \frac{\bar{y}_N - \bar{y}_D}{\tau}. \end{aligned}\right\} \quad \text{(OC CN1)}$$

For the convergence analysis later on we will set

$$-M\bar{p}_N = \alpha M_D\left(\bar{y}_N - \bar{y}_D\right)$$

and replace the last half step by

$$\left.\begin{aligned} M\frac{\bar{p}_N - \bar{p}_{N-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{N-\frac{1}{2}}}{2} &= \beta M_d \frac{\bar{y}_N - y_{d,N}}{2}, \\ -M\bar{p}_N &= \alpha M_D\left(\bar{y}_N - \bar{y}_D\right). \end{aligned}\right\} \quad \text{(OC CN1*)}$$

This approach is motivated by the terminal condition for the adjoint state in the continuous optimality conditions.

Remark 6.1. *At the beginning we had to choose a discretization of the cost functional. Another possible choice is the midpoint rule for both integrals in the cost functional. This gives the optimization problem*

$$\left.\begin{aligned} \min\ & \frac{\alpha}{2}\left\|M_D^{1/2}y_N - y_D\right\|_H^2 + \frac{\beta\tau}{2}\sum_{k=0}^{N-1}\left\|M_d^{1/2}\left(\frac{y_k+y_{k+1}}{2} - \frac{y_{d,k+1}+y_{d,k}}{2}\right)\right\|_H^2 + \\ & \qquad + \frac{\tau\nu}{2}\sum_{k=0}^{N-1}\left\|M_u^{1/2}u_{k+\frac{1}{2}}\right\|_H^2 \\ & M\frac{y_{k+1}-y_k}{\tau} + A\frac{y_{k+1}+y_k}{2} = Gu_{k+\frac{1}{2}} \\ & My_0 = Mv \end{aligned}\right\} \quad \text{(CN2)}$$

The corresponding Lagrange functional is

$$\begin{aligned} \mathcal{L}(\mathbf{y},\mathbf{u},\mathbf{p}) = {} & \frac{\alpha}{2}\left\|M_D^{1/2}y_N - y_D\right\|_H^2 + \frac{\beta\tau}{2}\sum_{k=0}^{N-1}\left\|M_d^{1/2}\left(\frac{y_k+y_{k+1}}{2} - \frac{y_{d,k+1}+y_{d,k}}{2}\right)\right\|_H^2 + \\ & + \frac{\tau\nu}{2}\sum_{k=0}^{N-1}\left\|M_u^{1/2}u_{k+\frac{1}{2}}\right\|_H^2 + \langle M(y_0-v), p_0\rangle_{H\times H} \\ & + \tau\sum_{k=0}^{N-1}\langle M\frac{y_{k+1}-y_k}{\tau} + A\frac{y_{k+1}+y_k}{2} - Gu_{k+\frac{1}{2}}, p_{k+\frac{1}{2}}\rangle_{H\times H} \end{aligned}$$

and the first order optimality conditions are

$$\left.\begin{aligned}
M\bar{y}_0 &= Mv,\\
M\frac{\bar{y}_{i+1}-\bar{y}_i}{\tau} + A\frac{\bar{y}_{i+1}+\bar{y}_i}{2} &= G\bar{u}_{i+\frac{1}{2}}\\
&\textit{for } i = 0,\dots,N-1,\\
\nu M_u \bar{u}_{i+\frac{1}{2}} &= G^*\bar{p}_{i+\frac{1}{2}}\\
&\textit{for } i = 0,\dots,N-1,\\
M\frac{\bar{p}_{\frac{1}{2}}-\bar{p}_0}{\tau} - A^*\frac{\bar{p}_{\frac{1}{2}}}{2} &= \beta M_d \frac{\frac{\bar{y}_0+\bar{y}_1}{2} - \frac{y_{d,0}+y_{d,1}}{2}}{2},\\
M\frac{\bar{p}_{i+\frac{1}{2}}-\bar{p}_{i-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{i+\frac{1}{2}}+\bar{p}_{i-\frac{1}{2}}}{2} &=\\
= \beta M_d \frac{\frac{\bar{y}_i+\bar{y}_{i-1}}{2} - \frac{y_{d,i-1}+y_{d,i}}{2}}{2} &+ \beta M_d \frac{\frac{\bar{y}_i+\bar{y}_{i+1}}{2} - \frac{y_{d,i+1}+y_{d,i}}{2}}{2}\\
&\textit{for } i = 1,\dots,N-2,\\
-M\frac{\bar{p}_{N-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{N-\frac{1}{2}}}{2} &= \beta M_d \frac{\frac{\bar{y}_{N-1}+\bar{y}_N}{2} - \frac{y_{d,N}+y_{d,N-1}}{2}}{2} +\\
&\quad + \alpha M_D \frac{\bar{y}_N - y_D}{\tau}.
\end{aligned}\right\} \qquad \text{(OC CN2)}$$

The difference in both system is the discretization of the right hand side of the adjoint system. The right hand side of the adjoint equation of (OC CN2)*,*

$$\frac{\frac{\bar{y}_i+\bar{y}_{i-1}}{2} - \frac{y_{d,i+1}+y_{d,i}}{2}}{2} + \frac{\frac{\bar{y}_i+\bar{y}_{i+1}}{2} - \frac{y_{d,i+1}+y_{d,i}}{2}}{2},$$

can be interpreted as averaged approximation of $\bar{y}_i - y_{d,i}$ (see Figure 6.2). It is also for this scheme possible to introduce an approximation of the adjoint state at the terminal time T by

$$-M\bar{p}_N = \alpha M_D \bar{y}_N - \bar{y}_D$$

and replace the last half step by

$$\left.\begin{aligned}
M\frac{\bar{p}_N-\bar{p}_{N-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{N-\frac{1}{2}}}{2} &= \beta M_d \frac{\frac{\bar{y}_{N-1}+\bar{y}_N}{2} - \frac{y_{d,N}+y_{d,N-1}}{2}}{2}\\
-M\bar{p}_N &= \alpha M_D\left(\bar{y}_N - \bar{y}_D\right).
\end{aligned}\right\} \qquad \text{(OC CN2*)}$$

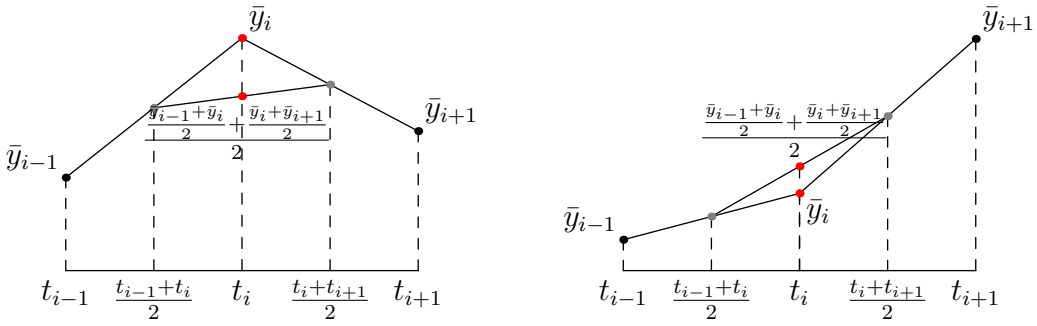


Figure 6.2.: $\frac{\frac{\bar{y}_{i-1}+\bar{y}_i}{2}+\frac{\bar{y}_i+\bar{y}_{i+1}}{2}}{2}$ vs $\bar{y}_i$

6.2. Optimize then discretize

In the first section of this chapter we discretized the optimal control Problem 5.2 and discussed the optimality conditions for the discrete optimization problem. Now we discretize the continuous optimality conditions of Problem 5.12, which are

$$\begin{aligned}
M\bar{y}_t + A\bar{y} &= G\bar{u}, \\
M\bar{y}(\cdot,0) &= Mv(\cdot), \\
M\bar{p}_t - A^*\bar{p} &= \beta M_d\left(\bar{y} - y_d\right), \\
M\bar{p}(\cdot,T) &= \alpha M_D\left(y_D(\cdot) - \bar{y}(\cdot,T)\right), \\
M_u\bar{u} &= \frac{1}{\nu}G^*\bar{p}.
\end{aligned}$$

In Section 5.2 we have seen that these optimality conditions form a Hamiltonian system. Therefore we apply the Störmer-Verlet scheme (SV) of Section 4.3 (on page 60)

$$\begin{aligned}
M\bar{y}_0 &= Mv, \\
M\frac{\bar{y}_{i+1}-\bar{y}_i}{\tau} + A\frac{\bar{y}_{i+1}+\bar{y}_i}{2} &= G\bar{u}_{i+\frac{1}{2}} && \text{for } i = 0,\dots,N, \\
\nu M_u\bar{u}_{i+\frac{1}{2}} &= G^*\bar{p}_{i+\frac{1}{2}} && \text{for } i = 0,\dots,N-1, \\
M\frac{\bar{p}_{\frac{1}{2}}-\bar{p}_0}{\tau} - A^*\frac{\bar{p}_{\frac{1}{2}}}{2} &= \beta M_d\frac{\bar{y}_0 - y_{d,0}}{2}, \\
M\frac{\bar{p}_{i+\frac{1}{2}}-\bar{p}_{i-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{i+\frac{1}{2}}+\bar{p}_{i-\frac{1}{2}}}{2} &= \beta M_d\left(\bar{y}_i - y_{d,i}\right) && \text{for } i = 0,\dots,N-2, \\
M\frac{\bar{p}_N-\bar{p}_{N-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{N-\frac{1}{2}}}{2} &= \beta M_d\frac{\bar{y}_N - y_{d,N}}{2} \\
M\bar{p}_N &= \alpha M_D\left(\bar{y}_D - \bar{y}_N\right)
\end{aligned}$$

to these conditions and observe that this system is the system (OC CN1) with terminal step (OC CN1*).

Remark 6.2. *The Störmer-Verlet scheme is a symplectic partitioned Runge-Kutta scheme. In [16, 17] Bonnans and Laurent-Varin discuss the application of such schemes to optimal control*

problems with ordinary differential equations. They use a slightly different Hamiltonian and prove order conditions. This implies that the scheme (OC CN1) *is an second order scheme, if we have sufficient the regularity. For the convergence proof of the Störmer-Verlet scheme with techniques known from ordinary differential we need* $\mathcal{C}^3$*-regularity with respect to the time variable. As this regularity assumption is rather high in the case of time dependent partial differential equations, we will discuss a convergence proof with less regularity later on. But first we see, that we can also get the scheme* (OC CN2) *with an optimize-then-discretize approach.*

6.3. Galerkin method

To show, that we can reach the scheme (OC CN2) with an optimize-then-discretize approach we use a Galerkin method. Galerkin methods are also popular discretizations of evolution equations. We start with the continuous Lagrange functional (5.2) and the corresponding optimality conditions (5.3), (5.4) and (5.5) given by the first variation of the Lagrange functional set to zero

$$\begin{aligned}
\int_0^T \langle M\bar{y}_{,t} + A\bar{y} - G\bar{u}, \varphi\rangle_{V^*\times V}\,\mathrm{d}\,t + \langle M\left(y(\cdot,0) - v, \varphi(\cdot,0)\right)\rangle_{H\times H} &= 0 \qquad \forall \varphi \in \mathcal{P},\\
\int_0^T \nu\langle M_u\bar{u}, \psi\rangle_{H\times H} - \langle \psi, G^*\bar{p}\rangle_{V^*\times V} &= 0 \qquad \forall \psi \in \mathcal{P},\\
\alpha\left\langle M_D\left(\bar{y}(\cdot,T) - y_D(\cdot)\right), \phi\right\rangle_{H\times H} + \beta\int_0^T \left\langle M_d\left(\bar{y} - y_d\right), \phi\right\rangle_{V^*\times V}\,\mathrm{d}\,t&\\
+ \int_0^T \langle M\phi_{,t} + A\phi, \bar{p}\rangle_{V^*\times V}\,\mathrm{d}\,t + \langle M\phi(\cdot,0), p(\cdot,0)\rangle_{H\times H} &= 0 \qquad \forall \phi \in \mathcal{Y}.
\end{aligned}$$

As the Lagrangian, this version of the optimality conditions is well defined for the state $y \in \mathcal{Y}$ and the adjoint state $p \in \mathcal{P}$. So we need no additional regularity assumption and therefore we avoid the partial integration in time direction for the adjoint equation. For the discretization we choose test functions

$$\phi_i \in \mathcal{Y}_1 = \left\{\phi \in \mathcal{Y} :\ \phi|_{t\in(t_i,t_{i+1})} \in \mathbb{P}^1(t_i, t_{i+1}, V)\right\}$$

and

$$\psi_{i+\frac{1}{2}},\ \varphi_{i+\frac{1}{2}} \in \mathcal{P}_0 = \left\{\phi \in \mathcal{P} :\ \phi|_{t\in(t_i,t_{i+1})} \in \mathbb{P}^0(t_i, t_{i+1}, V)\right\}$$

for the discretization, so that

$$\left.\begin{aligned}
\int_0^{t_1} -\langle M\phi_{1,t},\bar{p}\rangle_{V^*\times V} - \langle A^*\bar{p},\phi_1\rangle_{V^*\times V}\,\mathrm{d}\,t &= \\
\langle M\phi_1(\cdot,0),\bar{p}(\cdot,0)\rangle_{H\times H} &+ \beta\int_0^{t_1}\langle M_d(\bar{y}-y_d),\phi_1\rangle_{V^*\times V}\,\mathrm{d}\,t,\\
\int_{t_{i-1}}^{t_{i+1}} -\langle M\phi_{i,t},\bar{p}\rangle_{V^*\times V} - \langle A^*\bar{p},\phi_i\rangle_{V^*\times V}\,\mathrm{d}\,t &= \beta\int_{t_{i-1}}^{t_{i+1}}\langle M_d(\bar{y}-y_d),\phi_i\rangle_{V^*\times V}\,\mathrm{d}\,t\\
&\text{for } i=2,\dots,N-2,\\
\int_{t_{N-1}}^{t_N} -\langle M\phi_{N-1,t},\bar{p}\rangle_{V^*\times V} - \langle A^*\bar{p},\phi_{N-1}\rangle_{V^*\times V}\,\mathrm{d}\,t &= \\
\alpha\langle M_D(\bar{y}(\cdot,T)-y_D(\cdot),\phi_{N-1}\rangle_{H\times H} &+ \int_{t_{N-1}}^{t_N}\langle M(\bar{y}-y_d),\phi_{N-1}\rangle_{V^*\times V}\,\mathrm{d}\,t,\\
\langle \bar{M}y(\cdot,0)-Mv,\varphi_{-\frac{1}{2}}(\cdot,0)\rangle &= 0,\\
\int_{t_i}^{t_{i+1}}\langle M\bar{y}_{,t},\varphi_{i+\frac{1}{2}}\rangle_{V^*\times V} + \langle A\bar{y},\varphi_{i+\frac{1}{2}}\rangle_{V^*\times V}\,\mathrm{d}\,t &= \int_{t_i}^{t_{i+1}}\langle G\bar{u},\varphi_{i+\frac{1}{2}}\rangle_{V^*\times V}\,\mathrm{d}\,t\\
&\text{for } i=0,\dots,N-1,\\
\int_{t_i}^{t_{i+1}}\nu\langle M_u\bar{u},\psi_{i+\frac{1}{2}}\rangle_{V^*\times V}\,\mathrm{d}\,t &= \int_{t_i}^{t_{i+1}}\langle G^*\bar{p},\psi_{i+\frac{1}{2}}\rangle_{V^*\times V}\,\mathrm{d}\,t\\
&\text{for } i=0,\dots,N-1.
\end{aligned}\right\}\tag{6.2}$$

For the time discretization of these equations we need also to discretize the remaining functions. If we discretize with $y,\ y_d\in\mathcal{Y}_1,\ u,\ p\in\mathcal{P}_0$ and evaluate the integrals exactly (for the computation of the integrals see Appendix D), this yields the weak form of the system

$$\left.\begin{aligned}
M\bar{y}_0 &= Mv,\\
M\frac{\bar{y}_{i+1}-\bar{y}_i}{\tau} + A\frac{\bar{y}_{i+1}+\bar{y}_i}{2} &= G\bar{u}_{i+\frac{1}{2}}\\
&\text{for } i=1,\dots,N-1,\\
M\frac{\bar{p}_{\frac{1}{2}}-\bar{p}_0}{\tau} - A^*\frac{\bar{p}_{\frac{1}{2}}}{2} &= \frac{2}{6}\beta M_d\left(\bar{y}_0-y_{d,0}\right)+\frac{1}{6}\beta M_d\left(\bar{y}_1-y_{d,1}\right),\\
M\frac{\bar{p}_{i+\frac{1}{2}}-\bar{p}_{i-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{i+\frac{1}{2}}+\bar{p}_{i-\frac{1}{2}}}{2} &= \\
= \frac{1}{6}\beta M_d\left(\bar{y}_{i-1}-y_{d,i-1}\right)+\frac{4}{6}\beta M_d\left(\bar{y}_i-y_{d,i}\right)&+\frac{1}{6}\beta M_d\left(\bar{y}_{i+1}-y_{d,i+1}\right)\\
&\text{for } i=1,\dots,N-2,\\
-\frac{\bar{p}_{N-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{N-\frac{1}{2}}}{2} &= \\
= \frac{1}{6}\beta M_d\left(\bar{y}_{N-1}-y_{d,N-1}\right)+\frac{2}{6}\beta M_d\left(\bar{y}_N-y_{d,N}\right)&+\frac{\alpha}{\tau}M_D\left(\bar{y}_N-y_D\right),\\
G^*\bar{p}_{i+\frac{1}{2}} &= \nu M_u\bar{u}_{i+\frac{1}{2}}\\
&\text{for } i=1,\dots,N-1.
\end{aligned}\right\}\tag{OC G1}$$

We observe that the left hand side of this system coincides with the discretizations (OC CN1) and (OC CN2), but the right hand side of the adjoint equation is a different discretization, which is also a Crank-Nicolson discretization in the sense of Remark 4.21 .

We discuss now, whether it is possible to get a more convenient Crank-Nicolson discretization as Galerkin scheme. Therefore we have a closer look on the right hand side of the system (6.2) and (OC G1). The right hand side of the state equation is given by the integral

$$\int_{t_i}^{t_{i+1}} \langle G\bar{u}, \varphi_{i+\frac{1}{2}} \rangle_{V^* \times V} \,\mathrm{d}\, t.$$

With the chosen discretization this is an integral in time over the product of a piecewise constant function times a piecewise constant function. This choice of ansatz and test function just yields a convenient Crank-Nicolson discretization.

Also the discretization of the right hand side with a piecewise linear function with piecewise constant functions as test functions would yield a Crank-Nicolson discretization.

Whereas the integral on the right hand side of the adjoint equation

$$\int_{t_{i-1}}^{t_{i+1}} \langle M_d(\bar{y} - y_d), \phi_i \rangle_{V^* \times V} \,\mathrm{d}\, t$$

is an integral over a product of two piecewise linear functions in time. If we want to reach a more common Crank-Nicolson discretization for this equation we need to modify this equation so that we have an integral over the product of a piecewise constant times a piecewise linear function in time. As the integrals on the left hand sides yield an Crank-Nicolson discretization, we can not alter the test function without changing the discretization on the left hand side. But if we project $(y - y_d) \in \mathcal{Y}_1$ onto a piecewise constant function $z \in \mathcal{P}_0$, we have an integral in time over the product of a piecewise constant times a piecewise linear function.

To this end, we consider the algebraically equivalent optimal control problem

$$\begin{aligned} \min_u \frac{\alpha}{2} \left\| M_D^{1/2} \left(y(\cdot, T) - y_D \right) \right\|_H^2 + \frac{\beta}{2} \int_0^T \left\| M_d^{1/2} z \right\|_H^2 \,\mathrm{d}\, t + \int_0^T \frac{\nu}{2} \left\| M_u^{1/2} u \right\|_H^2 \,\mathrm{d}\, t, \\ \text{s.t. } M y_{,t} + A y = G u, \\ M y(0) = M v, \\ M z = M (y - y_d). \end{aligned}$$

This corresponds to the Lagrange functional

$$\begin{aligned} \tilde{\mathcal{L}}(y, z, u, p, q) = \frac{\alpha}{2} \left\| M_D^{1/2} \left(y(\cdot, T) - y_D \right) \right\|_H^2 + \frac{\beta}{2} \int_0^T \left\| M_d^{1/2} z \right\|_H^2 \,\mathrm{d}\, t + \int_0^T \frac{\nu}{2} \left\| M_u^{1/2} u \right\|_H^2 \,\mathrm{d}\, t \\ + \int_0^T \langle M y_{,t} + A y - G u, p \rangle_{V^* \times V} \,\mathrm{d}\, t + \langle M y(\cdot, 0) - M v, p(\cdot, 0) \rangle_{H \times H} \\ + \int_0^T \langle M q, y - y_d - z \rangle_{H \times H} \,\mathrm{d}\, t \end{aligned}$$

with an additional Lagrange multiplier $q \in L^2((0,T),H)$. This functional is well defined for $y \in \mathcal{Y}$, $y_d \in L^2((0,T),H)$ and $u,\ p,\ z \in \mathcal{P}$. The advantage is that we can discretize y and z differently as described above. The first order optimality conditions are

$$
\begin{aligned}
\left.\frac{\partial \tilde{\mathcal{L}}(\bar{y}+\varepsilon\phi,\bar{z},\bar{u},\bar{p},\bar{q})}{\partial\varepsilon}\right|_{\varepsilon=0} &= \alpha\langle M_D\left(y(\cdot,T)-y_D\right),\phi\rangle_{H\times H} \\
&\quad + \int_0^T \langle M\phi_t + A\phi, \bar{p}\rangle_{V^\star\times V} + \langle M\bar{q},\phi\rangle_{H\times H}\,\mathrm{d}\,t \\
&\quad + \langle M\phi(\cdot,0),\bar{p}(\cdot,0)\rangle_{H\times H} = 0 && \forall \phi\in\mathcal{Y},\\
\left.\frac{\partial \tilde{\mathcal{L}}(\bar{y},\bar{z}+\varepsilon\vartheta,\bar{u},\bar{p},\bar{q})}{\partial\varepsilon}\right|_{\varepsilon=0} &= \int_0^T \beta\langle M_d\bar{z},\vartheta\rangle_H - \langle M\bar{q},\vartheta\rangle_H\,\mathrm{d}\,t = 0 && \forall \vartheta\in\mathcal{P},\\
\left.\frac{\partial \tilde{\mathcal{L}}(\bar{y},\bar{z},\bar{u}+\varepsilon\psi,\bar{p},\bar{q})}{\partial\varepsilon}\right|_{\varepsilon=0} &= \int_0^T \langle \nu M_u\bar{u} - G^*\bar{p},\psi\rangle_{V^\star\times V}\,\mathrm{d}\,t = 0 && \forall \psi\in\mathcal{P},\\
\left.\frac{\partial \tilde{\mathcal{L}}(\bar{y},\bar{z},\bar{u},\bar{p}+\varepsilon\varphi,\bar{q})}{\partial\varepsilon}\right|_{\varepsilon=0} &= \int_0^T \langle M\bar{y}_t + A\bar{y} - G\bar{u},\varphi\rangle_{V^\star\times V}\,\mathrm{d}\,t + \\
&\quad + \langle M\bar{y}(\cdot,0) - Mv,\varphi(\cdot,0)\rangle_{H\times H} = 0 && \forall \varphi\in\mathcal{P},\\
\left.\frac{\partial \tilde{\mathcal{L}}(\bar{y},\bar{z},\bar{u},\bar{p},\bar{q}+\varepsilon\eta)}{\partial\varepsilon}\right|_{\varepsilon=0} &= \int_0^T \langle M\eta,\bar{y}-y_d-\bar{z}\rangle_{H\times H}\,\mathrm{d}\,t = 0 && \forall \eta\in\mathcal{P}.
\end{aligned}
$$

This system is equivalent to (6.2) which justifies the use of the formal Lagrange approach. If we choose for the discretization $\phi,\ y \in \mathcal{Y}_1$ and for all other functions the discretization space $\mathcal{P}_0$ we have to solve, after elimination of the additional variables z and q

$$
\left.
\begin{aligned}
M\bar{y}_0 &= Mv\\
M\frac{\bar{y}_{i+1}-\bar{y}_i}{\tau} + A\frac{\bar{y}_{i+1}-\bar{y}_i}{2} &= G\bar{u}_{i+\frac12}\\
&\text{for } i = 1,\dots,N-1,\\
M\frac{\bar{p}_{\frac12}-\bar{p}_0}{\tau} - A^*\frac{\bar{p}_{\frac12}}{2} &= \beta M_d\frac{\frac{\bar{y}_0+\bar{y}_1}{2} - \frac{y_{d,0}+y_{d,1}}{2}}{2}\\
M\frac{\bar{p}_{i+\frac12}-\bar{p}_{i-\frac12}}{\tau} - A^*\frac{\bar{p}_{i-\frac12}+\bar{p}_{i+\frac12}}{2} &= \beta M_d\frac{\frac{\bar{y}_{i-1}+\bar{y}_i}{2} - \frac{y_{d;i-1}+y_{d;i}}{2}}{2}\\
&\quad + \beta M_d\frac{\frac{\bar{y}_i+\bar{y}_{i+1}}{2} - \frac{y_{d;i}+y_{d;i+1}}{2}}{2}\\
&\text{for } i = 1,\dots,N-2,\\
-M\frac{\bar{p}_{N-\frac12}}{\tau} - A^*\frac{\bar{p}_{N-\frac12}}{2} &= \beta M_d\frac{\frac{\bar{y}_N+\bar{y}_{N-1}}{2} - \frac{y_{d,N}+y_{d,N-1}}{2}}{2}\\
&\quad + \frac{\alpha}{\tau}M_D\left(\bar{y}_N - y_D\right),\\
\nu M_u\bar{u}_{i+\frac12} &= G^*\bar{p}_{i+\frac12}\\
&\text{for } i = 1,\dots,N-1.
\end{aligned}
\right\} \qquad \text{(OC G2)}
$$

Hence, this approach is equivalent to (OC CN2).

Remark 6.3. *Another possible discretization of the equations* (6.2) *is obtained by again using* $y \in \mathcal{Y}_1$, $u,\ p \in \mathcal{P}_0$ *but the approximate evaluation of the integral with the midpoint rule for the intervals* $[t_{i-1}, t_i]$ *and* $[t_i, t_{i+1}]$. *This yields*

$$\begin{aligned}
M\bar{y}_0 &= Mv \\
M\frac{\bar{y}_{i+1}-\bar{y}_i}{\tau} + A\frac{\bar{y}_{i+1}-\bar{y}_i}{2} &= G\bar{u}_{i+\frac{1}{2}} \\
\text{for } i &= 1,\dots,N-1, \\
M\frac{\bar{p}_{\frac{1}{2}}-\bar{p}_0}{\tau} - A^*\frac{\bar{p}_{\frac{1}{2}}}{2} &= \beta M_d \frac{\frac{\bar{y}_0+\bar{y}_1}{2} - \frac{y_{d;0}+y_{d;1}}{2}}{2} \\
M\frac{\bar{p}_{i+\frac{1}{2}}-\bar{p}_{i-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{i-\frac{1}{2}}+\bar{p}_{i+\frac{1}{2}}}{2} &= \beta M_d \frac{\frac{\bar{y}_{i-1}+\bar{y}_i}{2} - \frac{y_{d;i-1}+y_{d;i}}{2}}{2} \\
&\quad + \beta M_d \frac{\frac{\bar{y}_i+\bar{y}_{i+1}}{2} - \frac{y_{d;i}+y_{d;i+1}}{2}}{2} \\
\text{for } i &= 1,\dots,N-2, \\
-M\frac{\bar{p}_{N-\frac{1}{2}}}{\tau} - A^*\frac{\bar{p}_{N-\frac{1}{2}}}{2} &= \beta M_d \frac{\frac{\bar{y}_N+\bar{y}_{N-1}}{2} - \frac{y_{d,N}+y_{d,N}}{2}}{2} + \frac{\alpha}{\tau} M_D\left(\bar{y}_N - y_D\right), \\
\nu M_u \bar{u}_{i+\frac{1}{2}} &= G^*\bar{p}_{i+\frac{1}{2}} \\
\text{for } i &= 1,\dots,N-1,
\end{aligned}$$

which is (OC CN2).

In summary, the Galerkin method with exact integration led to a new scheme, (OC G1), which can be interpreted as another variant of the Crank-Nicolson scheme. The Galerkin method with quadrature or projection reproduced scheme (OC CN2). For the scheme (OC CN1) we did not find a quadrature rule with which it is a Galerkin scheme.

6.4. Convergence analysis

Now we discuss the approximation properties of the numerical schemes. Therefore let us specify the conditions, for which the convergence proof is done.

Assumption 6.4.

- *Let the Assumptions 5.6 for parabolic optimal control problems hold.*
- *Let* $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$ *and* $H = L^2(\Omega)$.
- *Let* A *be a second order differential operator which fulfills Gårding's inequality* (3.20) *and additionally*

$$0 \le a(y, y).$$

- *Let* Ω *be convex and* $W = V \cap H^2(\Omega)$.
- *Let for the solution of the corresponding stationary problems*

$$a(y,\varphi) = \langle u, \varphi\rangle_{H\times H} \qquad \forall \varphi \in V, \qquad a(y_h,\varphi) = \langle u_h, \varphi\rangle_{H\times H} \qquad \forall \varphi \in V_h,$$

and the corresponding dual problems

$$a(\varphi, p) = \langle u, \varphi \rangle_{H\times H} \qquad \forall \varphi \in V, \qquad a(\varphi, p_h) = \langle u, \varphi \rangle_{H\times H} \qquad \forall \varphi \in V_h,$$

the error estimates

$$\|y - y_h\|_H \lesssim h^2 \|y\|_W , \qquad \|p - p_h\|_H \lesssim h^2 \|p\|_W$$

hold.

- *Let for the initial data $v \in W$ hold.*
- *Let for the exact solution $\bar{y}, \bar{p}, \bar{u} \in H^3((0,T), L^2(\Omega)) \cap H^2((0,T), H^2(\Omega))$ hold.*
- *Let*

$$C_1(\bar{y}, \bar{p}) = \|v\|_{H^2(\Omega)} + \int_0^T \|\bar{y}_{,t}(\cdot, s)\|_{W)} + \|\bar{p}_{,t}(\cdot, s)\|_W \,\mathrm{d}\, s \tag{6.3}$$

$$\left.\begin{aligned} C_2(\bar{y}, \bar{p}) = &\int_0^T \|\bar{y}_{,ttt}(\cdot, s)\|_H + \|A\bar{y}_{,tt}(\cdot, s)\|_H \,\mathrm{d}\, s \\ &+ \int_0^T \|\bar{p}_{,ttt}(\cdot, s)\|_H + \|A\bar{p}_{,tt}(\cdot, s)\|_H + \|\bar{u}_{,tt}(\cdot, s)\|_H \,\mathrm{d}\, s \\ &+ \|\bar{p}_{,tt}(\cdot, T)\|_H + \|A\bar{p}_{,t}(\cdot, T)\|_H + \|\bar{p}_{,tt}(\cdot, 0)\|_H + \|A\bar{p}_{,t}(\cdot, 0)\|_H \\ &+ \|\bar{p}_{,tt}\|_{L^2(0,T,W)} , \end{aligned}\right\} \tag{6.4}$$

Theorem 6.5. *Let Assumptions 6.4 hold. If the scheme* (OC CN1) *is applied to the optimal control conditions of Problem 5.12 the error can be estimated by*

$$\begin{aligned} \|\bar{y}_{h,i}(\cdot) - \bar{y}(\cdot, t_i)\|_H + \|\bar{y}_{h,0}(\cdot) - \bar{y}(\cdot, 0)\|_H &\lesssim C_1(\bar{y}, \bar{p})h^2 + C_2(\bar{y}, \bar{p})\tau^2, \\ \left\|\bar{p}_{h,i-\frac{1}{2}}(\cdot) - \bar{p}(\cdot, t_{i-\frac{1}{2}})\right\|_H + \|\bar{p}_{h,0}(\cdot) - \bar{p}(\cdot, 0)\|_H &\lesssim C_1(\bar{y}, \bar{p})h^2 + C_2(\bar{y}, \bar{p})\tau^2, \\ \|\bar{p}_{h,N}(\cdot) - \bar{p}(\cdot, T)\|_H &\lesssim C_1(\bar{y}, \bar{p})h^2 + C_2(\bar{y}, \bar{p})\tau^2, \\ i &= 1, \dots, N \end{aligned}$$

with C_1 and C_2 as in (6.3) *and* (6.4)*, i.e. we have a scheme of second order in h and τ.*

In preparation for the proof we will prove four lemmas. In these lemmas we discuss the convergence of the state with a given control, the error splitting for the adjoint state, the approximation of the adjoint state with a given state and the convergence of the control.

Remark 6.6. *For the proof we assume without loss of generality that the operator $M : H \to H$ is the identity. This assumption is without loss of generality as*

$$\langle\langle u, v \rangle\rangle_{H\times H} = \langle Mu, v \rangle_{H\times H}$$

defines an equivalent $L^2(\Omega)$-scalar product which introduces an equivalent $L^2(\Omega)$-norm.

Remark 6.7 (Regularity)**.** *In our analysis we assume*

$$\bar{y},\ \bar{p} \in H^3((0,T),L^2(\Omega)) \cap H^2((0,T),H^2(\Omega)).$$

For such a regularity in a problem with parabolic partial differential equations we need a smooth right hand side and further compatibility conditions on initial and boundary conditions, see Theorem 3.39 and Theorem 3.40.

In the example of a smooth domain Ω, e.g. if the domain is one dimensional, one obtains from Theorem 3.40

$$\bar{y} \in L^2((0,T),H^2(\Omega)) \cap L^\infty((0,T),H_0^1(\Omega)) \cap H^1((0,T),L^2(\Omega))$$

and hence

$$\bar{p} \in H^2((0,T),L^2(\Omega)) \cap H^1((0,T),H^2(\Omega)) \cap L^2((0,T),H^4(\Omega))$$

and with a bootstrapping argument

$$\bar{y},\ \bar{p} \in H^3((0,T),L^2(\Omega)) \cap H^2((0,T),H^2(\Omega)) \cap H^1((0,T),H^4(\Omega)) \cap L^2((0,T),H^6(\Omega))$$

under the assumptions

$$\begin{aligned} y_d &\in H^2((0,T),L^2(\Omega)) \cap H^1((0,T),H^2(\Omega)) \cap L^2((0,T),H^4(\Omega)) \\ y_D,\ v &\in H_0^1(\Omega),\ Av,\ Ay_D \in H_0^1(\Omega),\ AAv,\ AAy_D \in H_0^1(\Omega),\ AAAv, AAAy_D\ \in L^2(\Omega). \end{aligned}$$

Lemma 6.8. *Assume that we have a given control with*

$$\left\| \bar{u}(\cdot, t_{i+\frac{1}{2}}) - u_{h,i+\frac{1}{2}} \right\|_{L^2(\Omega)} \lesssim h^2 C_1 + \tau^2 C_2.$$

If we solve the numerical schemes (OC CN1)*,* (OC CN2) *or* (OC G1) *with this given control for the state, then the error estimate*

$$\|\bar{y}(\cdot,t_i) - y_{h,i}\|_{L^2(\Omega)} \lesssim h^2 C_1 + \tau^2 C_2 \qquad \text{for } i = 1,\dots,N$$

holds for the corresponding state.

Proof. We recall that the three discretization scheme differ in the discretization of the adjoint state, but coincide in the discretization of the state.

The error bound of this lemma is the result of Theorem 4.22. □

With a given approximation of the state, we can discuss the approximation of the adjoint state.

We use again error splitting techniques. For the splitting we use again the elliptic projection R_h which was defined in Definition 4.24. We split the errors in the adjoint state into the difference between the exact solution and its projection

$$\rho_0^{\bar{p}}(\cdot) = R_h\bar{p}(\cdot,0) - \bar{p}(\cdot,0), \qquad \rho_{i-\frac{1}{2}}^{\bar{p}}(\cdot) = R_h\bar{p}(\cdot,t_{i-\frac{1}{2}}) - \bar{p}(\cdot,t_{i-\frac{1}{2}}),$$

and the difference between the projection and the numerical approximation

$$\theta_0^p(\cdot) = p_{h,0}(\cdot) - R_h\bar{p}(\cdot,0), \qquad \theta_{i-\frac{1}{2}}^p(\cdot) = p_{h,i-\frac{1}{2}}(\cdot) - R_h\bar{p}(\cdot,t_{i-\frac{1}{2}}).$$

Lemma 6.9. *The error between the adjoint state and its projection can be estimated by*

$$\left\|\rho^{\bar{p}}_{i-\frac{1}{2}}\right\|_H = \left\|R_h\bar{p}(\cdot,t_{i-\frac{1}{2}}) - \bar{p}(\cdot,t_{i-\frac{1}{2}})\right\|_H \lesssim h^2\left\|\bar{p}(\cdot,T)\right\|_W + h^2\int_{t_{i-\frac{1}{2}}}^{T}\left\|\bar{p}_{,t}(\cdot,s)\right\|_W \mathrm{d}\, s,$$

$$\left\|\rho^{\bar{p}}_0\right\|_{L^2(\Omega)} = \left\|R_h\bar{p}(\cdot,0) - \bar{p}(\cdot,0)\right\|_H \lesssim h^2\left\|\bar{p}(\cdot,T)\right\|_W + h^2\int_{0}^{T}\left\|\bar{p}_{,t}(\cdot,s)\right\|_W \mathrm{d}\, s.$$

Proof. For the state equation this result is proven in Lemma 4.27. For the adjoint state we integrate backward in time from $t = T$ to $t = t_{i-\frac{1}{2}}$ or $t = 0$. □

Lemma 6.10. *For a given discretized state $y_{h,i}$ with $\|y_{h,i} - \bar{y}(\cdot,t_i)\|_{L^2(\Omega)} \leq C_1h^2 + C_2\tau^2$ with C_1 and C_2 specified in (6.3) and (6.4), the error of the numerical approximation of the adjoint state with the scheme* (OC CN1) *is bounded by*

$$\left\|p_{h,i-\frac{1}{2}}(\cdot) - \bar{p}(\cdot,t_{i-\frac{1}{2}})\right\|_{L^2(\Omega)} \lesssim C_1h^2 + C_2\tau^2,$$

$$\left\|p_{h,0}(\cdot) - \bar{p}(\cdot,0)\right\|_{L^2(\Omega)} \lesssim C_1h^2 + C_2\tau^2.$$

Proof. Due to Lemma 6.9 we only need to discuss the error between the projection of the adjoint state and the numerical approximation. The proof is done with an analogous arguments as the proof Lemma 6.8 and Theorem 4.22.

We look at the first half time step, the inner time steps and the last half time step separately. For a bound of the error θ^p_0 the use of the numerical scheme gives

$$\begin{aligned}
&\left\langle \frac{\theta^p_{\frac{1}{2}} - \theta^p_0}{\tau}, \varphi \right\rangle_{H\times H} - a\left(\frac{\theta_{\frac{1}{2}}}{2}, \varphi\right)\\
&= \left\langle \beta M_d\frac{y_0 - y_{d,0}}{2} - \beta M_d\frac{\bar{y}(\cdot,0) - y_d(\cdot,0)}{2}, \varphi\right\rangle_{H\times H} + \left\langle \frac{\bar{p}_t(\cdot,0)}{2} - \frac{\bar{p}(\cdot,t_{\frac{1}{2}}) - \bar{p}(\cdot,t_0)}{\tau}, \varphi\right\rangle_{H\times H}\\
&+ \left\langle I - R_h\frac{\bar{p}(\cdot,t_{\frac{1}{2}}) - \bar{p}(\cdot,0)}{\tau}, \varphi\right\rangle_{H\times H} + a\left(-\frac{\bar{p}(\cdot,0)}{2} + R_h\frac{\bar{p}(\cdot,0)}{2}, \varphi\right),
\end{aligned}$$

for the equality we have used the definition of the time stepping scheme and have added two zeros, once we used the continuous adjoint equation for $t = 0$ and the other time we added and subtracted the same expression. Due to the definition of the projection R_h the term $a\left(-\frac{\bar{p}(\cdot,0)}{2} + R_h\frac{\bar{p}(\cdot,0)}{2}, \varphi\right)$ vanishes, if we use test functions $\varphi \in V_h$. The use of $-\theta^p_0$ as test function yields

$$\begin{aligned}
\frac{1}{\tau}\left\|\theta^p_0\right\|^2_H &= \left\langle \theta^p_{\frac{1}{2}}, \theta^p_0\right\rangle_{H\times H} + \beta\left\langle M_d\frac{y_0 - y_{d,0}}{2} - M_d\frac{\bar{y}(\cdot,0) - y_d(\cdot,0)}{2}, -\theta^p_0\right\rangle_{H\times H}\\
&+ \left\langle \frac{\bar{p}_t(\cdot,0)}{2} - \frac{\bar{p}(\cdot,t_{\frac{1}{2}}) - \bar{p}(\cdot,t_0)}{\tau}, -\theta^p_0\right\rangle_{H\times H}\\
&+ \left\langle I - R_h\frac{\bar{p}(\cdot,t_{\frac{1}{2}}) - \bar{p}(\cdot,0)}{\tau}, -\theta^p_0\right\rangle_{H\times H} + a\left(\frac{\theta^p_{\frac{1}{2}}}{2}, -\theta^p_0\right).
\end{aligned}$$

The last term can be estimated by

$$a\left(\frac{\theta^p_{\frac{1}{2}}}{2}, -\theta^p_0\right) = \left\langle A\frac{\theta^p_{\frac{1}{2}}}{2}, -\theta^p_0\right\rangle_{H\times H} \lesssim \|\theta^p_0\|_H \left\|\theta^p_{\frac{1}{2}}\right\|_H.$$

With the Cauchy-Schwarz inequality and after cancellation of the common factor this yields the estimate

$$\|\theta^p_0\|_H \lesssim \left\|\theta^p_{\frac{1}{2}}\right\|_H + \frac{\tau}{2}\left\|\theta^p_{\frac{1}{2}}\right\|_H + \tau\|\omega_0\|_H + \tau C(h^2+\tau^2), \tag{6.5}$$

with

$$\omega_0 = (I-R_h)\frac{\bar{p}(\cdot,0)-\bar{p}(\cdot,\frac{\tau}{2})}{\tau} + \frac{\bar{p}(\cdot,\frac{\tau}{2})-\bar{p}(\cdot,0)}{\tau} - \frac{\bar{p}_{,t}(\cdot,0)}{2}.$$

For a bound of $\theta^p_{i+\frac{1}{2}}$ we discuss the inner time discretization nodes of the scheme. The scheme gives

$$\begin{aligned}
&\left\langle \frac{\theta^p_{i+\frac{1}{2}}-\theta^p_{i-\frac{1}{2}}}{\tau}, \varphi\right\rangle_{H\times H} - a\left(\frac{\theta^p_{i+\frac{1}{2}}+\theta^p_{i-\frac{1}{2}}}{2}, \varphi\right)\\
&= \beta\left\langle M_d\left(y_i - y_{d,i}\right) + M_d\left(y_d(\cdot,t_i) - \bar{y}(\cdot,t_i)\right), \varphi\right\rangle_{H\times H}\\
&+ \left\langle \bar{p}_t(\cdot,t_i) - \frac{\bar{p}(\cdot,t_{i+\frac{1}{2}})-\bar{p}(\cdot,t_{i-\frac{1}{2}})}{\tau}, \varphi\right\rangle_{H\times H} + \left\langle (I-R_h)\frac{\bar{p}(\cdot,t_{i+\frac{1}{2}})-\bar{p}(\cdot,t_{i-\frac{1}{2}})}{\tau}, \varphi\right\rangle_{H\times H}\\
&+ \left\langle A\frac{\bar{p}(\cdot,t_{i+\frac{1}{2}})+\bar{p}(\cdot,t_{i-\frac{1}{2}})}{2} - A\bar{p}(\cdot,t_i), \varphi\right\rangle_{H\times H},
\end{aligned}$$

for the equality we have used the definition of the time stepping scheme and have added two zeros, once we used the continuous adjoint equation for $t = t_i$ and the other time we added and subtracted the same expression. The use of $-\frac{\theta^p_{i+\frac{1}{2}}+\theta^p_{i-\frac{1}{2}}}{2}$ as test function yields

$$\left\|\theta^p_{i-\frac{1}{2}}\right\|_H \le \left\|\theta^p_{i+\frac{1}{2}}\right\|_H + \tau\left\|\omega_{i-\frac{1}{2}}\right\|_H + \tau C\left(h^2+\tau^2\right)$$

with

$$\begin{aligned}
\omega_{i-\frac{1}{2}} = (I-R_h)&\frac{\bar{p}(\cdot,t_{i-\frac{1}{2}})-\bar{p}(\cdot,t_{i+\frac{1}{2}})}{\tau} + \frac{\bar{p}(\cdot,t_{i+\frac{1}{2}})-\bar{p}(\cdot,t_{i-\frac{1}{2}})}{\tau} - \bar{p}_t(\cdot,t_i) +\\
&+ \frac{A\bar{p}(\cdot,t_{i-\frac{1}{2}})+A\bar{p}(\cdot,t_{i+\frac{1}{2}})}{2} - A\bar{p}(\cdot,t_i).
\end{aligned}$$

Finally we investigate the last step. For this we use the formulation with the additional

terminal adjoint state as in (OC CN1*). To derive a bound for $\theta^p_{N-\frac{1}{2}}$ we look at

$$\begin{aligned}
&\left\langle \frac{\theta^p_N - \theta^p_{N-\frac{1}{2}}}{\tau}, \varphi \right\rangle_{H\times H} - a\left(\frac{\theta^p_{N-\frac{1}{2}}}{2}, \varphi\right) \\
&= \left\langle \beta M_d \frac{y_N - y_{d,N}}{2} - \beta M_d \frac{\bar{y}(\cdot,t_N) - y_d(\cdot,t_N)}{2}, \varphi \right\rangle \\
&+ \left\langle \frac{\bar{p}(\cdot,t_N) - \bar{p}(\cdot,t_{N-\frac{1}{2}})}{\tau} - \frac{R_h\bar{p}(\cdot,t_N) - R_h\bar{p}(\cdot,t_{N-\frac{1}{2}})}{\tau}, \varphi \right\rangle + \frac{1}{2}a\left(R_h\bar{p}(\cdot,t_{N-\frac{1}{2}}), \varphi\right) \\
&+ \left\langle \frac{1}{2}\bar{p}_t(\cdot,t_N) - \frac{\bar{p}(\cdot,t_N) - \bar{p}(\cdot,t_{N-\frac{1}{2}})}{\tau}, \varphi \right\rangle - \frac{1}{2}a\left(\bar{p}(\cdot,t_N), \varphi\right),
\end{aligned}$$

where we have done the same steps as in the other cases. With $-\theta^p_{N-\frac{1}{2}}$ as test function we have

$$\left\|\theta^p_{N-\frac{1}{2}}\right\|^2_H + \left\langle \theta^p_N, -\theta^p_{N-\frac{1}{2}} \right\rangle_{H\times H} \leq \left\langle \frac{\theta^p_N - \theta^p_{N-\frac{1}{2}}}{\tau}, -\theta^p_{N-\frac{1}{2}} \right\rangle_{H\times H} - \frac{1}{2}a\left(\theta^p_{N-\frac{1}{2}}, -\theta^p_{N-\frac{1}{2}}\right)$$

due to the assumptions on the bilinear form $a(\cdot,\cdot)$. Therefore we get with the Cauchy-Schwarz inequality and cancellation of a common factor the estimate

$$\left\|\theta^p_{N-\frac{1}{2}}\right\|_H \leq \left\|\theta^p_N\right\|_H + \tau\left\|\omega_{N-\frac{1}{2}}\right\|_H + \tau C\left(h^2+\tau^2\right)$$

with

$$\begin{aligned}
\omega_{N-\frac{1}{2}} = (I - R_h)\frac{\bar{p}(\cdot,t_N) - \bar{p}(t_{N-\frac{1}{2}})}{\tau} + \frac{1}{2}\bar{p}_{,t}(\cdot,t_N) - \frac{\bar{p}(\cdot,t_N) - \bar{p}(\cdot,t_{N-\frac{1}{2}})}{\tau} \\
+ \frac{A\bar{p}(\cdot,t_{N-\frac{1}{2}}) - A\bar{p}(\cdot,t_N)}{2}.
\end{aligned}$$

Summing up, yields the estimates

$$\begin{aligned}
\|\theta^p_0\|_{L^2(\Omega)} &\lesssim \left\|\theta^p_{\frac{1}{2}}\right\|_{L^2(\Omega)} + \frac{\tau}{2}\left\|\theta^p_{\frac{1}{2}}\right\|_H + \tau\|\omega_0\|_H + \tau C(h^2+\tau^2) \\
&\lesssim \frac{\tau}{2}\left\|\theta^p_{\frac{1}{2}}\right\|_H + \left\|\theta^p_N\right\|_H + \tau\|\omega_0\|_H + \tau\sum_{j=1}^{N}\left\|\omega_{N-\frac{1}{2}}\right\|_H + C\left(h^2+\tau^2\right), \\
\left\|\theta^p_{i-\frac{1}{2}}\right\|_H &\leq \left\|\theta^p_{i+\frac{1}{2}}\right\|_H + \tau\left\|\omega_{i-\frac{1}{2}}\right\|_H + \tau C\left(h^2+\tau^2\right) \\
&\leq \left\|\theta^p_N\right\|_H + \tau\sum_{j=i}^{N}\left\|\omega_{N-\frac{1}{2}}\right\|_H + C\left(h^2+\tau^2\right), \\
\left\|\theta^p_{N-\frac{1}{2}}\right\|_H &\leq \left\|\theta^p_N\right\|_H + \tau\left\|\omega_{N-\frac{1}{2}}\right\|_H + \tau C\left(h^2+\tau^2\right)
\end{aligned}$$

with, as defined above,

$$\begin{aligned}
\omega_0 &= (I - R_h)\frac{\bar{p}(\cdot,0) - \bar{p}(\cdot,\frac{\tau}{2})}{\tau} + \frac{\bar{p}(\cdot,\frac{\tau}{2}) - \bar{p}(\cdot,0)}{\tau} - \frac{\bar{p}_{,t}(\cdot,0)}{2},\\
\omega_{i-\frac{1}{2}} &= (I - R_h)\frac{\bar{p}(\cdot,t_{i-\frac{1}{2}}) - \bar{p}(\cdot,t_{i+\frac{1}{2}})}{\tau} + \frac{\bar{p}(\cdot,t_{i+\frac{1}{2}}) - \bar{p}(\cdot,t_{i-\frac{1}{2}})}{\tau} - \bar{p}_t(\cdot,t_i) +\\
&\quad + \frac{A\bar{p}(\cdot,t_{i-\frac{1}{2}}) + A\bar{p}(\cdot,t_{i+\frac{1}{2}})}{2} - A\bar{p}(\cdot,t_i),\\
\omega_{N-\frac{1}{2}} &= (I - R_h)\frac{\bar{p}(\cdot,t_N) - \bar{p}(t_{N-\frac{1}{2}})}{\tau} + \frac{1}{2}\bar{p}_{,t}(\cdot,t_N) - \frac{\bar{p}(\cdot,t_N) - \bar{p}(\cdot,t_{N-\frac{1}{2}})}{\tau}\\
&\quad + \frac{A\bar{p}(\cdot,t_{N-\frac{1}{2}}) - A\bar{p}(\cdot,t_N)}{2}.
\end{aligned}$$

Due the assumptions on the approximation properties of y_N we have

$$\left\|\theta_N^p\right\| \leq C\left(\tau^2 + h^2\right),$$

thus we only need to bound the differences ω_i.

As in the proof of Theorem 4.22 we can bound $\tau\sum_{j=i}^{N-1}\left\|\omega_{i-\frac{1}{2}}\right\|_{L^2(\Omega)}$ as in [125, Theorem 1.6]. So we only need to bound the errors ω_0 and $\omega_{N-\frac{1}{2}}$ in the first and the last step.

The first term of ω_0 and $\omega_{N-\frac{1}{2}}$ is the error of a projection and therefore of order h^2, where we used also the cancellation of the factor $\frac{1}{\tau}$ with the factor τ. The other terms are of order τ^2 as

$$\begin{aligned}
-\frac{\tau}{2}\bar{p}_{,t}(\cdot,0) + \bar{p}\left(\cdot,\frac{\tau}{2}\right) - \bar{p}(\cdot,0) &= \frac{\tau^2}{8}\bar{p}_{,tt}(\cdot,0) + \frac{1}{2}\int_0^{\frac{\tau}{2}}\left(s - \frac{\tau}{2}\right)^2\bar{p}_{,ttt}(\cdot,s)\,\mathrm{d}\,s,\\
-\frac{\tau}{2}\bar{p}_{,t}(\cdot,t_N) + \bar{p}(\cdot,t_N) - \bar{p}(\cdot,t_{N-\frac{1}{2}}) &= -\frac{\tau^2}{8}\bar{p}_{,tt}(\cdot,t_N) + \frac{1}{2}\int_{t_{N-\frac{1}{2}}}^{T}\left(s - t_{N-\frac{1}{2}}\right)^2\bar{p}_{,ttt}(\cdot,s)\,\mathrm{d}\,s,\\
\tau\frac{A\bar{p}(\cdot,t_{N-\frac{1}{2}}) - A\bar{p}(\cdot,t_N)}{2} &= -\frac{\tau^2}{4}A\bar{p}_{,t}(\cdot,t_N) + \frac{\tau}{2}\int_{t_{N-\frac{1}{2}}}^{T}\left(s - t_{N-\frac{1}{2}}\right)A\bar{p}_{,tt}(\cdot,s)\,\mathrm{d}\,s
\end{aligned}$$

hold. So all terms of the error estimate are bounded with order $h^2 + \tau^2$. □

Remark 6.11. *Lemma 6.10 can also be used to investigate how accurate the data for a right hand side of the optimality system need to be evaluated so that the Crank-Nicolson scheme preserves second order convergence.*

Finally we need to assure that the control approximation is of second order. To this end we introduce some further notation. The interpolation operator J_τ is defined by

$$\begin{aligned}
&J_\tau u(\cdot,T) = u(\cdot,T),\ J_\tau u(\cdot,0) = u(\cdot,0),\\
&J_\tau u(\cdot,t_{k+\frac{1}{2}}) = u(\cdot,t_{k+\frac{1}{2}}), && \forall k = 0,\dots,N-1,\\
&\quad J_\tau u \text{ linear in } (0,t_{\frac{1}{2}}),\ (t_{k-\frac{1}{2}},t_{k+\frac{1}{2}}),\ (t_{N-\frac{1}{2}},t_N), && \forall k = 1,\dots,N-1,\\
&\quad J_\tau u \text{ continuous in } [0,T].
\end{aligned}$$

The optimal control is denoted by $\bar{u}$ and $\bar{u}_{h,\tau}$ at the discrete level. Finally we recall the standard interpolation error estimate

$$\|u - J_\tau u\|_{L^2(Q)} \lesssim \tau^2 \|u_{,tt}\|_{L^2(0,T,H^2(\Omega))} \qquad \forall u \in H^2((0,T), H^2(\Omega)). \tag{6.6}$$

Lemma 6.12. *If for the optimal control $\bar{u} \in H^3((0,T), L^2(\Omega)) \cap H^2((0,T), H^2(\Omega))$ holds, then the error of the control approximation can be bounded by*

$$\|J_\tau \bar{u}_{h,\tau} - \bar{u}\|_{L^2(Q)} \lesssim C_1 h^2 + C_2 \tau^2.$$

Proof. We follow the proof of [81, Theorem 6.1] and start with the weak form of the optimality condition

$$\int_0^T \langle \nu M_u \bar{u}(\cdot,t) - G^* \bar{p}(\cdot,t), \varphi \rangle_{H\times H} \,\mathrm{d}\, t = 0 \qquad \forall \varphi \in L^2(\Omega),$$

and its discretization

$$\begin{aligned}
\frac{\tau}{2} \langle \nu M_u \bar{u}_{h,0} - G^* \bar{p}_{h,0}, \varphi \rangle_{H\times H} &= 0 && \forall \varphi \in V_h. \\
\tau \left\langle \nu M_u \bar{u}_{h,k+\frac{1}{2}} - G^* \bar{p}_{h,k+\frac{1}{2}}, \varphi \right\rangle_{H\times H} &= 0 && \forall \varphi \in V_h, \text{ for } k = 0, \dots, N-1, \\
\frac{\tau}{2} \langle \nu M_u \bar{u}_{h,N} - G^* \bar{p}_{h,N}, \varphi \rangle_{H\times H} &= 0 && \forall \varphi \in V_h.
\end{aligned}$$

This is equivalent to the optimality condition of (OC CN1). Now we insert some admissible control u into the reduced continuous cost functional

$$j(u) = \frac{\alpha}{2} \|M_D S u(\cdot,T) - M_D y_D\|_H^2 + \int_0^T \frac{\beta}{2} \|M_d S u - M_d y_d\|_H^2 + \frac{\nu}{2} \|M_u u\|_H^2 \,\mathrm{d}\, t$$

with the (linear) solution operator S of the corresponding parabolic initial boundary value problem. As seen in Theorem 5.18 the first derivative of this functional can be written with the adjoint state in the form

$$j'(u)(\varphi) = \int_0^T \langle \nu M_u u(\cdot,t) - G^* p(\cdot,t;u), \varphi \rangle_{H\times H} \,\mathrm{d}\, t \tag{6.7}$$

with $p(\cdot,\cdot;u)$ being the adjoint state corresponding to the control u. The second derivative of the cost functional is

$$j''(u)(\varphi,\varphi) = \int_0^T \nu \langle M_u \varphi, \varphi \rangle_{H\times H} + \langle G^* S\varphi, S\varphi \rangle_{H\times H} \,\mathrm{d}\, t \geq \int_0^T \nu \|\varphi\|_{L^2(\Omega)}^2 \,\mathrm{d}\, t \tag{6.8}$$

and therefore independent of u. The discrete optimality condition

$$\begin{aligned}
\frac{\tau}{2} \langle \nu M_u \bar{u}_{h,0} - G^* \bar{p}_{h,0}, \varphi \rangle_{H\times H} &= 0 && \forall \varphi \in V_h. \\
\tau \left\langle \nu M_u \bar{u}_{h,k+\frac{1}{2}} - \bar{G}^* p_{h,k+\frac{1}{2}}, \varphi \right\rangle_{H\times H} &= 0 && \forall \varphi \in V_h, \text{ for } k = 0, \dots, N-1, \\
\frac{\tau}{2} \langle \nu M_u \bar{u}_{h,N} - G^* \bar{p}_{h,N}, \varphi \rangle_{H\times H} &= 0 && \forall \varphi \in V_h.
\end{aligned}$$

implies, as the space V_h is finite dimensional,

$$\begin{aligned} \nu M_u \bar{u}_{h,0} - G^* \bar{p}_{h,0} &= 0 \\ \nu M_u \bar{u}_{h,k+\frac{1}{2}} - G^* \bar{p}_{h,k+\frac{1}{2}} &= 0 \qquad \text{for } k = 0, \dots, N-1. \\ \nu M_u \bar{u}_{h,N} - G^* \bar{p}_{h,N} &= 0. \end{aligned}$$

And therefore we have the optimality condition also for the interpolant in time

$$\nu J_\tau M_u \bar{u}_{h,\tau} - J_\tau G^* \bar{p}_{h,\tau} = 0.$$

We define the functional $j'_{h\tau}(J_\tau M_u \bar{u}_{h,\tau})$ as the weak form of this optimality condition

$$j'_{h\tau}(J_\tau M_u \bar{u}_{h,\tau})(J_\tau \varphi_{h,\tau}) = \int_0^T \langle \nu J_\tau M_u \bar{u}_{h,\tau} - J_\tau G^* \bar{p}_{h,\tau}(\cdot,\cdot,\bar{u}_{h,\tau}), J_\tau \varphi_{h,\tau} \rangle_{H \times H} \,\mathrm{d}\, t.$$

For the optimal solution we have

$$j'_{h\tau}(J_\tau M_u \bar{u}_{h,\tau})(J_\tau \varphi_{h,\tau}) = 0 \qquad \forall J_\tau \varphi_{h,\tau} \in \mathcal{Y}_{1,h}. \tag{6.9}$$

with the space

$$\mathcal{Y}_{1,h} = \left\{ y \in \mathcal{C}\left([0,T], V_h\right), \;\; y|_{(t_i,t_{i+1})} \in \mathbb{P}_1\left((t_i,t_{i+1}), V_h\right) \;\; \forall i \in \{0,1,\dots,N\} \right\}.$$

For the error between the projection of the exact solution and the numerical approximation

$$e_{h,\tau} = J_\tau R_h \bar{u} - J_\tau \bar{u}_{h,\tau}$$

we have with any $\varphi \in H^3((0,T), L^2(\Omega)) \cap H^3((0,T), H^2(\Omega))$ and with (6.8)

$$\nu \|e_{h,\tau}\|^2_{L^2(Q)} \leq j''(\varphi)(e_{h,\tau}, e_{h,\tau}) = j'\left(J_\tau R_h \bar{u}\right)(e_{h,\tau}) - j'\left(J_\tau \bar{u}_{h,\tau}\right)(e_{h,\tau}).$$

Due to the optimality conditions (6.9) and

$$j'(\bar{u})(\varphi(\cdot)) = 0 \qquad \forall \varphi \in L^2(Q) \supset \mathcal{Y}_{1,h},$$

this is equal to

$$\begin{aligned} \nu \left\|e_{h,\tau}\right\|^2_H \leq\; & j'(J_\tau R_h \bar{u})(e_{h,\tau}) - j'(\bar{u})(e_{h,\tau}) \\ & + j'_{h\tau}\left(J_\tau \bar{u}_{h,\tau}\right)(e_{h,\tau}) - j'\left(J_\tau \bar{u}_{h,\tau}\right)(e_{h,\tau}). \end{aligned}$$

After the application of the Cauchy-Schwarz inequality for $L^2(Q)$ in (6.7) this is

$$\begin{aligned} \nu \left\|e_{h,\tau}\right\|^2_{L^2(Q)} \lesssim\; & \nu \left\|M_u J_\tau R_h \bar{u} - M_u \bar{u}\right\|_{L^2(Q)} \left\|e_{h,\tau}\right\|_{L^2(Q)} \\ & + \left\|G^* \bar{p}(\cdot,\cdot;\bar{u}) - G^* p(\cdot,\cdot;J_\tau R_h \bar{u})\right\|_{L^2(Q)} \left\|e_{h,\tau}\right\|_{L^2(Q)} \\ & + \left\|G^* p(\cdot,\cdot,J_\tau \bar{u}_{h,\tau}) - G^* J_\tau \bar{p}_{h,\tau}(\cdot,\cdot;J_\tau \bar{u}_{h,\tau})\right\|_{L^2(Q)} \left\|e_{h,\tau}\right\|_{L^2(Q)} \\ & + \nu \left\|M_u J_\tau \bar{u}_{h,\tau} - M_u J_\tau \bar{u}_{h,\tau}\right\|_{L^2(Q)} \left\|e_{h,\tau}\right\|_{L^2(Q)}, \end{aligned}$$

where $J_\tau \bar{p}_{h,\tau}(\cdot,\cdot; J_\tau \bar{u}_{h,\tau})$ denotes the discrete adjoint state corresponding to the control $J_\tau \bar{u}_{h,\tau}$. Thus we proved for the error between the projection of the optimal control and its numerical approximation the estimate

$$\begin{aligned}\nu \left\| e_{h,\tau} \right\|_{L^2(Q)} \lesssim \nu \left\| M_u \left(J_\tau R_h \bar{u} - \bar{u} \right) \right\|_{L^2(Q)} \\ + \left\| G^* \left(\bar{p}(\cdot,\cdot;\bar{u}) - p(\cdot,\cdot; J_\tau R_h \bar{u}) \right) \right\|_{L^2(Q)} \\ + \left\| G^* \left(p(\cdot,\cdot; J_\tau \bar{u}_{h,\tau}) - J_\tau \bar{p}_{h,\tau}(\cdot,\cdot; J_\tau \bar{u}_{h,\tau}) \right) \right\|_{L^2(Q)}.\end{aligned}$$

With this result, the error between the optimal control and its numerical approximation can be estimated with the the triangle inequality. This yields

$$\begin{aligned}\left\| \bar{u} - J_\tau \bar{u}_{h,\tau} \right\|_{L^2(Q)} &\leq \left\| \bar{u} - R_h \bar{u} \right\|_{L^2(Q)} + \left\| R_h \bar{u} - J_\tau R_h \bar{u} \right\|_{L^2(Q)} + \left\| J_\tau R_h \bar{u} - J_\tau \bar{u}_{h,\tau} \right\|_{L^2(Q)} \\ &\lesssim \left\| I + M_u \right\| \left\| \bar{u} - R_h \bar{u} \right\|_{L^2(Q)} + \left\| I + M_u \right\| \left\| R_h \bar{u} - J_\tau R_h \bar{u} \right\|_{L^2(Q)} \\ &+ \frac{1}{\nu} \left\| G^* \right\| \left\| \bar{p}(\cdot,\cdot;\bar{u}) - p(\cdot,\cdot; J_\tau R_h \bar{u}) \right\|_{L^2(Q)} \\ &+ \frac{1}{\nu} \left\| G^* \right\| \left\| p(\cdot,\cdot; J_\tau \bar{u}_{h,\tau}) - J_\tau \bar{p}_{h,\tau}(\cdot,\cdot; J_\tau \bar{u}_{h,\tau}) \right\|_{L^2(Q)}.\end{aligned}$$

The first term, the error between the optimal control and its projection, can be bounded as in Lemma 6.9. The second term is bounded by an interpolation result as in (6.6). The last term can be bounded due to the Theorem 4.22 and Lemma 6.10. For the third term we discuss $\tilde{p} = \bar{p}(\cdot,\cdot;\bar{u}) - p(\cdot,\cdot; J_\tau R_h \bar{u})$. By subtraction the differential equations for $\bar{p}(\cdot,\cdot;\bar{u})$ and $p(\cdot,\cdot; J_\tau R_h \bar{u})$ we see, that this functions fulfills the differential equation

$$\begin{aligned}\tilde{p}_{,t} - A^* \tilde{p} &= \beta M_d \tilde{y} \\ \tilde{p}(\cdot, T) &= \alpha M_D \tilde{y}(\cdot, T)\end{aligned}$$

with homogeneous boundary conditions. The right hand side $\tilde{y}$ is the solution of

$$\begin{aligned}\tilde{y}_{,t} + A\tilde{y} &= G\bar{u} - G J_\tau R_h \bar{u} \\ \tilde{y}(\cdot, 0) &= 0\end{aligned}$$

with the same homogeneous boundary conditions on the spatial boundary. With the a priori estimate of Theorem 3.40 we obtain

$$\left\| \tilde{y} \right\|_{L^2(Q)} \leq \left\| G \right\| \left\| \bar{u} - J_\tau R_h \bar{u} \right\|_{L^2(Q)} \leq \left\| G \right\| \left\| \bar{u} - R_h \bar{u} \right\|_{L^2(Q)} + \left\| R_h \bar{u} - J_\tau R_h \bar{u} \right\|_{L^2(Q)}.$$

This estimate can also be applied to the adjoint equation, as the adjoint equation can be transformed to a parabolic initial boundary value problem with time $\tilde{t} = T - t$. Therefore we have the estimate

$$\left\| \tilde{p} \right\|_{L^2(Q)} \leq \left\| \tilde{y} \right\|_{L^2(Q)} \leq \left\| \bar{u} - J_\tau R_h \bar{u} \right\|_{L^2(Q)}.$$

So we have bound all terms in the estimate for $\left\| \bar{u}(\cdot,\cdot) - J_\tau \bar{u}_{h,\tau}(\cdot,\cdot) \right\|_{L^2(Q)}$. □

All together we have proven Theorem 6.5.

Proof of Theorem 6.5. The convergence of the control follows with Lemma 6.12. The convergence of the control implies the convergence of the state (Lemma 6.8), which implies the convergence of the adjoint state (Lemma 6.10). □

Finally we transfer the result to the other schemes.

Remark 6.13. *We have not shown that the schemes (OC G2) and (OC CN2) are second order schemes. The schemes only differ from (OC CN1) in the right hand side of the adjoint state and for the non-final time steps we have (by Taylor expansion)*

$$\begin{aligned}\frac{\frac{y_{i-1}+y_i}{2} - \frac{y_{d,i-1}+y_{d,i}}{2}}{2} + \frac{\frac{y_i+y_{i+1}}{2} - \frac{y_{d,i}+y_{d,i+1}}{2}}{2} &= y_i - y_{d,i} + \mathcal{O}(\tau^2),\\ \frac{y_{i-1} - y_{d,i-1}}{6} + \frac{4}{6}(y_i - y_{d,i}) + \frac{y_{i+1} - y_{d,i+1}}{6} &= y_i - y_{d,i} + \mathcal{O}(\tau^2).\end{aligned}$$

Therefore the assumptions of Lemma 6.10 hold for these time steps. But for the final step we can only show

$$\begin{aligned}\frac{\frac{y_{N-1}+y_N}{2} - \frac{y_{d,N}+y_{d,N-1}}{2}}{2} &= \frac{y_N - y_{d,N}}{2} + \mathcal{O}(\tau),\\ \frac{1}{6}(y_{N-1} - y_{d,N-1}) + \frac{2}{6}(y_N - y_{d,N}) &= \frac{y_N - y_{d,N}}{2} + \mathcal{O}(\tau).\end{aligned}$$

by Taylor expansions. Nevertheless we see in the numerical example in Section 6.6 that all the schemes seem to be of second order.

For Crank-Nicolson discretizations of parabolic partial differential equation with irregular initial data it is known that even two first order implicit Euler starting steps do not destroy the convergence [100, Theorem 2]. We hope that we can transfer similar results to the schemes (OC G2) and (CN2). This is work of further research.

6.5. Variable time steps

As mentioned before, it is well known that for higher regularity of the solution of parabolic partial differential equations additional compatibility conditions for the initial and boundary data are needed. If the compatibility conditions are not fulfilled or the initial data are non-smooth, graded time step sizes are in use, see e. g. [117, Section 5.2] for the h-version (or better τ-version in this context) of a discontinuous Galerkin scheme in time or [95] for different approaches for the Crank-Nicolson scheme and non-smooth initial data.

These incompatibilities can appear in the state for $t = 0$ and in the adjoint state for $t = T$. Therefore the error analysis for appropriate time step generating function is done in Section 6.5.2.

6.5.1. Generalization to variable time step sizes

We generalize our scheme (OC CN1) to variable time step sizes. Therefore let $\tau_i = t_i - t_{i-1}$ for $i = 1, \dots, N$. The Crank-Nicolson discretization with variable time step size is

$$M\frac{y_{k+1} - y_k}{\tau_{k+1}} + A\frac{y_{k+1} + y_k}{2} = Gu_{k+\frac{1}{2}}$$

and for the cost functional we chose the trapezoidal rule for the first integral and the midpoint rule for the second integral, which yields

$$\frac{\beta\tau_1}{4}\left\|M_d^{1/2}\left(y_0-y_{d,0}\right)\right\|_H^2+\sum_{k=1}^{N-1}\beta\frac{\tau_k\left\|M_d^{1/2}(y_k-y_{d,k})\right\|_H^2+\left\|M_d^{1/2}\tau_{k+1}(y_k-y_{d,k})\right\|_H^2}{4}$$
$$+\frac{\beta\tau_N}{4}\left\|M_d^{1/2}(y_N-y_{D,N})\right\|_H^2+\sum_{k=0}^{N-1}\frac{\nu\tau_{k+1}}{2}\left\|M_u^{1/2}u_{k+\frac12}\right\|_H^2.$$

The corresponding Lagrangian is

$$\begin{aligned}\mathcal{L}(\mathbf{y},\mathbf{u},\mathbf{p}) = {}& \frac{\beta\tau_1}{4}\left\|M_d^{1/2}\left(y_0-y_{d,0}\right)\right\|_H^2\\ &+\sum_{k=1}^{N-1}\beta\frac{\tau_k\left\|M_d^{1/2}(y_k-y_{d,k})\right\|_H^2+\left\|M_d^{1/2}\tau_{k+1}(y_k-y_{d,k})\right\|_H^2}{4}\\ &+\frac{\beta\tau_N}{4}\left\|M_d^{1/2}(y_N-y_{d,N})\right\|_H^2+\frac{\alpha}{2}\left\|M_D^{1/2}(y_N-y_N)\right\|_H^2\\ &+\sum_{k=0}^{N-1}\frac{\nu\tau_{k+1}}{2}\left\|M_u^{1/2}u_{k+\frac12}\right\|_H^2+\left\langle M\left(y_0-v\right),p_0\right\rangle_{H\times H}\\ &+\sum_{k=0}^{N-1}\tau_{k+1}\left\langle M\frac{y_{k+1}-y_k}{\tau_{k+1}}+A\frac{y_{k+1}+y_k}{2}-Gu_{k+\frac12},p_{k+\frac12}\right\rangle_{H\times H}.\end{aligned}$$

The first order conditions $\frac{\partial\mathcal{L}}{\partial p_0}$, $\frac{\partial\mathcal{L}}{\partial p_{i+\frac12}}$ (for $i=0,\dots N-1$), $\frac{\partial\mathcal{L}}{\partial u_{i+\frac12}}$ (for $i=0,\dots N-1$) and $\frac{\partial\mathcal{L}}{\partial y_i}$ (for $i=0,\dots N$) are

$$\begin{aligned}My_0 &= Mv\\ M\frac{\bar y_{i+1}-\bar y_i}{\tau_{i+1}}+A\frac{\bar y_{i+1}+\bar y_i}{2} &= G\bar u_{i+\frac12} && \text{for } i=0,\dots N-1,\\ \nu M_u\bar u_{i+\frac12} &= G^*\bar p_{i+\frac12} && \text{for } i=0,\dots N-1,\\ M\frac{\bar p_{\frac12}-\bar p_0}{\tau_1}+A\frac{\bar p_{\frac12}}{2} &= \frac{\beta}{2}M_d(\bar y_0-y_{d;0})\\ M\bar p_{i+\frac12}-M\bar p_{i-\frac12}-A\frac{\tau_i\bar p_{i-\frac12}+\tau_{i+1}\bar p_{i+\frac12}}{2} &= \beta\frac{\tau_i+\tau_{i+1}}{2}M_d(\bar y_i-y_{d;i}) && \text{for } i=1,\dots,N-2,\\ -M\frac{\bar p_{N-\frac12}}{\tau_N}-A\frac{\bar p_{N-\frac12}}{2} &= \frac{\beta}{2}M_d(\bar y_N-y_{d;N})\\ &\quad+\frac{\alpha}{\tau_N}M_D\left(y_N-y_D\right).\end{aligned}$$

The only difference to the optimality system (OC CN1), in addition to the replacement of τ by τ_i, is the discretization of the adjoint state for the indices $i=1,\dots,N-2$.

6.5.2. Convergence analysis

In this section we show that second order convergence is also possible in the case of variable time step sizes. The Lemmas 6.8 and 6.10 of the previous section, which also discuss perturbed

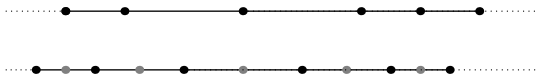


Figure 6.3.: Comparison of the discretization time nodes of y and p. Midpoints of the intervals $[p_{i-\frac{1}{2}}, p_{i+\frac{1}{2}}]$ are additionally inserted in gray. First line y, second line p

approximations, are the key to our analysis (see also Remark 6.11).

The forward equation is the same as in the case of constant time step sizes. Therefore we achieve second order convergence if $\bar{p}_{i+\frac{1}{2}}$ is a second order approximation.

So we only need to discuss the approximation of the backward equation. If we show second order convergence for $\bar{p}_{i+\frac{1}{2}}$ we are done. We only have problems if the time-step sizes change. In these cases the midpoint of the interval $[t_{i-\frac{1}{2}}, t_{i+\frac{1}{2}}]$ and t_i do not coincide anymore (see Figure 6.3). The time step size between $\bar{p}_{i-\frac{1}{2}}$ and $\bar{p}_{i+\frac{1}{2}}$ is equal to $\frac{\tau_{i+1}+\tau_i}{2} = \frac{t_{i+1}-t_{i-1}}{2}$. For simplicity and shortness we assume $y_{d,i} = 0$. All the arguments for y_i carry over to $y_{d,i}$ if y_d is smooth enough. So we discuss the scheme

$$-M\bar{p}_{i-\frac{1}{2}} + M\bar{p}_{i+\frac{1}{2}} + A\frac{\tau_i\bar{p}_{i-\frac{1}{2}} + \tau_{i+1}\bar{p}_{i+\frac{1}{2}}}{2} = \beta\frac{\tau_i + \tau_{i+1}}{2}M_d\bar{y}_i$$

for $\bar{p}_{i-\frac{1}{2}}$ and show that this scheme is a $\mathcal{O}\left(\overline{\tau}_i^2\right)$ perturbation of the midpoint-rule

$$M\frac{-\bar{p}_{i-\frac{1}{2}} + \bar{p}_{i+\frac{1}{2}}}{\frac{t_{i+1}-t_{i-1}}{2}} + A\bar{p}\left(\frac{t_{i-\frac{1}{2}} + t_{i+\frac{1}{2}}}{2}\right) = \beta M_d\bar{y}\left(\frac{\frac{t_{i-1}+t_i}{2} + \frac{t_i+t_{i+1}}{2}}{2}\right).$$

If we have proven this we can use Lemma 6.8 and are done. Therefore we divide our scheme by the time step size $\frac{t_{i+1}-t_{i-1}}{2}$,

$$M\frac{-\bar{p}_{i-\frac{1}{2}} + \bar{p}_{i+\frac{1}{2}}}{\frac{t_{i+1}-t_{i-1}}{2}} + \frac{1}{\frac{t_{i+1}-t_{i-1}}{2}}A\frac{\tau_i\bar{p}_{i-\frac{1}{2}} + \tau_{i+1}\bar{p}_{i+\frac{1}{2}}}{2} = \beta M_d\bar{y}_i.$$

Lemma 6.14. *If the changes of time step size are of order*

$$\tau_{i+1} - \tau_i = \mathcal{O}\left(\overline{\tau}^2\right), \tag{6.10}$$

then

$$y\left(\frac{\frac{t_{i-1}+t_i}{2} + \frac{t_i+t_{i+1}}{2}}{2}\right) - y(t_i) = \mathcal{O}\left(\overline{\tau}^2\right),$$

$$Ap\left(\frac{t_{i-\frac{1}{2}} + t_{i+\frac{1}{2}}}{2}\right) - \frac{1}{\frac{t_{i+1}-t_{i-1}}{2}}A\frac{\tau_i p(t_{i-\frac{1}{2}}) + \tau_{i+1}p(t_{i+\frac{1}{2}})}{2} = \mathcal{O}\left(\overline{\tau}^2\right).$$

Proof. For the proof we use Taylor expansions. With the assumption (6.10) we have

$$y\left(\frac{\frac{t_{i-1}+t_i}{2} + \frac{t_i+t_{i+1}}{2}}{2}\right) - y(t_i) = \dot{y}(t_i)\frac{t_{i-1} - 2t_i + t_{i+1}}{4} + h.o.t. = \mathcal{O}\left(\overline{\tau}_i^2\right).$$

For the other term we compare Taylor expansions

$$\frac{1}{\frac{t_{i+1}-t_{i-1}}{2}} A \frac{\tau_i p(t_{i-\frac{1}{2}}) + \tau_{i+1} p(t_{i+\frac{1}{2}})}{2} =$$

$$= \frac{1}{\tau_i + \tau_{i+1}} A \left(\tau_i \left(p(t_i) - \frac{\tau_i}{2} \dot{p}(t_i) + h.o.t. \right) + \tau_{i+1} \left(p(t_i) + \frac{\tau_i}{2} \dot{p}(t_i) + h.o.t. \right) \right) =$$

$$= Ap(t_i) + \frac{\tau_{i+1}^2 - \tau_i^2}{\tau_i + \tau_{i+1}} \frac{1}{2} A\dot{p}(t_i) + h.o.t. = Ap(t_i) + \frac{\tau_{i+1} - \tau_i}{2} A\dot{p}(t_i) + h.o.t.$$

and

$$Ap\left(\frac{t_{i-\frac{1}{2}} + t_{i+\frac{1}{2}}}{2} \right) = Ap(t_i) + A\dot{p}(t_i) \left(\frac{t_i + t_{i-1} + t_i + t_{i+1}}{4} - t_i \right) + h.o.t.$$

$$= Ap(t_i) + A\dot{p}(t_i) \left(\frac{\tau_{i+1} - \tau_i}{4} \right) + h.o.t.$$

As above the difference of the two expansions is of order $\mathcal{O}\left(\overline{\tau}_i^2\right)$. □

Altogether we have proven the convergence for variable time step sizes.

Theorem 6.15. *The scheme with variable time step sizes is a second order scheme if the time grid satisfies assumption (6.10).*

Corollary 6.16. *Last we mention a method to provide a variable time step distribution which fulfills equation (6.10). Therefore we choose a monotone mesh generating function k which fulfills*

$$k \in \mathcal{C}^2([0,1],[0,T]) \qquad k(0) = 0 \qquad k(1) = T \qquad t_i = k\left(\frac{i}{N}\right).$$

The resulting time step sizes τ_i fulfill the condition (6.10) of theorem 6.15.

Proof. We use Taylor expansions of both sides of (6.10). For the left hand side we have

$$\begin{aligned}
\tau_{i+1} - \tau_i = t_{i+1} - 2t_i + t_{i-1} &= k\left(\frac{i+1}{N}\right) - 2k\left(\frac{i}{N}\right) + k\left(\frac{i-1}{N}\right) \\
&= k\left(\frac{i}{N}\right) + k'\left(\frac{i}{N}\right)\frac{1}{N} + \frac{1}{2}k''(\xi_1)\frac{1}{N^2} - 2k\left(\frac{i}{N}\right) \\
&\quad + k\left(\frac{i}{N}\right) - k'\left(\frac{i}{N}\right)\frac{1}{N} + \frac{1}{2}k''(\xi_2)\frac{1}{N^2} \\
&= \frac{1}{2}\left(k''(\xi_1) + k''(\xi_2)\right)\frac{1}{N^2} = \mathcal{O}\left(\frac{1}{N^2}\right)
\end{aligned}$$

with some $\xi_1 \in [\frac{i}{N}, \frac{i+1}{N}]$ and $\xi_2 \in [\frac{i-1}{N}, \frac{i}{N}]$. And for the right hand side we compute

$$\begin{aligned}
\tau_i = t_i - t_{i-1} &= k\left(\frac{i}{N}\right) - \left(k\left(\frac{i}{N}\right) - k'\left(\frac{i}{N}\right)\frac{1}{N} + \frac{1}{2}k''(\xi_3)\frac{1}{N^2}\right) \\
&= k'\left(\frac{i}{N}\right)\frac{1}{N} - \frac{1}{2}k''(\xi_3)\frac{1}{N^2} = \mathcal{O}\left(\frac{1}{N}\right).
\end{aligned}$$

This finishes this proof as $\tau_{i+1} - \tau_i$ is of higher order then needed for (6.10). □

Remark 6.17. *Rösch discusses in [103] a parabolic optimal control problem with a terminal objective functional and control constraints. He uses*

$$k(t) = T - T\,(1-t)^4$$

as grading function and shows that with this grading towards $t = T$ the convergence of the control is of order $\frac{3}{2}$. With our scheme he would have obtained order 2 for the case without constraints.

Remark 6.18. *For the simulation of parabolic partial differential equations with discontinuous Galerkin schemes as time discretization, Schötzau and Schwab introduce*

$$k(t) = T \cdot t^{(2r+3)/\theta}$$

as mesh generating function in [117, Section 5.2]. The constant r is the polynomial degree of the discontinuous Galerkin scheme in time and the constant $\theta \in (0, 1]$ corresponds to the smoothness of the initial data, so that $y_0 \in H^\theta(\Omega)$. Clearly this function fulfills also our conditions.

6.6. Numerical examples

6.6.1. Solution Algorithm

As we discuss a problem without control or state constraints it is possible to eliminate the optimality condition in the discrete system. Altogether for (OC CN1) we have to solve the linear system

$$\left(\begin{array}{c|c} I & II \\ \hline III & IV \end{array}\right) \left(\begin{array}{c} \bar{Y}_\tau \\ \hline \bar{P}_\tau \end{array}\right) = \left(\begin{array}{c} R_1 \\ \hline R_2 \end{array}\right) \tag{6.11}$$

with the sub-matrices and the sub-vectors given by

$$I = \begin{pmatrix} K & & & \\ L & K & & \\ & \ddots & \ddots & \\ & & L & K \end{pmatrix}, \qquad II = \begin{pmatrix} -C & & & \\ & \ddots & & \\ & & \ddots & \\ & & & -C \end{pmatrix},$$

$$III = \begin{pmatrix} -\beta M_d & & & \\ & \ddots & & \\ & & -\beta M_d & \\ & & & -\frac{\alpha}{\tau} M_D - \frac{\beta}{2} M_d \end{pmatrix}, \quad IV = \begin{pmatrix} -K & -L & & \\ & \ddots & \ddots & \\ & & -K & -L \\ & & & -K \end{pmatrix},$$

$$\bar{Y}_\tau = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \vdots \\ \bar{y}_N \end{pmatrix}, \qquad \bar{P}_\tau = \begin{pmatrix} \bar{p}_{\frac{1}{2}} \\ \vdots \\ \vdots \\ \bar{p}_{N-\frac{1}{2}} \end{pmatrix},$$

$$R_1 = \begin{pmatrix} -Lv \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad R_2 = \begin{pmatrix} -\beta M_d y_{d,1} \\ \vdots \\ -\beta M_d y_{d,N-1} \\ -\frac{\alpha}{\tau} M_D y_D - \frac{\beta}{2} M_d y_{d,N} \end{pmatrix}$$

with $K = \frac{M}{\tau} + \frac{A}{2} \in \mathbb{R}^{n \times n}$, $C = \frac{1}{\nu} G M_u^{-1} G^* \in \mathbb{R}^{n \times n}$ and $L = -\frac{M}{\tau} + \frac{A}{2} \in \mathbb{R}^{n \times n}$, where the operators have been replaced by their discrete counterpart, i.e. M is the mass matrix and A the stiffness matrix and so on. As the adjoint state $\bar{p}_0$ does not influence the further computations, it can be computed afterwards.

If we choose (OC CN2) the lower left sub-matrix and the lower part of the right hand side have to be replaced by

$$III = \begin{pmatrix} -\frac{\beta}{2} M_d & -\frac{\beta}{4} M_d & & & \\ -\frac{\beta}{4} M_d & -\frac{\beta}{2} M_d & -\frac{\beta}{4} M_d & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{\beta}{4} M_d & -\frac{\beta}{2} M_d & -\frac{\beta}{4} M_d \\ & & & -\frac{\beta}{4} M_d & -\frac{\alpha}{\tau} M_D - \frac{\beta}{4} M_d \end{pmatrix},$$

$$R_2 = \begin{pmatrix} -\beta M_d \frac{y_{d,0} + 2y_{d,1} + y_{d,2}}{4} \\ \vdots \\ -\beta M_d \frac{y_{d,N-2} + 2y_{d,N-1} + y_{d,N}}{4} \\ -\frac{\alpha}{\tau} M_D y_D - \beta M_d \frac{y_{d,N-1} + y_{d,N}}{4} \end{pmatrix},$$

and for the discretization (OC G1) the lower matrices are replaced by

$$III = \begin{pmatrix} -\frac{4}{6}\beta M_d & -\frac{\beta}{6}M_d & & & \\ -\frac{\beta}{6}M_d & -\frac{4}{6}\beta M_d & -\frac{\beta}{6}M_d & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{\beta}{6}M_d & -\frac{4}{6}\beta M_d & -\frac{\beta}{6}M_d \\ & & & -\frac{\beta}{6}M_d & -\frac{\alpha}{\tau}M_D y_D - \frac{2}{6}\beta M_d \end{pmatrix},$$

$$R_2 = \begin{pmatrix} -\beta M_d \frac{y_{d,0}+4y_{d,1}+y_{d,2}}{6} \\ \vdots \\ -\beta M_d \frac{y_{d,N-2}+4y_{d,N-1}+y_{d,N}}{6} \\ -\frac{\alpha}{\tau}M_D y_D - \beta M_d \frac{y_{d,N-1}+2y_{d,N}}{6} \end{pmatrix}.$$

6.6.2. Tracking over the full space time cylinder

As first numerical example we choose the tracking over the full space-time cylinder and no tracking of a terminal state, i.e. $\alpha = 0$ and $\beta = 1$ in the cost functional of Problem 5.2.

Example 6.19. *The first numerical example is taken from [4, 5]. We choose a test problem with parabolic partial differential equations with homogeneous Neumann boundary conditions*

$$\left.\begin{aligned} \min \int_0^1 \frac{1}{2}\|y - y_d\|_H^2 + \frac{\nu}{2}\|u\|_H^2 \,\mathrm{d}\,t & \\ y_{,t} - \Delta y = u & \quad \text{in } \Omega \times (0,T], \\ \frac{\partial}{\partial n} y = 0 & \quad \text{on } \partial\Omega \times (0,T], \\ y = 0 & \quad \text{in } \Omega \times \{0\}. \end{aligned}\right\} \tag{6.12}$$

For our numerical example we study $\Omega \times (0,T] = (0,1)^2 \times (0,1]$.

We measure the error by

$$\max\left\{\left((\bar{y}_{hi} - \bar{y}(t_i,x))^T M (\bar{y}_{hi} - \bar{y}(t_i,x))\right)^{1/2}, i = 0,\dots,N\right\},$$

$$\text{and } \max\left\{\left((\bar{p}_{hi} - \bar{p}(t_i,x))^T M (\bar{p}_{hi} - \bar{p}(t_i,x))\right)^{1/2}, i = 0,\frac{1}{2},\frac{3}{2},\dots,N-\frac{1}{2}\right\}.$$

For these expressions we have proven error bounds of order τ^2. *It can be interpreted as a discretization of the* $L^\infty\left([0,T],L^2(\Omega)\right)$*-error between the numerical approximation and the interpolant of the exact solution. Inspired by [81], where a Dirichlet problem is given as numerical example, we choose for our Neumann problem*

$$\begin{aligned} y_d(t,x_1,x_2) &= c_7 w_a(t,x_1,x_2) + c_8 w_b(t,x_1,x_2) + \\ &\quad + c_9 w_a(0,x_1,x_2) + c_{10} w_b(0,x_1,x_2) + \\ &\quad + c_{11} w_a(1,x_1,x_2) + c_{12} w_b(1,x_1,x_2) \\ \text{with } w_a(t,x_1,x_2) &= \mathrm{e}^{\frac{1}{3}\pi^2 t}\cos(\pi x_1)\cos(\pi x_2) \\ \text{and } w_b(t,x_1,x_2) &= \mathrm{e}^{-\frac{1}{3}\pi^2 t}\cos(\pi x_1)\cos(\pi x_2). \end{aligned} \tag{6.13}$$

c_1	c_2	c_3	$c_4 = c_5 = c_6$
$\frac{-5\left(5\,\mathrm{e}^{-\frac{1}{3}\pi^2}-6\right)}{-6+7\,\mathrm{e}^{\frac{1}{3}\pi^2}}$	5	$-\frac{1}{4}\frac{7+141\,\mathrm{e}^{\frac{1}{3}\pi^2}+7\left(\mathrm{e}^{\frac{1}{3}\pi^2}\right)^2-6-106\,\mathrm{e}^{-\frac{1}{3}\pi^2}}{-6+7\,\mathrm{e}^{\frac{1}{3}\pi^2}}$	$\frac{1}{4}$

(a) Coefficients for y.

c_7	c_8	c_9	$c_{10} = c_{11} = c_{12}$
$-\frac{5}{9}\frac{\left(9+35\nu\pi^4\right)\left(5\,\mathrm{e}^{-\frac{1}{3}\pi^2}-6\right)}{-6+7\,\mathrm{e}^{\frac{1}{3}\pi^2}}$	$5+\frac{175}{9}\nu\pi^4$	$4\cdot c_3\cdot c_{10}$	$\frac{1}{4}+\nu\pi^4$

(b) Coefficients $c_7,\ c_8, c_9,\ \ c_{10},\ c_{11},\ c_{12}$ for y_d.

Table 6.1.: Coefficients for the numerical Example 6.20.

The exact solution (y, p) of this optimal control problem can be represented by a linear combination of $w_a(t, x_1, x_2)$, $w_b(t, x_1, x_2)$, $w_a(0, x_1, x_2)$, $w_b(0, x_1, x_2)$, $w_a(1, x_1, x_2)$ and $w_b(1, x_1, x_2)$,

$$\begin{aligned}\bar{y}(t, x_1, x_2) = c_1 w_a(t, x_1, x_2) + c_2 w_b(t, x_1, x_2) + \\ + c_3 w_a(0, x_1, x_2) + c_4 w_b(0, x_1, x_2) + c_5 w_a(1, x_1, x_2) + c_6 w_b(1, x_1, x_2).\end{aligned} \tag{6.14}$$

The coefficients must be chosen such that the optimality system is satisfied. The choice of the coefficients is not unique, using Maple we computed the solution displayed in Table 6.1.

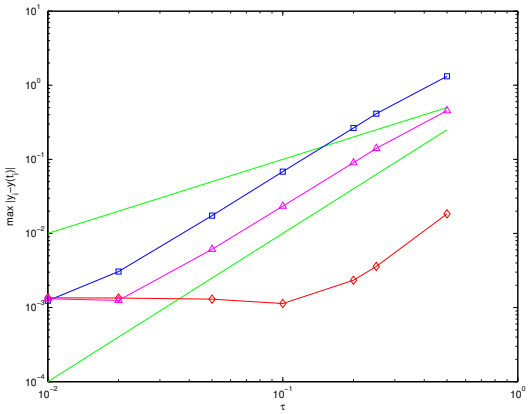
For clarity of presentation of the numerical results the convergence plots are split into two parts. On the left hand side we always plot the error in the state $\bar{y}$, on right hand side the error of the adjoint state $\bar{p}$.

We nicely observe second order convergence for different ν in Figure 6.4 for the example (6.12), (6.13), (6.14). As the same spatial discretization is used for all examples the error is dominated by this for small time step sizes. We see also that different problems are solved for different ν and that the error constants become larger for decreasing ν.

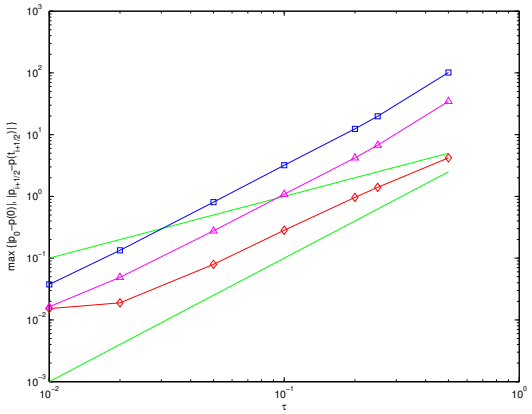
6.6.3. Terminal state tracking

As second numerical example we consider an example with terminal state tracking and no tracking of a desired state over the full time interval, i.e. $\alpha = 1$ and $\beta = 0$ in the cost functional of Problem 5.2.

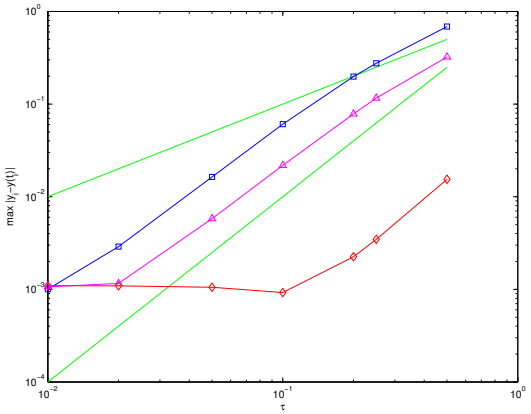
In this case all three numerical schemes (OC CN1), (OC CN2) and (OC G1) coincide, as the three schemes differ only in the discretization of the right hand side of the adjoint equation which refers to the tracking of a desired state over the full time interval.



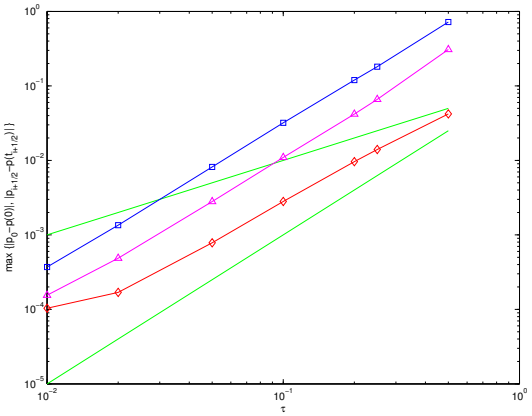
(a) Convergence of y for $\nu = 1$



(b) Convergence of p for $\nu = 1$



(c) Convergence of y for $\nu = 1 \cdot 10^{-2}$

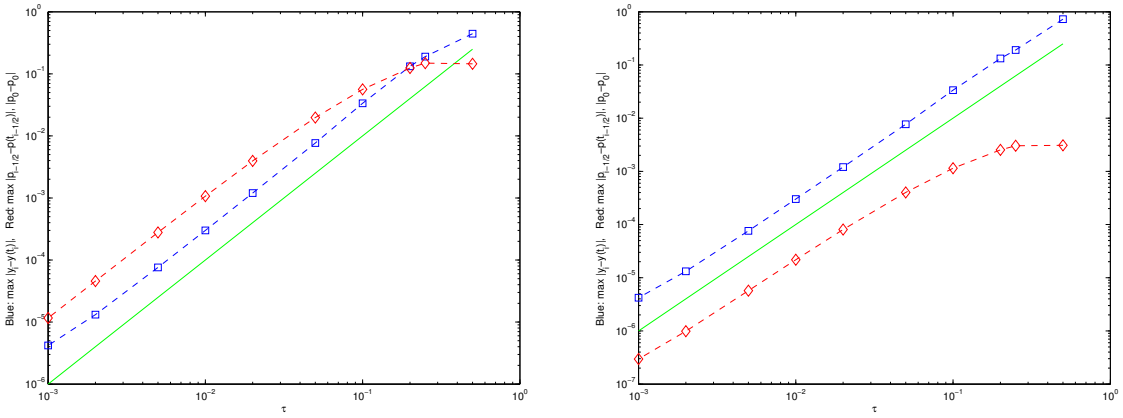


(d) Convergence of p for $\nu = 1 \cdot 10^{-2}$

Figure 6.4.: Plot of the error against the step size τ for different ν for the Example 6.20 Spatial discretization is for all time step sizes the same. On the left side the errors for the approximation of the state y and on the right side the errors for the approximation of the adjoint state p are plotted. The green lines indicate τ, τ^2, blue with square the scheme (OC CN1), red with diamonds the scheme (OC CN2)≡(OC G2) and magenta with triangles the scheme (OC G1).

$y_{0,1}$	$y_{D,1}$	$C_{1,1}$	$C_{2,1}$	$C_{3,1}$
1	1	$-\frac{-1+\mathrm{e}^{-\pi^2}}{2\nu\pi^2\,\mathrm{e}^{\pi^2}+\mathrm{e}^{\pi^2}-\mathrm{e}^{-\pi^2}}$	$\frac{-1+\mathrm{e}^{\pi^2}+2\nu\pi^2\,\mathrm{e}^{\pi^2}}{2\nu\pi^2\,\mathrm{e}^{\pi^2}+\mathrm{e}^{\pi^2}-\mathrm{e}^{-\pi^2}}$	$2\pi^2\nu C_{1,1}$

Table 6.2.: Coefficients for the numerical Example 6.20.



(a) Convergence for $\nu = 1$ (b) Convergence for $\nu = 0.001$

Figure 6.5.: Plot of the error against the step size τ for different ν for the example with data according to Table 6.2. For the spatial discretization linear FEM were chosen, the mesh parameter h was chosen independent of τ. The error for the state y is plotted in blue with square markers and the error of the adjoint state is plotted in red with diamond markers.

Example 6.20. *The problem for this example is given by*

$$\begin{aligned} \min \frac{1}{2}\|y(\cdot,T)-y_D\|_H^2+\frac{\nu}{2}\int_0^T \|u\|_{L^2(\Omega)}^2 \,\mathrm{d}\,t, & \\ y_{,t}-\Delta y = u & \quad \text{in } \Omega\times(0,T), \\ \frac{\partial}{\partial n}y = 0 & \quad \text{on } \partial\Omega\times(0,T), \\ y = v & \quad \text{in } \Omega\times\{0\}, \end{aligned}$$

with $\Omega = (0,1)$ *and* $T = 1$. *Further we choose* $v = \sqrt{2}\cos(\pi x)$ *and* $y_D = \sqrt{2}\cos(\pi x)$. *The solution has the representation*

$$\bar{y} = C_{1,1}\cos(\pi x)\,\mathrm{e}^{\pi t}+C_{2,1}\cos(\pi x)\,\mathrm{e}^{-\pi t}, \qquad \bar{p} = C_{3,1}\cos(\pi x)\,\mathrm{e}^{\pi t}.$$

The coefficients for this example can be found in Table 6.2.

For the spatial discretization we choose a finite element discretization where the mesh parameter h *was chosen independent of* τ. *The numerical experiments confirm nicely the order of convergence, see Figure 6.5.*

$y_{0,i}$	$y_{D,i}$	$C_{1,i}$	$C_{2,i}$	$C_{3,i}$
a_i	b_i	$\frac{-b_i+a_i\,\mathrm{e}^{-\lambda_i}}{-2\nu\lambda_i\,\mathrm{e}^{\lambda_i}-\mathrm{e}^{\lambda_i}+\mathrm{e}^{-\lambda_i}}$	$-\frac{-b_i+a_i\,\mathrm{e}^{\lambda_i}+2\nu\lambda_i a_i\,\mathrm{e}^{\lambda_i}}{-2\nu\lambda_i\,\mathrm{e}^{\lambda_i}-\mathrm{e}^{\lambda_i}+\mathrm{e}^{-\lambda_i}}$	$2\lambda_i\nu C_{1,i}$

Table 6.3.: Coefficients for the exact solution (6.16) to the data (6.15).

Remark 6.21. *Hou, Imanuvilov and Kwon give in their paper [64] for the optimal control problem*

$$\begin{aligned} &\min \frac{1}{2}\,\|y-y_D\|_H^2+\frac{\nu}{2}\int_0^T \|u\|_H^2\,\mathrm{d}\,t \\ &y_t-\nabla\cdot(A(x)\nabla y)=0 && \text{for } (t,x)\in Q,\\ &y(t,x)=0 && \text{for } (t,x)\in\Sigma,\\ &y(0,x)=v(x) && \text{for } (t,x)\in\Omega, \end{aligned}$$

the solution of optimal state $\bar{y}$ as eigenfunction expansion. They assume Dirichlet boundary conditions and that the function $A(x)$ is symmetric matrix-valued $C^1(\bar{\Omega})$-function that is uniformly positive definite. We extend their representation to the solution for the state $\bar{y}$ and the adjoint state $\bar{p}$ for more general symmetric operators and boundary conditions. Therefore we introduce the $L^2(\Omega)$-orthonormal eigenfunctions $\{e_i\}_{i=1}^{\infty}$ of the spatial operator A with the corresponding eigenvalues $\{\lambda_i\}_{i=0}^{\infty}$. Let the initial value and the desired state be given as eigenfunction expansions

$$v=\sum_{k=0}^{\infty} y_{0,k}e_i, \qquad\qquad y_D=\sum_{k=0}^{\infty} y_{D,k}e_i. \tag{6.15}$$

The optimal control problem decouples into problems for every eigenfunction e_i and has the solution

$$\bar{y}=\sum_{i=0}^{\infty} C_{1,i}e_i\,\mathrm{e}^{\lambda_i t}+C_{2,i}e_i\,\mathrm{e}^{-\lambda_i t}, \qquad\qquad \bar{p}=\sum_{i=0}^{\infty} C_{3,i}e_i\,\mathrm{e}^{\lambda_i t}\,. \tag{6.16}$$

For given $y_{0,i}$ and $y_{D,i}$ these coefficients can be computed to the values given in Table 6.3. These coefficients have been computed with Maple.

Remark 6.22. *In Remark 6.21 a representation of the solution as eigenfunction expansion is given and for our numerical example we have chosen an example where the exact solution is known. Nevertheless the numerical analysis of the optimal control problem with terminal observation is of interest. On the one hand the eigenfunctions are only known for special operators and special domains. And on the other hand the optimal control problem with terminal observation is only a special case of the optimal control Problem 5.2. For the more general choices of the parameter, i.e. $\alpha\neq 0$ and $\beta\neq 0$ we are not able to construct a solution in general.*

6.7. Summary

In this chapter we have introduced three time discretization schemes for the discretization of optimal control problems with parabolic partial differential equations.

The scheme (OC CN1) was derived with discretize-then-optimize approach with a Crank-Nicolson discretization of the state equation. On the other hand we have seen that the scheme (OC CN1) is also the Störmer-Verlet discretization of the optimality condition, so for this approach discretize-then-optimize and optimize-then-discretize lead to the same discrete scheme.

The discretization (OC CN2) was also introduced by the discretize-then-optimize approach. Later on we have seen, that this approach is also a Galerkin discretization with a quadrature rule or with additional variables. So we have also for this approach the commutation of discretization and optimization.

The third discretization (OC G1) is a Galerkin discretization, which was introduced by optimizing and then discretizing. We have no time stepping scheme introduced and not discussed whether discretization and optimization commute or this scheme.

For the scheme (OC CN1) we have proven second order convergence and we have proven that the discretization of the inner time steps of the schemes (OC CN2) and (OC G1) are second order perturbations of the scheme (OC CN1).

The numerical examples confirm the second order convergence.

7. Space time finite elements for approximation of the optimal state

Contents

In Chapter 6 we discussed Crank-Nicolson discretizations for the first discretize then optimize and the first optimize then discretize approach. These two approaches deal with the discretization of the optimal control problem and the optimality conditions, and end up in a system containing state variables y, adjoint state variables p and control variables u. But we have seen in Section 5.3 that we can eliminate two of the functions in the optimality conditions and discuss a equation only for the state y or the adjoint state p. In this chapter we discretize these $H^{(2,1)}(Q)$-elliptic equations directly with finite elements and transfer the ideas of the elimination of two variables to the discrete scheme (OC CN2).

This approach is of interest when we are just interested in the optimal control but not in the optimal state (as we have an optimal control problem and not the problem of the optimal state) or in the optimal state but not the optimal control (so we are solving the problem of the optimal state).

We discuss a conforming finite element and a mixed finite element method for this approach. We see that the mixed approximation can also be reached with a first-discretize-then-optimize or first-optimize-then-discretize approach for the optimality conditions.

For simplicity and clarity of presentation we will assume a self adjoint operator A, $M_D = M_d = M_u = G = M$ and only tracking of a state over the full space time cylinder and no terminal tracking, i.e. $\alpha = 0$ in the cost functional of the problem (5.1).

7.1. A Conforming Finite Element Method

As seen in Section 5.3 the solution of a parabolic optimal control problem is equivalent to the solution of an $H^{(2,1)}(Q)$-elliptic boundary value problem in the space-time domain. The corresponding equations are stated in Problem 5.22 and Problem 5.23.

For a spatially one dimensional domain a conforming discretization of $H^{(2,1)}(Q)$-elliptic equations with Hermite Lagrange tensor product finite elements has been investigated in Section 4.1.2. The boundary conditions in Section 4.1.2 were chosen so that they coincide with the boundary conditions for Problem 5.23 in the case of a desired state on the full space time domain and no desired terminal state. So the numerical Example 4.19 is also a numerical example for the discretization of optimal control problems as $H^{(2,1)}(Q)$-elliptic equations.

The only difference in boundary conditions for Problem 5.22, compared to the conditions of Problem 5.23, are inhomogeneities in the boundary conditions. The difference for a desired terminal state also appears only in the boundary condition. Therefore we do not discuss these problems here in detail.

7.2. Mixed finite element approximations

In this Section we introduce a mixed discretization for this equation and show that this discretization is connected to the Crank-Nicolson-discretization (OC CN2) of the previous Chapter.

7.2.1. Mixed discretization as Galerkin approximation

Mixed formulation

For a mixed discretization of the $H^{(2,1)}(Q)$-elliptic equation in Problem 5.23 we use the mixed formulation (5.23) for the state y of Remark 5.26, i.e.

$$\begin{aligned} A\bar{y} &= M\bar{z}, \\ M\bar{y} &= M\bar{w}, \\ -\nu M\bar{y}_{tt} + \nu A\bar{z} + \beta M\bar{w} &= \beta M y_d. \end{aligned}$$

The boundary conditions for this problem are

$$\begin{aligned} M\bar{y} &= Mv && \text{in } \Omega \times \{0\}, \\ M\bar{y}_t + A\bar{y} &= 0 && \text{in } \Omega \times \{T\}, \\ \bar{y} &= 0 && \text{on } \Sigma_1, && \frac{\partial \bar{y}}{\partial n} = 0 && \text{on } \Sigma_2, \\ \bar{z} &= 0 && \text{on } \Sigma_1, && \frac{\partial \bar{z}}{\partial n} = 0 && \text{on } \Sigma_2. \end{aligned}$$

As usual in the numerical analysis of boundary value problems, we assume that the Dirichlet boundary values are homogeneous and therefore $v = 0$. With test functions with $\phi(x, 0) = 0$ and partial integration in time, given by

$$\begin{aligned} \int_0^T -\nu \langle M\bar{y}_{tt}, \phi \rangle_{H\times H} \,\mathrm{d}\, t &= \int_0^T \nu \langle M\bar{y}_t, \phi_t \rangle_{H\times H} \,\mathrm{d}\, t \\ &\quad - \langle M\bar{y}_t(x,T), \phi(x,T) \rangle_{H\times H} + \langle M\bar{y}_t(x,0), \phi(x,0) \rangle_{H\times H} \\ &= \int_0^T \nu \langle M\bar{y}_t, \phi_t \rangle_{H\times H} \,\mathrm{d}\, t - \langle M\bar{y}_t(x,T), \phi(x,T) \rangle_{H\times H}, \end{aligned}$$

the weak form of the system is given by

$$\left.\begin{aligned}
\int_0^T a(\bar{y},\varphi)\,\mathrm{d}\,t &= \int_0^T \langle Mz,\varphi\rangle_{H\times H}\,\mathrm{d}\,t \\
&\quad \forall \varphi \in L^2(0,T;V), \\
\int_0^T \langle M\bar{y},\psi\rangle_{H\times H}\,\mathrm{d}\,t &= \int_0^T \langle Mw,\psi\rangle_{H\times H}\,\mathrm{d}\,t \\
&\quad \forall \psi \in L^2(0,T;H), \\
\int_0^T \langle My_d,\psi\rangle_{H\times H}\,\mathrm{d}\,t &= \int_0^T \langle Mw_d,\psi\rangle_{H\times H}\,\mathrm{d}\,t \\
&\quad \forall \psi \in L^2(0,T;H), \\
a(\bar{y}(x,T),\phi(x,T)) & \\
+\int_0^T \nu\langle M\bar{y}_t,\phi_t\rangle_{H\times H} + \nu a(\bar{z},\phi) + \beta\langle M\bar{w},\phi\rangle_{H\times H}\,\mathrm{d}\,t &= \int_0^T \beta\langle Mw_d,\phi\rangle_{H\times H}\,\mathrm{d}\,t \\
&\quad \forall \phi \in H^1(0,T;V):\ \phi(x,0)=0,
\end{aligned}\right\} \tag{7.1}$$

where for the solution $(\bar{w},\bar{y},\bar{z}) \in L^2(0,T;H)\times\{y\in H^1(0,T;V):\ y|_{t=0}=0\}\times L^2(0,T;V)$ holds. For the discretization of this system we choose a tensor product ansatz.

Structure of the matrices

Prior a discussion of the discretization of the mixed system (7.1), we investigate the general structure of the matrices of a tensor product Petrov-Galerkin finite element discretization. To that end let for the spatial discretization the set $\{\chi_i^t(x)\}_{i=0}^n$ be a basis for the test space and the set $\{\chi_i^a(x)\}_{i=0}^n$ a basis of the ansatz space. For the temporal discretization let the set $\{\rho_j^t(t)\}_{j=0}^N$ be a basis for the test space and the set $\{\rho_j^a(t)\}_{j=0}^N$ a basis of the ansatz space.

Tensor product bases for ansatz space are given by the set $\{\chi_i^a(x)\cdot\rho_j^a(t)\}_{i=0,j=0}^{n,N}$ and for the test space by the set $\{\chi_i^t(x)\cdot\rho_j^t(t)\}_{i=0,j=0}^{n,N}$, where every spatial ansatz function or spatial test function is multiplied with every temporal ansatz function or temporal test function, respectively. A function f in the ansatz space can be represented as

$$f = \sum_{j=0}^{N}\sum_{i=0}^{n} \chi_i^a(x)\rho_j^a(t)f_{ij}$$

with coefficients $f_{ij}\in\mathbb{R}$. For the assembly of the matrices of the weak form (7.1) we need the evaluation of integrals with some spatial operator K, in our case $K=A$ or $K=M$, applied to a function f of the ansatz space multiplied with some test function. In general we have to compute

$$\int_0^T \langle Kf,\chi_k^t(x)\rho_l^t(t)\rangle_{V^*\times V}\,\mathrm{d}\,t \tag{7.2}$$

for all combinations of $\chi_k^t\in\{\chi_i^t(x)\}_{i=0}^n$ and $\rho_l^t\in\{\rho_j^t(t)\}_{j=0}^N$. As the linear spatial operator K

does not depend on t this is for some fixed choice of $(k,l) \in \{0,\dots,n\} \times \{0,\dots,N\}$

$$\begin{aligned}\int_0^T \langle Kf, \chi_k^t(x)\rho_l^t(t)\rangle_{V^*\times V}\,\mathrm{d}\,t &= \int_0^T \left\langle K\left(\sum_{j=0}^{N}\sum_{i=0}^{n}\chi_i^a(x)\rho_j^a(t)f_{ij}\right), \chi_k^t(x)\rho_l^t(t)\right\rangle_{V^*\times V}\mathrm{d}\,t \\ &= \sum_{j=0}^{N}\left(\int_0^T \rho_j^a(t)\rho_l^t(t)\,\mathrm{d}\,t\cdot\sum_{i=0}^{n}\left\langle K\chi_i^a(x),\chi_k^t(x)\right\rangle_{V^*\times V} f_{ij}\right). \quad (7.3)\end{aligned}$$

Further let K_h and f_j be the matrix and the vector given by

$$K_h = \begin{pmatrix} K_{h,00} & \cdots & K_{h,0n} \\ \vdots & & \vdots \\ K_{h,n0} & \cdots & K_{h,nn}\end{pmatrix}, \qquad \text{with } K_{h,ki} = \langle K\chi_i^a(x), \chi_k^t(x)\rangle_{V^*\times V}, \quad (7.4)$$

$$f_j = \begin{pmatrix} f_{0j} \\ \vdots \\ f_{nj}\end{pmatrix}.$$

With this matrix and vector we can replace the inner sum of (7.3) by the $k-th$ line of a matrix vector product as

$$\begin{pmatrix} \sum_{i=0}^{n}\left\langle K\chi_i^a(x),\chi_0^t(x)\right\rangle_{V^*\times V} f_{ij} \\ \vdots \\ \sum_{i=0}^{n}\left\langle K\chi_i^a(x),\chi_n^t(x)\right\rangle_{V^*\times V} f_{ij}\end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n} K_{h,0i}f_{ij} \\ \vdots \\ \sum_{i=0}^{n} K_{h,ni}f_{ij}\end{pmatrix} = K_h f_j.$$

So we see that the computation of the spatial and temporal integrals decouples as

$$\begin{pmatrix} \int_0^T \langle Kf, \chi_0^t(x)\rho_l^t(t)\rangle\,\mathrm{d}\,t \\ \vdots \\ \int_0^T \langle Kf, \chi_n^t(x)\rho_l^t(t)\rangle\,\mathrm{d}\,t\end{pmatrix} = \sum_{j=0}^{N}\int_0^T \rho_j^a(t)\rho_l^t(t)\,\mathrm{d}\,t K_h f_j.$$

Therefore the tensor product structure of the discretization is also transferred to the structure of the matrices.

Discretization

After the analysis of the structure of the matrices we discuss the lowest order conforming ansatz and test spaces for discretization. The integrals for the temporal discretization can be computed easily, which is done in Appendix D. We choose the lowest order conforming test space and ansatz space for the temporal discretization, i.e.

$$\begin{aligned} y_\tau, \phi &\in \left\{ v \in \mathcal{C}(0,T;\mathcal{C}(0,X)) : \ v|_{t\in(t_i,t_{i+1})} \in \mathbb{P}_1(t_i,t_{i+1};V_h)\right\}, \\ w_\tau, z_\tau, \varphi, \psi &\in \left\{ L^2(0,T);\mathcal{C}(0,X) : \ v|_{t\in(t_i,t_{i+1})} \in \mathbb{P}_0(t_i,t_{i+1};V_h)\right\}. \end{aligned}$$

The full discretization of (7.1) with this choice is given by

$$\left.\begin{aligned}
A_h\frac{\bar{y}_{h,1}}{2} &= M_h\bar{z}_{h,\frac{1}{2}},\\
A_h\frac{\bar{y}_{h,i-1}+y_{h,i}}{2} &= M_h\bar{z}_{h,i-\frac{1}{2}},\\
&\text{for } i=2,\dots,N,\\
M_h\frac{\bar{y}_{h,1}}{2} &= M_h\bar{w}_{h,\frac{1}{2}},\\
M_h\frac{\bar{y}_{h,i-1}+\bar{y}_{h,i}}{2} &= M_h\bar{w}_{h,i-\frac{1}{2}},\\
&\text{for } i=2,\dots,N,\\
M_h\frac{y_{dh,i-1}+y_{dh,i}}{2} &= M_h\bar{w}_{dh,i-\frac{1}{2}},\\
&\text{for } i=1,\dots,N,\\
\nu M_h\frac{-y_{h,i-1}+2y_{h,i}-y_{h,i+1}}{\tau^2}+\nu A_h\frac{\bar{z}_{h,i-\frac{1}{2}}+\bar{z}_{h,i+\frac{1}{2}}}{2}&\\
+\beta M_h\frac{w_{h,i-\frac{1}{2}}+w_{h,i+\frac{1}{2}}}{2} &= \beta M_h\frac{w_{dh,i-\frac{1}{2}}+w_{dh,i+\frac{1}{2}}}{2},\\
&\text{for } i=1,\dots,N-1,\\
\nu M_h\frac{y_{h,N}-y_{h,N-1}}{\tau^2}+\nu A_h z_{h,N-\frac{1}{2}}+\beta M_h w_{h,N-\frac{1}{2}}+Ay_{h,N} &= \beta M_h w_{dh,N-\frac{1}{2}},
\end{aligned}\right\}\tag{7.5}$$

where the matrices A_h and M_h are the spatial finite element discretization of the operators A and M. For the lowest order spatial discretization one can choose continuous, piecewise linear finite elements.

7.2.2. Crank-Nicolson discretization as a mixed approximation

In Section 5.3 we eliminated two of the three unknown functions of an optimal control problem in the continuous setting. Now we repeat this approach for the discrete optimization problem (OC CN2).

We recall the full discretization of system (OC CN2) with terminal step (OC CN2*) for the case of A self adjoint, $M_d = M_D = M_u = G = M$, $\alpha = 0$ and $v = 0$

$$M_h\bar{y}_{h,0} = 0 \tag{7.6}$$

$$M_h\frac{\bar{y}_{h,i+1}-\bar{y}_{h,i}}{\tau}+A_h\frac{\bar{y}_{h,i+1}+\bar{y}_{h,i}}{2} = \frac{1}{\nu}M_h\bar{p}_{h,i+\frac{1}{2}} \tag{7.7}$$

$$\text{for } i=0,\dots,N-1,$$

$$M_h\frac{\bar{p}_{h,\frac{1}{2}}-\bar{p}_{h,0}}{\tau}-A_h\frac{\bar{p}_{h,\frac{1}{2}}}{2} = \beta M_h\frac{\frac{\bar{y}_{h,0}+\bar{y}_{h,1}}{2}-\frac{y_{dh,0}+y_{dh,1}}{2}}{2}, \tag{7.8}$$

$$\begin{aligned}
M_h\frac{\bar{p}_{h,i+\frac{1}{2}}-\bar{p}_{h,i-\frac{1}{2}}}{\tau}-A_h\frac{\bar{p}_{h,i+\frac{1}{2}}+\bar{p}_{h,i-\frac{1}{2}}}{2} &=\\
= \beta M_h\frac{\frac{\bar{y}_{h,i}+\bar{y}_{h,i-1}}{2}-\frac{y_{dh,i-1}+y_{dh,i}}{2}}{2}+\beta M_h\frac{\frac{\bar{y}_{h,i}+\bar{y}_{h,i+1}}{2}-\frac{y_{dh,i+1}+y_{dh,i}}{2}}{2}&
\end{aligned}\tag{7.9}$$

$$\text{for } i=1,\dots,N-2,$$

$$M_h \frac{\bar{p}_{h,N} - \bar{p}_{h,N-\frac{1}{2}}}{\tau} - A_h \frac{\bar{p}_{h,N-\frac{1}{2}}}{2} = \beta M_h \frac{\frac{\bar{y}_{h,N-1}+\bar{y}_{h,N}}{2} - \frac{y_{dh,N}+y_{dh,N-1}}{2}}{2} \tag{7.10}$$

$$M_h p_{h,N} = 0, \tag{7.11}$$

where the matrices A_h and M_h are the discretizations of the operators A and M. Now we use (OC CN2) for the definition of the adjoint state and eliminate state and control of the optimality system. We use the state equation (7.7) as definition for the adjoint state

$$\frac{1}{\nu} M_h \bar{p}_{h,i+\frac{1}{2}} = M_h \frac{\bar{y}_{h,i+1} - \bar{y}_{h,i}}{\tau} + A_h \frac{\bar{y}_{h,i+1} + \bar{y}_{h,i}}{2} \qquad \text{for } i = 0, \dots, N-1.$$

and insert this to the adjoint equation (7.9)–(7.11). As the first half step of the adjoint equation (7.8) is not needed for a solution of of the state y (see Section 6.6), we do not use this equation.

For the inner time steps (7.9) this elimination of the adjoint state yields

$$\begin{aligned}
&\nu \frac{M_h \frac{\bar{y}_{h,i+1}-\bar{y}_{h,i}}{\tau} + A_h \frac{\bar{y}_{h,i+1}+\bar{y}_{h,i}}{2} - M_h \frac{\bar{y}_{h,i}-\bar{y}_{h,i-1}}{\tau} - A_h \frac{\bar{y}_{h,i}+\bar{y}_{h,i-1}}{2}}{\tau} - \\
&- \nu A_h \frac{\frac{\bar{y}_{h,i+1}-\bar{y}_{h,i}}{\tau} + M_h^{-1} A_h \frac{\bar{y}_{h,i+1}+\bar{y}_{h,i}}{2} + \frac{\bar{y}_{h,i}-\bar{y}_{h,i-1}}{\tau} + M_h^{-1} A_h \frac{\bar{y}_{h,i}+\bar{y}_{h,i-1}}{2}}{2} = \\
&= \beta M_h \frac{\frac{\bar{y}_{h,i}+\bar{y}_{h,i-1}}{2} - \frac{y_{dh,i-1}+y_{dh,i}}{2}}{2} + \beta M_h \frac{\frac{\bar{y}_{h,i}+\bar{y}_{h,i+1}}{2} - \frac{y_{dh,i+1}+y_{dh,i}}{2}}{2}
\end{aligned}$$

and after simplification

$$\begin{aligned}
-\nu M_h \frac{\bar{y}_{h,i-1} - 2\bar{y}_{h,i} + \bar{y}_{h,i+1}}{\tau^2} + \nu A_h M_h^{-1} A_h \frac{\bar{y}_{h,i+1} + 2\bar{y}_{h,i} + \bar{y}_{h,i-1}}{4} \\
+ \beta M_h \frac{\bar{y}_{h,i+1} + 2\bar{y}_{h,i} + \bar{y}_{h,i-1}}{4} = \beta M_h \frac{y_{dh,i+1} + 2y_{dh,i} + y_{dh,i-1}}{4}.
\end{aligned} \tag{7.12}$$

To avoid the inversion of the matrix M_h we introduce $\bar{z}_{h,i+\frac{1}{2}}$ as the solution of

$$A_h \frac{\bar{y}_{h,i+1} + \bar{y}_{h,i}}{2} = M_h \bar{z}_{h,i+\frac{1}{2}}, \qquad \text{for } i = 0 \dots, N-2,$$

so that

$$A \frac{\bar{y}_{i+1} + 2\bar{y}_i + \bar{y}_{i-1}}{4} = M \frac{\bar{z}_{i+\frac{1}{2}} + \bar{z}_{i-\frac{1}{2}}}{2}.$$

Similarly we introduce $\bar{w}_{\frac{1}{2}}$ and $w_{d,\frac{1}{2}}$ as solution of

$$M_h \frac{\bar{y}_{h,i+1} + \bar{y}_{h,i}}{2} = M_h \bar{w}_{h,i+\frac{1}{2}}, \qquad \text{for } i = 0 \dots, N-2.$$

$$M_h \frac{y_{dh,i+1} + y_{dh,i}}{2} = M_h w_{dh,i+\frac{1}{2}}. \qquad \text{for } i = 0 \dots, N-2.$$

So the inner time steps (7.12) are equivalent to the solution of the mixed system

$$\left.\begin{aligned} A_h\frac{\bar y_{h,i+1}+\bar y_{h,i}}{2} &= M_h\bar z_{h,i+\frac12},\\ \text{for } i&=0,\dots N-2\\ M_h\frac{\bar y_{h,i+1}+\bar y_{h,i}}{2} &= M_h\bar w_{h,i+\frac12},\\ \text{for } i&=0,\dots N-2\\ M_h\frac{y_{dh,i+1}+y_{dh,i}}{2} &= M_h w_{dh,i+\frac12},\\ \text{for } i&=0,\dots N-2\\ \nu M_h\frac{-\bar y_{h,i-1}+2\bar y_{h,i}-\bar y_{h,i+1}}{\tau^2}+\nu A_h\frac{\bar z_{h,i+\frac12}+\bar z_{h,i-\frac12}}{2}&\\ +\beta M_h\frac{\bar w_{h,i+\frac12}+\bar w_{h,i-\frac12}}{2} &= \beta M_h\frac{w_{dh,i+\frac12}+w_{dh,i-\frac12}}{2}.\\ \text{for } i&=1,\dots N-1 \end{aligned}\right\} \tag{7.13}$$

These are the corresponding equations of the mixed formulation (7.5) of the previous section.

For the final half time step (7.11) the elimination gives

$$\begin{aligned} M_h\frac{p_{h,N}}{\tau}-\nu\frac{M_h\frac{\bar y_{h,N}-\bar y_{h,N-1}}{\tau}+A_h\frac{\bar y_{h,N}+\bar y_{h,N-1}}{2}}{\tau}-\nu A_h\frac{\frac{\bar y_{h,N}-\bar y_{h,N-1}}{\tau}+M_h^{-1}A_h\frac{\bar y_{h,N}+\bar y_{h,N-1}}{2}}{2}&\\ =\beta M_h\frac{\frac{\bar y_{h,N-1}+\bar y_{h,N}}{2}-\frac{y_{dh,N}+y_{dh,N-1}}{2}}{2},&\\ M_h p_{h,N}=0.& \end{aligned}$$

which is equivalent to

$$\begin{aligned} \nu M_h\frac{\bar y_{h,N}-\bar y_{h,N-1}}{\tau^2}+A_hM_h^{-1}A_h\frac{\bar y_{h,N}+\bar y_{h,N-1}}{4}+\nu A_h\frac{\bar y_{h,N}}{\tau}+\beta M_h\frac{\bar y_{h,N-1}+\bar y_{h,N}}{4}=&\\ =\beta M_h\frac{y_{dh,N}+y_{dh,N-1}}{4}.& \end{aligned}$$

As before we write this as mixed system

$$\left.\begin{aligned} A_h\frac{\bar y_{h,N}+\bar y_{h,N-1}}{2} &= M_h\bar z_{h,N-\frac12},\\ M_h\frac{\bar y_{h,N}+\bar y_{h,N-1}}{2} &= M_h\bar w_{h,N-\frac12},\\ M_h\frac{y_{dh,N}+y_{dh,N-1}}{2} &= M_h w_{dh,N-\frac12},\\ -\nu M_h\frac{\bar y_{h,N-1}-\bar y_{h,N}}{\tau^2}+\nu A_h\frac{\bar z_{h,N-\frac12}}{2}+\nu A_h\frac{\bar y_{h,N}}{\tau}+\beta M_h\frac{\bar w_{h,N-\frac12}}{2} &= \beta M_h\frac{w_{dh,N-\frac12}}{2} \end{aligned}\right\} \tag{7.14}$$

So the system of the Crank-Nicolson discretization (OC CN2) can be written as mixed

problem consisting of the systems of equations (7.13) and (7.14). Altogether we have to solve

$$\left.\begin{aligned}
A_h\frac{\bar{y}_{h,1}}{2} &= M_h\bar{z}_{h,\frac{1}{2}},\\
A_h\frac{\bar{y}_{h,i}+\bar{y}_{h,i+1}}{2} &= M_h\bar{z}_{h,i+\frac{1}{2}},\\
&\text{for } i=1,\dots,N-1,\\
M_h\frac{\bar{y}_{h,1}}{2} &= M_h\bar{w}_{h,\frac{1}{2}},\\
M_h\frac{\bar{y}_{h,i+1}+\bar{y}_{h,i}}{2} &= M_h\bar{w}_{h,i+\frac{1}{2}}\\
&\text{for } i=1,\dots,N-1,\\
M_h\frac{y_{dh,i+1}+y_{dh,i}}{2} &= M_h w_{dh,i+\frac{1}{2}},\\
&\text{for } i=1,\dots,N-1,\\
-\nu M_h\frac{\bar{y}_{h,i-1}-2\bar{y}_{h,i}+\bar{y}_{h,i+1}}{\tau^2}+\nu A_h\frac{\bar{z}_{h,i+\frac{1}{2}}+\bar{z}_{h,i-\frac{1}{2}}}{2}&\\
+\beta M_h\frac{\bar{w}_{h,i+\frac{1}{2}}+\bar{w}_{h,i-\frac{1}{2}}}{2} &= \beta M_h\frac{w_{dh,i+\frac{1}{2}}+w_{dh,i-\frac{1}{2}}}{2},\\
&\text{for } i=1,\dots,N-1,,\\
-\nu M_h\frac{\bar{y}_{h,N-1}-\bar{y}_{h,N}}{\tau^2}+\nu A_h\frac{\bar{z}_{h,N-\frac{1}{2}}}{2}&\\
+\nu A_h\frac{\bar{y}_{h,N}}{\tau}+\beta M_h\frac{\bar{w}_{h,N-\frac{1}{2}}}{2} &= \beta M_h\frac{w_{dh,N-\frac{1}{2}}}{2}.
\end{aligned}\right\} \tag{7.15}$$

This system is just the same system, which was obtained by the mixed discretization of the $H^{(2,1)}(Q)$-elliptic equation for y in (7.5). So we have seen the following equivalence.

Theorem 7.1 (Equivalence of (OC CN2) and mixed finite element discretization)**.** *For a optimal control problem with parabolic partial differential equations with tracking over the full space time cylinder and no tracking of the terminal state the Crank-Nicolson discretization* (OC CN2) *is equivalent to a finite element discretization of the mixed problem for the state* (7.1) *given by* (7.5)*.*

Remark 7.2. *With this equivalence it is possible to prove the convergence of the numerical scheme* (OC CN2) *with the convergence of the mixed discretization of the $H^{(2,1)}(Q)$-elliptic equation for the state y given in Problem 5.23 and vice versa.*

The convergence of the mixed approximation and the convergence of the time stepping scheme (OC CN2) *are still open. In Section 6.4 it was only shown that the inner time steps of the scheme* (OC CN2) *are a pertubation of the scheme* (OC CN1) *of order τ^2.*

Remark 7.3. *We chose to eliminate the controls $\bar{u}_{i+\frac{1}{2}}$ and the adjoint states $\bar{p}_{i+\frac{1}{2}}$ and not the controls $\bar{u}_{i+\frac{1}{2}}$ and the states $\bar{y}_i$ in the numerical scheme* (OC CN2) *as the the elimination of the adjoint state can be done without solving a non-trivial linear system.*

7.2.3. No mixed formulation based on (OC CN1)

In Section 7.2.2 we have discussed that the discretization scheme (OC CN2) can be interpreted as mixed Galerkin scheme if we eliminate the adjoint state. One could ask if the same is also possible for the scheme (OC CN1), in which the adjoint state or the state can be eliminated. In this section we will see where the problems for a identification of the scheme (OC CN1) with a Galerkin scheme are. Therefore we recall the system (OC CN1) for the case $M_d = M_D = M_u = G = M$ and $\alpha = 0$

$$\begin{aligned}
M_h \bar{y}_0 &= M_h v, \\
M_h \frac{\bar{y}_{h,i+1} - \bar{y}_{h,i}}{\tau} + A_h \frac{\bar{y}_{h,i+1} + \bar{y}_{h,i}}{2} &= \frac{1}{\nu} M_h \bar{p}_{h,i+\frac{1}{2}}, \qquad (7.16)\\
&\text{for } i = 0, \dots, N-1, \\
M_h \frac{\bar{p}_{h,\frac{1}{2}} - \bar{p}_{h,0}}{\tau} - A_h \frac{\bar{p}_{h,\frac{1}{2}}}{2} &= \beta M_h \frac{\bar{y}_{h,0} - y_{hd,0}}{2}, \\
M_h \frac{\bar{p}_{h,i+\frac{1}{2}} - \bar{p}_{h,i-\frac{1}{2}}}{\tau} - A_h \frac{\bar{p}_{h,i+\frac{1}{2}} + \bar{p}_{h,i-\frac{1}{2}}}{2} &= \beta M_h \bar{y}_{h,i} - \beta M y_{dh,i} \qquad (7.17)\\
&\text{for } i = 0, \dots, N-2, \\
-M_h \frac{\bar{p}_{h,N-\frac{1}{2}}}{\tau} - A_h \frac{\bar{p}_{h,N-\frac{1}{2}}}{2} &= \beta M_h \frac{\bar{y}_{h,N} - y_{dh,N}}{2},
\end{aligned}$$

where A_h and M_h are the spatial discretizations of A and M. For the discussion of the problems which occur if we interpret the discretization (OC CN1) as mixed formulation, we focus on the equations for the inner time nodes (7.16) and (7.17).

As in the previous section we have, due to (7.16)

$$\frac{1}{\nu} M_h \bar{p}_{h,i+\frac{1}{2}} = M_h \frac{\bar{y}_{h,i+1} - \bar{y}_{h,i}}{\tau} + A_h \frac{\bar{y}_{h,i+1} + \bar{y}_{h,i}}{2}.$$

If we insert this definition of $\bar{p}_{h,i+\frac{1}{2}}$ into the inner time steps of the adjoint equation (7.17) this yields

$$-M_h \frac{\bar{y}_{h,i-1} - 2\bar{y}_{h,i} + \bar{y}_{h,i+1}}{\tau^2} + A_h M_h^{-1} A_h \frac{y_{h,i-1} + 2y_{h,i} + y_{h,i+1}}{4} + M_h y_{h,i} = -M_h y_{dh,i}.$$

This is equivalent to the mixed formulation

$$\begin{aligned}
A_h \frac{\bar{y}_{h,i+1} + \bar{y}_{h,i}}{2} &= M_h \bar{z}_{h,i+\frac{1}{2}} \qquad (7.18)\\
M_h \bar{y}_{h,i} &= M_h \bar{w}_{h,i} \qquad (7.19)\\
-M_h \frac{\bar{y}_{h,i-1} - 2\bar{y}_{h,i} + \bar{y}_{h,i+1}}{\tau^2} + A_h \frac{\bar{z}_{h,i+\frac{1}{2}} + \bar{z}_{h,i-\frac{1}{2}}}{2} + M_h \bar{w}_{h,i} &= -M y_{dh,i} \qquad (7.20)
\end{aligned}$$

We want now to reproduce this discretization as finite element discretization of the mixed

system (7.1), given by

$$\begin{aligned}
\int_0^T a(\bar{y}_h, \phi)\,\mathrm{d}\,t &= \int_0^T \langle M_h z_h, \phi\rangle_{H\times H}\,\mathrm{d}\,t,\\
\int_0^T \langle M_h \bar{y}_h, \psi\rangle_{H\times H}\,\mathrm{d}\,t &= \int_0^T \langle M_h \bar{w}_h, \psi\rangle_{H\times H}\,\mathrm{d}\,t,\\
\int_0^T \langle M_h y_{dh}, \psi\rangle_{H\times H}\,\mathrm{d}\,t &= \int_0^T \langle M_h w_{dh}, \psi\rangle_{H\times H}\,\mathrm{d}\,t,\\
a(\bar{y}_h(x,T), \varphi(x,T))&\\
+\int_0^T \nu\langle M_h \bar{y}_{h,t}, \varphi_t\rangle_{H\times H} + \nu a(\bar{z}_h, \varphi) + \beta\langle M_h \bar{w}_h, \varphi\rangle_{H\times H}\,\mathrm{d}\,t &= \int_0^T \beta\langle M_h w_{dh}, \varphi\rangle_{H\times H}\,\mathrm{d}\,t.
\end{aligned}$$

For a conforming ansatz we have to choose a piecewise linear and continuous ansatz in time for $\bar{y}$ and a piecewise linear and continuous ansatz for φ. With a picewise linear and continuous anstatz for this two functions we are able to produce the first term of (7.20). The choice of piecewise constant functions for z and ϕ is also quite clear. This implies (7.18) and the second term of (7.20).

For the reproduction of (7.18)–(7.20) as finite element method we would still need a set of (polynomial) test functions ψ and a set of (polynomial) ansatz functions for $\bar{w}$, so that

$$M_h \bar{y}_{h,i}\tau = \int_{\mathrm{supp}(\psi_i)} \langle M_h \bar{y}_h, \psi_i\rangle_{H\times H}\,\mathrm{d}\,t = \int_{\mathrm{supp}(\psi_i)} \langle M_h \bar{w}_h, \psi_i\rangle_{H\times H}\,\mathrm{d}\,t = M_h \bar{w}_i\tau \tag{7.21}$$

$$\int_{t_i}^{t_{i+1}} \langle M_h \bar{w}_h, \varphi_i\rangle_{H\times H}\,\mathrm{d}\,t = M_h \bar{y}_{h,i}\tau \tag{7.22}$$

holds with piecwise linear test functions φ_i.

If we choose a piecewise linear and continuous ansatz for $\bar{w}_h$, the condition (7.21) would imply the identity $\bar{w}_{h,i} = \bar{y}_{h,i}$. But as the integral of piecewise linear and continuous ansatz functions with piecewise linear and continuous test functions is given by (see also Appendix D)

$$\int_{t_{i-1}}^{t_{i+1}} \langle M_h \bar{y}_h, \varphi\rangle_{H\times H}\,\mathrm{d}\,t = \frac{\tau}{6} M_h \bar{y}_{h,i-1} + \frac{4}{6}\tau M_h \bar{y}_{h,i} + \frac{\tau}{6} M_h \bar{y}_{h,i+1},$$

this choice does not lead to 7.22.

If we choose on the other hand a piecewise constant ansatz for $\bar{w}_h$ with $\bar{w}_h|_{(t_i,t_{i+1})} = \bar{w}_{h,i+\frac{1}{2}}$, the integration would yield

$$\int_{t_{i-1}}^{t_{i+1}} \langle M_h \bar{w}_h, \varphi\rangle_{H\times H}\,\mathrm{d}\,t = \frac{\tau}{2} M_h \bar{w}_{h,i-\frac{1}{2}} + \frac{\tau}{2} M_h \bar{w}_{h,i+\frac{1}{2}},$$

which is also not equivalent to 7.22.

So it is not clear if such sets of polynomial test and ansatz functions exists, which fulfill (7.21) and (7.22) for piecewise linear and continuous functions $\bar{y}_h$ and φ_i. So we have at this point no equivalence of (OC CN1) and a mixed finite element discretization of a corresponding $H^{(2,1)}(Q)$-elliptic equation.

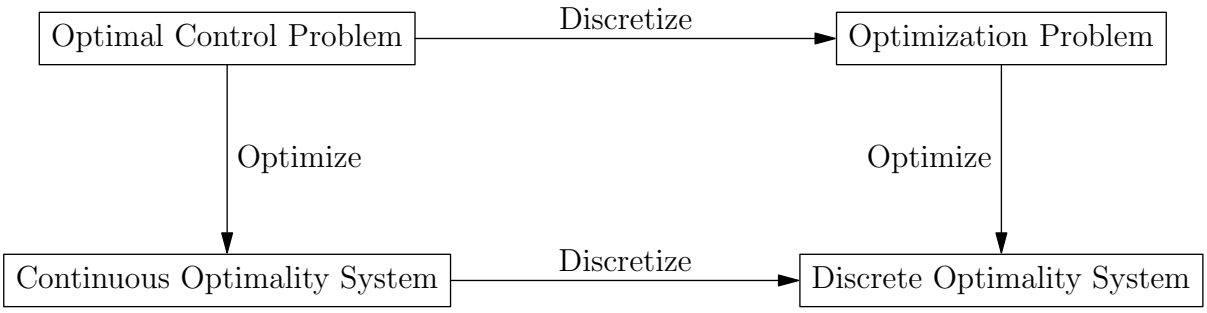


Figure 7.1.: For (OC CN1) optimization and discretization commute.

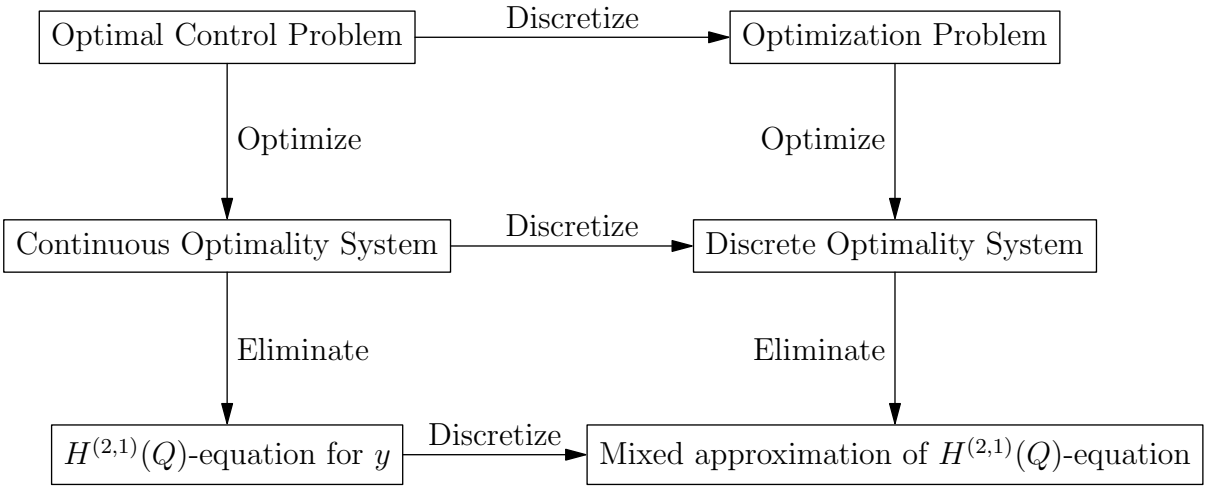


Figure 7.2.: For (OC CN2) discretization and elimination commute additionally.

7.3. Summary

Now we can review the paths in the graph of Figure 5.1 on page 73.

For the scheme (OC CN1) we have seen in Chapter 6 that optimization and discretization commute, but in the previous section we did not find a mixed discretization which is equivalent to the elimination of adjoint state and control in (OC CN1). So for this scheme only discretization and optimization commute, see Figure 7.1.

In Chapter 6 we have also observed that optimization and discretization commute for the scheme (OC CN2). Furthermore in Section 7.2.2 it was shown that the scheme (OC CN2) is also equivalent to the mixed discretization of the $H^{(2,1)}(Q)$-elliptic boundary value problem (5.23) for the optimal state $\bar{y}$. So for this scheme also the elimination of the adjoint state and the control commute with the discretization, see Figure 7.2.

8. Conclusions and outlook

Contents

8.1. Conclusions

In this thesis we have discussed several topics which are interconnected. The main goal was to demonstrate that there are relevant optimal control problems with parabolic partial differential equations and that there are methods for the solution of optimal control problems, which provide second or higher order of convergence.

We take the conclusions for the different topics separately.

Hydration of concrete.

First we focused on the hydration of concrete. We have seen that there are optimal control problems with parabolic partial differential equations of interest in applications. For some optimal control problems in this context we have proposed to discretize the problem and solve this problem with a solver for finite dimensional problems which considers the optimal control problem as finite dimensional optimization problem. For a further analysis of optimal control problems we did not discuss optimal control problems with a semilinear parabolic equation, but the still challenging optimal control problems with linear parabolic equations.

Functional analysis and numerical analysis.

In preparation of the discussion of parabolic optimal control problems, we discussed basic facts from functional analysis and have proven an a priori regularity estimate for an $H^{(2,1)}(Q)$-elliptic boundary value problem. Further we have shown an a priori estimate for a tensor product conforming finite element approximation for this equations.

Discretization of parabolic optimal control problems.

During the discussion of Crank-Nicolson and Störmer-Verlet discretizations of optimal control problems with parabolic partial differential equations, we have seen the importance of tailored approximations with different approximation of the state and the control. Due to this choice of discretization we were able to prove second order convergence. We presented a family of discretization schemes, and discussed the equivalence of the discretize-then-optimize approach and the optimize-then-discretize approach. Further we have seen that some schemes are also Galerkin schemes and can bee seen as mixed discretization of a semi-elliptic equation.

8.2. Outlook

As always in science every question answered poses new, still unanswered, questions. Again these questions are sorted by topic.

Hydration of concrete.

We have reviewed a model of the hydration of young concrete. The open question is the direct discussion and discretization of the connected full optimal control problems with control and state constraints. During this it may also be necessary to use a more complicated model for the mechanics of concrete.

Numerical analysis.

The numerical analysis of semi-elliptic partial differential equations was restricted to the case of a spatial one dimensional domain. The question about error estimates for higher spatial dimensions remains open. With two dimensional domain the space time domain Q is a three dimensional domain and with three spatial dimension the space time domain is even four dimensional. The Sobolev embedding theorem provides the continuous embedding $H^k(\Omega) \subseteq C^l(\bar{\Omega})$ for $k - l > \frac{d}{2}$ [131, Theorems 6.2 and 6.2]. So for the proof of an interpolation error one would need higher regularity assumption or the use of other techniques as quasi-interpolation.

The convergence of the pure time discretization of the Crank-Nicolson scheme with less regularity was proven by Schieweck [113]. Schieweck proves the second order convergence for parabolic partial differential equations with respect to the time discretization only if the right hand side is evaluated exactly. As his proof uses the usual duality technique of Theorem 4.7 this result can be transferred also to the case of an approximation of the right hand side. As Schieweck discusses only the time discretization error, one needs still the discussion of the spatial discretization error at this point if one considers the case of reduced regularity.

Discretization of parabolic optimal control problems.

In our discussion of the discretization error of the Crank-Nicolson discretization we have used a proof which discusses the Crank-Nicolson scheme as time stepping scheme. If we adopt the proof of [113] and use a similar technique for the convergence of the adjoint state, which is not trivial, one could decrease the regularity assumptions. This would be of interest if one discusses a case where the optimal solution has less regularity. Reasons for this could be less regular initial or desired terminal states. The Crank-Nicolson scheme in the case of irregular initial data has been discussed by Østerby [95]. As the Crank-Nicolson discretization can also bee seen as discontinuous Galerkin discretization, hence one could also think about the adoption of the results of Schötzau and Schwab [117] for discontinuous Galerkin time discretizations.

Non-convex polygonal domains might be another reason for a less regular solutions. Also in the case of optimal control problems with control (or even state) constraints, the optimal solution has reduced regularity in comparison with the unconstrained case.

Moreover, we have only discussed distributed control, so the discretization and convergence of boundary control problems remains another open question.

A. Riemann-Stieltjes integral

For the introduction of the Riemann-Stieltjes integral we follow the ideas of the textbook [112, Chapter 6] by Rudin with extensions found in textbook [119, Chapter 1] by Smirnow.

Definition A.1 (Riemann-Stieltjes integral)**.** *[112, Definition 6.2] Let P be a partition with the property*

$$P = \{x_i \in [a,b], i = 0, \dots, n : \ a = x_0 \leq x_1 \leq \cdots \leq x_n = b\}\,.$$

For a monotone increasing function α, which we is bounded on the interval $[a,b]$, we introduce the Riemann-Stieltjes upper and lower sums as

$$S(P,f,\alpha) = \sum_{i=1}^{N} \left(\sup_{x \in (x_{i-1}, x_i)} f(x) \right) \cdot (\alpha(x_i) - \alpha(x_{i-1}))\,,$$
$$s(P,f,\alpha) = \sum_{i=1}^{N} \left(\inf_{x \in (x_{i-1}, x_i)} f(x) \right) \cdot (\alpha(x_i) - \alpha(x_{i-1}))\,.$$

If the infimum over all possible partitions of the upper sum and the supremum of the lower sum coincide, i.e.

$$\inf_P S(P,f,\alpha) = \sup_P s(P,f,\alpha),$$

then we define the Riemann-Stieltjes integral as

$$\int_a^b f(x)\,\mathrm{d}\,\alpha(x) = \inf_P S(P,f,\alpha).$$

Remark A.2. *In the Definition A.1 we follow the definition of the Riemann-Stieltjes integral in [112, Definition 6.2]. But the Riemann-Stieltjes integral can be defined for a wider class of functions α as in this definition. As shown in [119, Chapter I.9] it is sufficient, that the function α has bounded variation, i.e.*

$$\lim_P \sum_{x_i \in P} |\alpha(x_k) - \alpha(x_{k+1})| < \infty,$$

where the limit is taken over all partitions P of the interval $[a,b]$. This can be easily understood as

$$\int_a^b f(x)\,\mathrm{d}\,(\alpha_1(x) + \alpha_2(x)) = \int_a^b f(x)\,\mathrm{d}\,\alpha_1(x) + \int_a^b f(x)\,\mathrm{d}\,\alpha_2(x),$$

(see [112, Theorem 6.12 (e)]) and as every function with bounded variation is the difference of two monotone increasing functions [119, Theorem I.8.5]. So we have for the function $\alpha(x) = \alpha_1(x) - \alpha_2(x)$*, that the Riemann-Stieltjes integral can expressed a*

$$\int_a^b f(x)\,\mathrm{d}\,\alpha(x) = \int_a^b f(x)\,\mathrm{d}\,\alpha_1(x) - \int_a^b f(x)\,\mathrm{d}\,\alpha_2(x),$$

as seen in [119, Chapter 1.9].

For the connection to the Riemann integral we recall the following Theorem.

Theorem A.3 (Connection of Riemann-Stieltjes integral and Riemann integral)**.** *[112, Theorem 6.17] For a monotone increasing function* α *with Riemann integrable derivative* α' *and a bounded function* f *on the interval* $[a, b]$ *the following both integrals coincide*

$$\int_a^b f(x)\,\mathrm{d}\,\alpha(x) = \int_a^b f(x)\alpha'(x)\,\mathrm{d}\,x.$$

B. A model for cooling pipes

In this appendix we discuss the model of Gehrmann [44] for the cooling pipes. The author of this PhD thesis was involved in the master's thesis [44] as supervisor.

For the development of a physical model of a concrete wall with cooling pipes, as sketched in Figure B.1a, we discuss only a small area around a pipe, Figure B.1b. As the wall of the pipe is rather small compared with the pipe and the concrete part, we neglect the wall of the pipe and assume that the rigid concrete and the fluid in the cooling pipes have direct contact (see Figure B.1c). The differential equations in the rigid body and in the fluid in the cooling pipe are well known, see e.g. the text book of Larsson and Thomée [76]. In the rigid part Ω_2 we have the usual heat equation

$$\varrho c \frac{\partial y_2}{\partial t} - \nabla \cdot (\lambda \nabla y_2) = p,$$

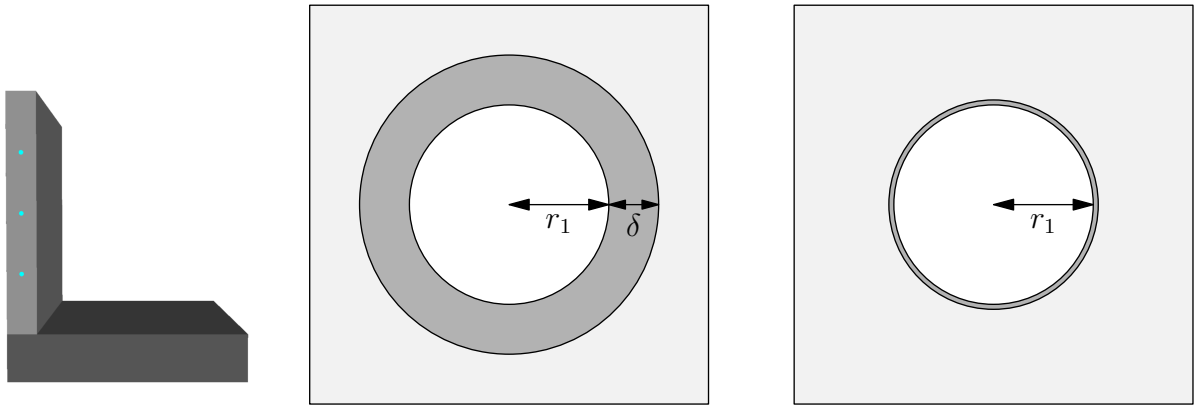
where ϱ is the density, c the heat capacity, λ the thermal conductivity and f the heat introduced by the hydration. In the fluid in the pipes the heat is also transported by convection, therefore the equation for the heat distribution in the pipe is

$$\varrho c \frac{\partial y_1}{\partial t} - \nabla \cdot (\lambda \nabla y_1) + \nabla \cdot (c \varrho \vec{v} y_1) = 0,$$

with the velocity field $\vec{v}$ of the fluid. The interesting part in the derivation of the equation are not the equation in the different domains, but the conditions on the interface. Therefore let Ω_d be an arbitrary smooth test domain which consist of two nonempty parts $\Omega_{d,1} = \Omega_d \cap \Omega_1$ and $\Omega_{d,2} = \Omega_d \cap \Omega_2$. For simplicity we can assume that these parts $\Omega_{d,1}$ and $\Omega_{d,2}$ are domains, i.e. open and connected. For the domain Ω_d the conservation of energy holds, i.e. the temporal changes of the energy in all test domains Ω_d equals the heat transfer over the boundary and the internal heat sources, so that

$$\frac{\mathrm{d}}{\mathrm{d}\, t} \int_{\Omega_d} e \,\mathrm{d}\, \omega = - \int_{\partial \Omega_d} j \cdot \vec{n} \,\mathrm{d}\, s + \int_{\Omega_d} p \,\mathrm{d}\, \omega, \tag{B.1}$$

with the energy density e, the heat flux f and the heat source p. For the derivation of an differential equation, we are going to apply the Gauß divergence theorem. As we distinguish the quantities in the different domains, we apply now the Gauß divergence theorem in each of the subdomains $\Omega_{d,1}$ and $\Omega_{d,2}$. The boundary integral in the conservation of energy (B.1) can be written as integrals over the boundary of the subdomains $\Omega_{d,1}$ and $\Omega_{d,2}$ and an integral over the common interface, which is part of both boundaries. So the boundary integral can be

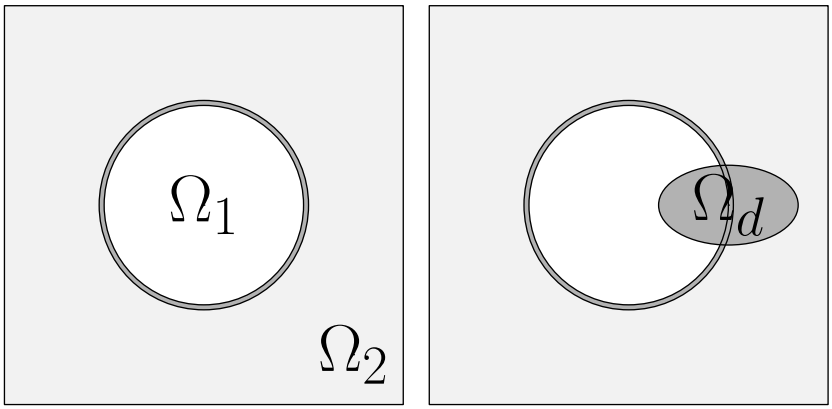


(a) Concrete wall with cooling pipes.

(b) Geometry of the concrete around a pipe with diameter r_1 and thickness δ of the wall of the pipe.

(c) Simplified geometry, if the thickness of the wall of the pipes goes to 0, i.e. $\delta \to 0$.

Figure B.1.: Cooling pipes in a concrete wall.



(a) Domain in the pipe is denoted Ω_1 and the concrete domain is Ω_2

(b) Geometry with an arbitrary test domain $\Omega_d \subset \Omega_1 \cup \Omega_2$.

Figure B.2.: Testdomain for pipe in concrete.

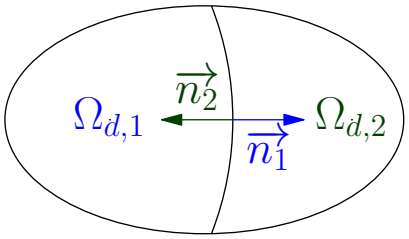


Figure B.3.: Outer normals of the subdomains $\Omega_{d,1}$ and $\Omega_{d,2}$ on the interface.

expressed as

$$\begin{aligned}
-\int_{\partial\Omega_d} j\cdot\vec{n}\,\mathrm{d}\,s &= -\int_{\partial\Omega_d} j\cdot\vec{n}\,\mathrm{d}\,s - \int_{\partial\Omega_{d,1}\setminus\partial\Omega_d} j_1\cdot\vec{n}_1\,\mathrm{d}\,s + \int_{\partial\Omega_{d,1}\setminus\partial\Omega_d} j_1\cdot\vec{n}_1\,\mathrm{d}\,s \\
&\quad - \int_{\partial\Omega_{d,2}\setminus\partial\Omega_d} j_2\cdot\vec{n}_2\,\mathrm{d}\,s + \int_{\partial\Omega_{d,2}\setminus\partial\Omega_d} j_2\cdot\vec{n}_2\,\mathrm{d}\,s \\
&= -\int_{\partial\Omega_{d,1}} j_1\cdot\vec{n}_1\,\mathrm{d}\,s - \int_{\partial\Omega_{d,2}} j_2\cdot\vec{n}_2\,\mathrm{d}\,s \\
&\quad + \int_{\partial\Omega_{d,1}\setminus\partial\Omega_d} j_1\cdot\vec{n}_1\,\mathrm{d}\,s + \int_{\partial\Omega_{d,2}\setminus\partial\Omega_d} j_2\cdot\vec{n}_2\,\mathrm{d}\,s.
\end{aligned}$$

The outer normals $\vec{n}_1$ and $\vec{n}_2$ of the both subdomains point in opposite directions (see also Figure B.3), so that $\vec{n}_2 = -\vec{n}_1$. With this geometric property and the Gauß divergence theorem the boundary integral is equivalent to

$$-\int_{\partial\Omega_d} j\cdot\vec{n}\,\mathrm{d}\,s = -\int_{\Omega_1}\nabla\cdot j\,\mathrm{d}\,\omega - \int_{\Omega_2}\nabla\cdot j\,\mathrm{d}\,\omega + \int_{\partial\Omega_{d,1}\setminus\partial\Omega_d}(j_1-j_2)\cdot\vec{n}_1\,\mathrm{d}\,s.$$

So the balance of energy can be written as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\,t}\int_{\Omega_{d,1}} e_1\,\mathrm{d}\,\omega + \frac{\mathrm{d}}{\mathrm{d}\,t}\int_{\Omega_{d,2}} e_2\,\mathrm{d}\,\omega &= -\int_{\Omega_{d,1}}\nabla\cdot j_1\,\mathrm{d}\,\omega - \int_{\Omega_{d,2}}\nabla\cdot j_2\,\mathrm{d}\,\omega \\
&\quad + \int_{\partial\Omega_{d,1}}(j_1-j_2)\cdot\vec{n}_1\,\mathrm{d}\,s + \int_{\Omega_d} p\,\mathrm{d}\,\omega.
\end{aligned}$$

The energy densities e_i and the heat fluxes j_i can be connected with the temperature y_i. The energy density is an affine linear function (see e.g. [76, formula (1.11)]), so that

$$e_i = e_{0,i} + \varrho c y_i.$$

for the heat flux in the rigid body we can use Fourier's law

$$j_2 = -\lambda_2\nabla y_2.$$

In the pipe the heat is transported by diffusion and convection, so that

$$j_1 = -\lambda_1\nabla y_1 + \vec{v}e,$$

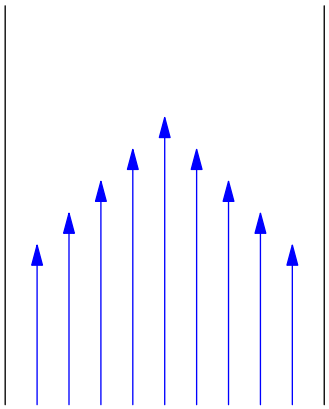


Figure B.4.: Velocity profile of a fluid or gas in a pipe. The streamlines are orthogonal to the outer normal vector.

with the velocity field $\vec{v}$ of the fluid.

As the integral equation does not only hold for a specific test domain Ω_d but for all possible test domain Ω_d, the equation must hold pointwise, so that the equations in the both domains are

$$\varrho c y_2 - \lambda_2 \Delta y_2 = p \qquad \text{in } \Omega_2, \tag{B.2}$$

$$\varrho c y_1 - \lambda_1 \Delta y_1 + \nabla \cdot \vec{v} \varrho c y_1 = p \qquad \text{in } \Omega_1. \tag{B.3}$$

On the boundary we have the identity

$$-\lambda_1 \nabla y_1 \cdot \vec{n}_1 + \vec{v}e \cdot \vec{n}_1 = -\lambda_2 \nabla y_2 \cdot \vec{n}_1 \qquad \text{on } \partial\Omega_1.$$

On the boundary the velocity of the fluid and the outer normal are orthogonal, as the fluid cannot cross the wall of the pipe (see Figure B.4). So the equation on the boundary simplifies to the condition

$$-\lambda_1 \nabla y_1 \cdot \vec{n}_1 = -\lambda_2 \nabla y_2 \cdot \vec{n}_1 \qquad \text{on } \partial\Omega_1. \tag{B.4}$$

The system (B.2), (B.3) and (B.4) can be solved numerically, but the following example shows that the system is still not complete.

Example B.1. *Let* $\Omega = (-1,1)^2$ *with* $\Omega_1 = \left(-\frac{1}{2}, \frac{1}{2}\right)^2$ *and* $\Omega_2 = \Omega \setminus \overline{\Omega_1}$. *Consider the initial value problem*

$$\begin{aligned}
y_{1t} - \Delta y_1 &= 0 && \text{in } (0,T] \times \Omega_1, \\
y_{2t} - \Delta y_2 &= 0 && \text{in } (0,T] \times \Omega_2, \\
\frac{\partial y_1}{\partial n} &= \frac{\partial y_2}{\partial n} && \text{on } (0,T] \times \partial\Omega_1 = (0,T] \times \{\partial\Omega_1 \cap \partial\Omega_2\}, \\
\frac{\partial y_2}{\partial n} &= 0 && \text{on } (0,T] \times \{\partial\Omega_2 \setminus \partial\Omega_1\}, \\
y_1(\cdot,0) &= 1 && \text{in } (0,T] \times \Omega_1, \\
y_2(\cdot,0) &= 0 && \text{in } (0,T] \times \Omega_2.
\end{aligned}$$

It has the solution $y_1 \equiv 1$ and $y_2 \equiv 0$. But this solution is physical not reasonable, as there is no diffusion across the interface due to the fact that $\frac{\partial y_1}{\partial n} = \frac{\partial y_2}{\partial n} = 0$, which represent perfect insulation of Ω_1 and Ω_2.

Remark B.2. *For elliptic boundary value problems with jumps in the coefficients one has two conditions on the interface*

$$\begin{aligned} y_1 &= y_2 && \text{on } \partial\Omega_1, \\ \lambda_1 \nabla y_1 \cdot \vec{n}_1 &= \lambda_2 \nabla y_2 \cdot \vec{n}_1 && \text{on } \partial\Omega_1. \end{aligned}$$

The first Dirichlet-like condition expresses the thermodynamical equilibrium.

If we look at our model, the first question is, whether the modeled system is in a thermodynamic equilibrium. There is no reason that the temperature of the water at the wall of the pipe and the temperature of the concrete at the other side of the pipe have the same temperature. It is reasonable that the temperature of the water in and the temperature of the inner wall of the pipe coincide just as the temperature of the concrete and the temperature of the outer wall of the pipe. But as we have neglected the diameter of the wall of the pipe we need an additional condition for the description of the coupling of the two temperatures.

For this condition we use Newton's law, that the heat flux is proportional to the temperature difference

$$j \cdot n = \sigma_p \left(y_i - y_a \right).$$

Together with Fourier's law on the rigid part and the consideration of the convection in the pipes this yields

$$\begin{aligned} \lambda_2 \nabla y_2 \cdot n_1 &= \sigma_p \left(y_2 - y_1 \right), \\ -\lambda_1 \nabla y_1 \cdot n_1 &= \sigma_p \left(y_1 - y_2 \right). \end{aligned}$$

In [44] these conditions are replaced by

$$y_1 = \kappa y_2 \qquad \text{on } (0,T] \times \partial\Omega_1.$$

where the coefficient κ describes the losses over the interface.

With this additional boundary condition the system is described by

$$\left.\begin{aligned} \varrho c y_2 - \lambda_2 \Delta y_2 &= p && \text{in } (0,T] \times \Omega_2, \\ \varrho c y_1 - \lambda_1 \Delta y_1 + \nabla \cdot \vec{v} \varrho c y_1 &= p && \text{in } (0,T] \times \Omega_1. \\ y_1 &= \kappa y_2 && \text{on } (0,T] \times \partial\Omega_1. \\ -\lambda_1 \nabla y_1 \cdot \vec{n}_1 &= -\lambda_2 \nabla y_2 \cdot \vec{n}_1 && \text{on } (0,T] \times \partial\Omega_1, \end{aligned}\right\} \tag{B.5}$$

where some additional Dirichlet, Neumann or Robin boundary conditions must be posed on the boundary $(0,T] \times \{\partial\Omega_2 \setminus \partial\Omega_1\}$.

C. The finite element method for elliptic partial differential equations

As example for the theory, which we repeated in Section 4.1.1, we consider a second order elliptic differential equation with constant coefficients and the variational formulation

$$a(y,\varphi) = \langle f,\varphi\rangle_{L^2(\Omega)\times L^2(\Omega)} \qquad \forall\varphi\in V,\ \text{where } H_0^1(\Omega)\subseteq V\subseteq H^1(\Omega)$$

with a V-elliptic, continuous bilinear form $a(\cdot,\cdot)$. Further we assume that the domain Ω is nonempty, convex and polygonal one, two or three dimensional domain.

Assumption C.1 (Assumptions on the spatial discretization)**.** *For the discretization we introduce a family of triangulations $\mathcal{T}_h$ of this domain into intervals, triangles or simplices (depending on the dimension) with the following properties:*

- *The triangulation covers the domain:* $\overline{\Omega} = \bigcup_{\theta\in\mathcal{T}_h} \theta$.
- *The intersection of two different elements $\theta_1\in\mathcal{T}_h$ and $\theta_2\in\mathcal{T}_h$ is either empty or a common node or a common edge or a common facet of the triangulation.*
- *For the ratio of the diameter h_k of an element and the radius ρ_k of the largest circle or ball, which is contained in an element, is bounded, i.e.* $\dfrac{h_k}{\rho_k}\le\sigma,\ \forall\theta\in\mathcal{T}_h$.

As discrete space we introduce

$$V_h = \left\{ v\in\mathcal{C}^0(\overline{\Omega})\cap V(\Omega):\ v|_\theta\in\mathbb{P}_1(\theta,\mathbb{R})\ \forall\theta\in\mathcal{T}_h,\ \right\}. \tag{C.1}$$

Let $\{x_1,\dots,x_n\}$ be the set of the nodes of the triangulation $\mathcal{T}_h$, then the nodal or Lagrangian basis of the space V_h is given by

$$\begin{aligned} \varphi_i(x_j) &= \delta_{ij}, && \forall i,j\in\{1,\dots,n\},\\ \varphi \text{ linear on } \theta, && && \forall\theta\in\mathcal{T}_h. \end{aligned}$$

For the approximation we replace f by its Lagrangian interpolation $f_h(x)=\sum_{i=1}^n f(x_i)\varphi_i(x)$. The interpolation error estimate

$$\|v-v_h\|_{L^2(\Omega)} \lesssim h^{k+1}\,|v|_{H^{k+1}(\Omega)}, \qquad k=0 \text{ or } 1 \tag{C.2}$$

is well known for functions $v\in H^2(\Omega)$ ([28, Theorem 3.1.6]).

The finite element approximation y_h is defined as solution of the finite dimensional problem

$$a(y_h,\varphi) = \langle f_h,\varphi\rangle_{L^2(\Omega)\times L^2(\Omega)} \qquad \forall\varphi\in V_h.$$

The error in the $H^1(\Omega)$-norm can be bounded due to the first Strang's Lemma, Theorem 4.5. As the bilinearforms $a(\cdot,\cdot)$ and $a_h(\cdot,\cdot)$ coincide the remaining terms can be bounded with the interpolation error estimate (C.2), and therefore

$$\|y - y_h\|_{H^1(\Omega)} \lesssim h \left(\|y\|_{H^2(\Omega)} + h \|f\|_{L^2(\Omega)} \right).$$

For an error estimate in the $L^2(\Omega)$-norm we can use the Aubin Nitsche Trick in Theorem 4.8 If the stability estimates (4.8)–(4.9) hold for this example and the approximation result (4.10) is fulfilled,the error can be estimated with

$$\|y - y_h\|_{L^2(\Omega)} \lesssim h^2 \left(\|y\|_{H^2(\Omega)} + \|f\|_{L^2(\Omega)} \right).$$

D. Integrals of Finite Element functions

Contents

For the realization of a finite element method we need to evaluate the integration of the product of ansatz functions with test functions. In this chapter we compute these integrals for some one dimensional combinations which are often used in this thesis.

D.1. Finite Element space

We consider functions mapping from an interval $[0,T]$ to $\mathbb{R}$. The generalization to ansatz functions mapping from the interval $[0,T]$ to the Hilbert space V is straightforward.

We recall the space of polynomials up to order k

$$\mathbb{P}_k((0,T)) = \operatorname{span}\left\{t^0,\ t^1, \cdots,\ t^k\right\}.$$

For the discretization we introduce a time grid $0 = t_0 < t_1 < \cdots < t_N = T$ with step size $\tau = t_{i+1} - t_i$. For the discretization in t we use continuous piecewise linear functions

$$\mathcal{Y}_1 = \left\{y \in \mathcal{C}([0,T]) : y|_{(t_i,t_{i+1})} \in \mathbb{P}_1((t_i,t_{i+1}))\forall i \in \{0,\cdots,N-1\}\right\} \tag{D.1}$$

and discontinuous piecewise constant functions

$$\mathcal{P}_0 = \left\{y \in L^2((0,T)) : y|_{(t_i,t_{i+1})} \in \mathbb{P}_0((t_i,t_{i+1}))\forall i \in \{0,\cdots,N-1\}\right\}. \tag{D.2}$$

The Lagrangian basis for the space $\mathcal{Y}_1$ is given as

$$\begin{aligned} &\varphi_i(t_j) = \delta_{ij}, && \forall i,j \in \{0,1,2,\ldots,N\},\\ &\quad \varphi_i \text{ linear in } [t_j,t_{j+1}] && \forall i,j \in \{0,1,2,\ldots,N-1\}. \end{aligned}$$

For the computation of integrals we need a representation of these basis functions, we use the

form

$$\varphi_0(t) = \begin{cases} \dfrac{t_1 - t}{t_1 - t_0} & \text{if } t \in [t_0; t_1], \\ 0 & \text{if } t \notin [t_0; t_1], \end{cases}$$

$$\varphi_i(t) = \begin{cases} \dfrac{t - t_{i-1}}{t_i - t_{i-1}} & \text{if } t \in [t_{i-1}; t_i), \\ \dfrac{t_{i+1} - t}{t_{i+1} - t_i} & \text{if } t \in [t_i; t_{1+1}], \\ 0 & \text{if } t \notin [t_{i-1}; t_{1+1}], \end{cases}$$

$$\varphi_N(t) = \begin{cases} \dfrac{t - t_{N-1}}{t_N - t_{N-1}} & \text{if } t \in [t_{N-1}, t_N], \\ 0 & \text{if } t \notin [t_{N-1}, t_N]. \end{cases}$$

Further we need to evaluate $y \in \mathcal{Y}_1$ for computing the integrals. The function y in the time intervals $[t_{i-1}, t_i]$ and $[t_i, t_{i+1}]$ has the representation

$$y = \begin{cases} y_{i-1} + \dfrac{t - t_{i-1}}{t_i - t_{i-1}}(y_i - y_{i-1}) & \text{for } t \in [t_{i-1}; t_i), \\ y_i + \dfrac{t - t_i}{t_{i+1} - t_i}(y_{i+1} - y_i) & \text{for } t \in [t_i; t_{i+1}). \end{cases}$$

Finally a basis for the space $\mathcal{P}_0$ is

$$\phi_{i+\frac{1}{2}} = \begin{cases} 1 & \text{if } t \in (t_i, t_{i+1}], \\ 0 & \text{if } t \notin (t_i, t_{i+1}]. \end{cases}$$

Now we have introduced everything we need for the computation of the integrals.

D.2. Integrals

D.2.1. Piecewise linear test and ansatz space

For the integrals which involve a piecewise linear test and ansatz space we compute for the product of the functions

$$\begin{aligned} \int_0^T y\varphi_i \,\mathrm{d}\, t &= \int_{t_{i-1}}^{t_i} \left(y_{i-1} + \frac{t - t_{i-1}}{t_i - t_{i-1}}(y_i - y_{i-1}) \right) \cdot \frac{t - t_{i-1}}{t_i - t_{i-1}} \,\mathrm{d}\, t + \\ &\quad + \int_{t_i}^{t_{i+1}} \left(y_i + \frac{t - t_i}{t_{i+1} - t_i}(y_{i+1} - y_i) \right) \cdot \frac{t_{i+1} - t}{t_{i+1} - t_i} \,\mathrm{d}\, t = \\ &= \frac{\tau}{6} y_{i-1} + \frac{4}{6}\tau y_i + \frac{\tau}{6} y_{i+1}, \\ \int_0^T y\varphi_0 \,\mathrm{d}\, t &= \int_0^{t_1} \left(y_0 + \frac{t}{\tau}(y_1 - y_0) \right) \cdot \frac{t_1 - t}{\tau} \,\mathrm{d}\, t = \frac{\tau}{6} y_0 + \frac{2}{6} y_1, \\ \int_0^T y\varphi_N \,\mathrm{d}\, t &= \int_{t_{N-1}}^{t_N} \left(y_{N-1} + \frac{t - t_{N-1}}{\tau}(y_N - y_{N-1}) \right) \cdot \frac{t - t_{N-1}}{\tau} \,\mathrm{d}\, t = \\ &= \frac{\tau}{6} y_{N-1} + \frac{2}{6}\tau y_N, \end{aligned}$$

and for the product of the derivatives

$$\begin{aligned}
\int_0^T y_t \varphi_{i,t} \,\mathrm{d}\, t &= \int_{t_{i-1}}^{t_i} \frac{y_i - y_{i-1}}{t_i - t_{i-1}} \cdot \frac{1}{t_i - t_{i-1}} \,\mathrm{d}\, t + \int_{t_i}^{t_{i+1}} \frac{y_{i+1} - y_i}{t_{i+1} - t_i} \cdot \frac{-1}{t_{i+1} - t_i} \,\mathrm{d}\, t = \\
&= \frac{-y_{i-1} + 2y_i - y_{i+1}}{\tau^2} \tau, \\
\int_0^T y_t \varphi_{0,t} \,\mathrm{d}\, t &= \int_0^{t_1} -\frac{y_1 - y_0}{(t_N - t_{N-1})^2} \,\mathrm{d}\, t = \frac{y_0 - y_1}{\tau^2} \tau, \\
\int_0^T y_t \varphi_{N,t} \,\mathrm{d}\, t &= \int_{t_{N-1}}^{t_N} \frac{y_N - y_{N-1}}{(t_N - t_{N-1})^2} \,\mathrm{d}\, t = \frac{y_N - y_{N-1}}{\tau^2} \tau.
\end{aligned}$$

D.2.2. Piecewise linear ansatz and piecewise constant test space

For the integral of the product of a piecewise linear function (or its derivative) with a piecewise constant function we have

$$\begin{aligned}
\int_0^T y \phi_i \,\mathrm{d}\, t &= \int_{t_{i-1}}^{t_i} y_{i-1} + \frac{t - t_{i-1}}{t_i - t_{i-1}} (y_i - y_{i-1}) \,\mathrm{d}\, t = \frac{\tau}{2} y_{i-1} + \frac{\tau}{2} y_i, \\
\int_0^T y_t \phi_i \,\mathrm{d}\, t &= \int_{t_{i-1}}^{t_i} \frac{1}{t_i - t_{i-1}} (y_i - y_{i-1}) \,\mathrm{d}\, t = y_i - y_{i-1}.
\end{aligned}$$

D.2.3. Piecewise constant test and ansatz space

Finally the integral of two piecewise constant functions is

$$\int_0^T p \phi_i \,\mathrm{d}\, t = \int_{t_{i-1}}^{t_i} p_{i-\frac{1}{2}} \,\mathrm{d}\, t = \tau p_{i-\frac{1}{2}}.$$

E. Software

Contents

E.1. Basic concept

The software for the numerical examples of this thesis is developed in Matlab. The pde-toolbox of Matlab is not used but a finite element implementation that uses ideas of the Matlab finite element codes by Alberty, Carstensen and Funken [1] and of the adaptive finite element implementation in Matlab by Chen and Zhang [25]. For the fast assembly of the matrices ideas presented by Davis [31] and Funken, Praetorius and Wissgott [41] are used.

The code does not use the native object model of Matlab for object orientation, but uses nevertheless ideas of object oriented software development. All the informations about a finite element mesh with its element nodes, elements, matrices and solutions are stored in a single structure, which we will call mesh structure. A typical function call during a finite element computation with the code has typically the form

```
mesh = do_something (mesh, some_other_arguments)
```

At this point the user interface does not differ dramatically from the user interface of object oriented Matlab programs as Matlab passes all arguments by value and not by reference, even objects in object orientated Matlab. One may wonder if it is efficient to pass large structures as arguments, but this is no problem as Matlab performs the copy action not at the moment of the function call but only if the function performs a write access to the object. Further for structures not the whole structure is copied but only the field, which is changed (see [118]).

Inspired by many Matlab-based Finite Element implementations the mesh structure has (at least) the following components.

`mesh.nodes`: A list with the coordinates of all nodes.

`mesh.elem`: In every line of this matrix there are the number of the nodes of an element.

`mesh.solu`: In this component the solution of the differential equation will be saved.

`mesh.solt`: For parabolic equations in this vector the time discretization points are given.

`mesh.type`: For compatibility with other Matlab finite element codes. If the corresponding node is the corner node of an element it contains the entry 1.

`mesh.Dirichlet`, `mesh.Neumann`, `mesh.Robin`: In these components the boundary meshes are saved. The structure is the same as in `mesh.elem`.

There are functions for the construction of simple one and two dimensional geometries and uniform refinements. Furthermore, there is a function for the import of meshes generated by the open source program `gmsh`[45]. With an additional C++-program by Gebhardt [42] it is possible to convert meshes produced by the `TetGen` and `NETGEN` to the `gmsh` format.

On top of this general concept there are implementations for the different partial differential equations.

E.2. Second order equations and optimal control for parabolic equations

For second order elliptic boundary value problems or parabolic initial boundary value problems the matrices are saved after assembling in the additional components

`mesh.A`, `mesh.K`, `mesh.M`, `mesh.R`,

of the structure. The mass matrix is saved in `mesh.M`, the matrix for the Robin boundary conditions in, `mesh.R` and the stiffness matrix in `mesh.A` and `mesh.K`. There are two copies for the stiffness matrix, so that one copy can be unchanged and one copy can be modified to implement Dirichlet boundary conditions by a penalty approach.

There are assembly routines for linear Lagrangian finite elements in one, two and three spatial dimensions. For the one dimensional case there also are quadratic and cubic Lagrangian elements available. The two and three dimensional functions use exact precomputed values for the integrals and the one dimensional uses numerical integration with the Gauß-Kronrod quadrature provided by the Matlab routine `quadgk`.

On top of this simple finite element code an extension for the simulation of optimal control problems exists.

The routines, which implement the Crank-Nicolson discretizations (OC CN1) (OC CN2) (OC G1) for parabolic optimal control problems, assume only that a structure with (at least) the components

`mesh.A`, `mesh.M`

is given and evaluations of the desired states are provided. This mesh structure can be provided by the finite element implementation described above, but every other structure or discretization with these components is also accepted. So the spatial discretization can be easily modified.

The linear system of the discretization, given by (6.11) (or the linear systems corresponding to the other discretizations), are assembled based on the existing matrices and solved with the standard linear solver of Matlab. Of course it is possible to replace the linear solver by another algorithm.

E.3. Hydration of concrete

For the simulation of the hydration of concrete the structure is enriched with the additional fields which contain material and model parameter. Further the fields

`mesh.maturity`, `mesh.hydration`

contain references to functions, which describe the maturity and the heat development in use. So the model in use can be easily adopted by changing these function references.

After the computation of the temperatures with the solution methods for initial value problems provided by Matlab, the stresses can be computed and are also stored in additional fields of the mesh structure.

There are functions for the visualization of the solution. For two dimensional domains the profiles with computed solution can be plotted along the time axis. Further the plot of a value at given point can be plotted over time.

For a more satisfactory visualization the solution can be exported in the vtk file format. The exported files can be visualized with vtk-viewers such as ParaView [122].

E.4. Fourth order elliptic equations and $H^{(2,1)}(Q)$-elliptic equations

In preparation of the Lagrange-Hermite tensor product finite elements for the $H^{(2,1)}$-elliptic equations, there exists an implementation of one dimensional cubic Hermite elements. In this implementation the structure possesses the additional field

`mesh.dof`

which doubles the nodes, as in every node two degree of freedom are allocated, one for the solution and one for the derivative of the solution, to provide a global continuous differentiable solution. Note that therefore for Hermite element also Neumann boundary conditions need a modification of the linear system as the derivative at the nodes is also a degree of freedom.

For the discretization of $H^{(2,1)}(Q)$-elliptic equations the `mesh` structure is once more enriched with an additional component. The field

`mesh.bnd`

contains a structure for the management of the boundary meshes and boundary conditions.

After assembly, where again numerical quadrature with Gauß-Kronrod quadrature provided by the Matlab routine `quadgk` is used, the linear system can be solved with the linear solver of Matlab.

Bibliography

[1] Jochen Alberty, Carsten Carstensen, and Stefan A. Funken. Remarks around 50 lines of Matlab: short finite element implementation. *Numerical Algorithms*, 20:117–137, 1999.

[2] Thomas Apel. Interpolation of non-smooth functions on anisotropic finite element meshes. *Mathematical Modelling and Numerical Analysis*, 33(6):1149–1185, 1999.

[3] Thomas Apel and Thomas G. Flaig. Simulation and mathematical optimization of the hydration of concrete for avoiding thermal cracks. In Klaus Gürlebeck and Carsten Könke, editors, *18th International Conference on the Application of Computer Science and Mathematics in Architecture and Civil Engineering*, Weimar, `http://euklid.bauing.uniweimar.de/ikm2009/paperDetails.php?ID=108`, 2009.

[4] Thomas Apel and Thomas G. Flaig. Crank-Nicolson schemes for optimal control problems with evolution equations. Preprint-Number SPP1253-113, DFG Priority Program 1253, Erlangen, `http://www.am.uni-erlangen.de/home/spp1253/wiki/images/b/b7/Preprint-SPP1253-113.pdf`, 2010.

[5] Thomas Apel and Thomas G. Flaig. Crank-Nicolson schemes for optimal control problems with evolution equations. *SIAM Journal on Numerical Analysis*, 50(3):1484–1512, 2012.

[6] Thomas Apel, Thomas G. Flaig, and Serge Nicaise. A priori error estimates for finite element methods for $H^{(2,1)}$-elliptic equations. Preprint-Number SPP1253-139, DFG Priority Program 1253, Erlangen, `http://www.am.uni-erlangen.de/home/spp1253/wiki/images/2/2d/Preprint-SPP1253-139.pdf`, 2012.

[7] Ralph A. Artino. On semielliptic boundary value problems. *Journal of Mathematical Analysis and Applications*, 42:610–626, 1973.

[8] Ralph A. Artino and J. Barros-Neto. Regular semielliptic boundary value problems. *Journal of Mathematical Analysis and Applications*, 61:40–57, 1977.

[9] J. R. Barber. *Elasticity*. Kluwer Academic Publishers, Dordrecht, 2002.

[10] Klaus-Jürgen Bathe. *Finite-Elemente-Methoden. Matrizen und lineare Algebra, die Methode der finiten Elemente, Lösung von Gleichgewichtsbedingungen und Bewegungsgleichungen*. Springer, Berlin, 1986.

[11] Roland Becker, Dominik Meidner, and Boris Vexler. Efficient Numerical Solution of Parabolic Optimization Problems by Finite Element Methods. *Optimization Methods and Software*, 22(5):813 – 833, 2007.

[12] Olaf Benedix. *Adaptive Numerical Solution of State Constrained Optimal Control Problems*. PhD thesis, TU München, `http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20110711-1078755-1-3`, 2011.

[13] M. H. Berger. Finite element analysis of flow in a gas-filled rotating annulus. *International Journal for Numerical Methods in Fluids*, 7:215–231, 1987.

[14] Martin Bernauer. *Motion Planning for the Two-Phase Stefan Problem in Level Set Formulationi*. PhD thesis, TU Chemnitz, `http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-63654`, 2010.

[15] Oleg Vladimirovich Besov, Valentin Petrovich Il'in, and Sergeĭ Michaĭlovič Nikol'skiĭ. *Integral representations of functions and imbedding Theorems - Volume I*. V. H. Winston & Sons, Washington, D.C.; Halsted Press [John Wiley & Sons], New York, 1978.

[16] J. Frédéric Bonnans and Julien Laurent-Varin. Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control. Rapport de recherche RR–5398, INRIA, `http://hal.inria.fr/docs/00/07/06/05/PDF/RR-5398.pdf`, 2004.

[17] J. Frédéric Bonnans and Julien Laurent-Varin. Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control - Order conditions for symplectic partitioned Runge-Kutta schemes (second revision). *Numerische Mathematik*, 103:1–10, 2006.

[18] Alfio Borzì. Multigrid methods for parabolic distributed optimal control problems. *Journal of Computational and Applied Mathematics*, 157(2):365–382, 2003.

[19] Alfio Borzì and Volker Schulz. Multigrid methods for PDE optimization. *SIAM Review*, 51(2):361–395, 2009.

[20] Alfio Borzì and Volker Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. SIAM, Philadelphia, 2012.

[21] Timm Braasch. *Herabsetzung des Risikos einer Rissbildung abschnittsweiser hergestellter Brückenbauten aus Beton*. PhD thesis, Universität Duisburg-Essen, `http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-12257/Diss%20Braasch.pdf`, 2003.

[22] Dietrich Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer, Berlin, second edition, 1997.

[23] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Springer, Berlin, third edition, 2008.

[24] Guido Büttner. *Ein Mehrgitterverfahren zur optimalen Steuerung parabolischer Probleme*. PhD thesis, TU Berlin, `http://opus.kobv.de/tuberlin/volltexte/2004/878/pdf/buettner_guido.pdf`, 2004.

[25] Long Chen and Chen-Song Zhang. AFEM@MATLAB: a MATLAB package of adaptive finite element methods. Technical report, University of Maryland, `http://www.math.umd.edu/~zhangcs/paper/AFEM%40matlab.pdf`, 2006.

[26] Chen Chuanmiao and Shih Tsimin. *Finite Element methods for Integrodifferential equations*, volume 9 of *Series on Applied Mathematics*. World Scientific, Singapore, 1998.

[27] Monique Chyba, Ernst Hairer, and Gilles Vilmart. The role of symplectic integrators in optimal control. *Optimal control applications and methods*, 30(4):367–382, 2009.

[28] Philippe G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Company, Amsterdam, 1979.

[29] Debora Clever and Jens Lang. Optimal control of radiative heat transfer in glass cooling with restrictions on the temperature gradient. *Optimal Control Applications and Methods*, 33:157–175, 2012.

[30] Robert Dautray and Jacques-Louis Lions. *Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution problems I*. Springer, Berlin, 1992.

[31] Tim Davis. Creating sparse finite-element matrices in MATLAB. Guest blog in: Loren Shure: Loren on the Art of MATLAB, `http://blogs.mathworks.com/loren/2007/03/01/creating-sparse-finite-element-matrices-in-matlab`, 2007.

[32] Rene de Vogelaere. Methods of integration which preserve the contact transformation property of the Hamilton equations. Technical Report 4, Department of Mathematics, University of Notre Dame, Notre Dame, Indiana, 1956.

[33] Klaus Deckelnick and Michael Hinze. Variational discretization of parabolic control problems in the presence of pointwise state constraints. Preprint-Number SPP1253-08-08, DFG Priority Program 1253, Erlangen, `http://www.am.uni-erlangen.de/home/spp1253/wiki/images/2/25/Preprint-spp1253-08-08.pdf`, 2009.

[34] Ewald Rudolf Dirnberger. *Zur thermischen Zwangsbeanspruchung von rückverankerten und unverankerten Unterwasserbeton-Sohlen*. PhD thesis, Universität der Bundeswehr München, `http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:706-2053`, 2009.

[35] Asen L. Dontchev, William W. Hager, and Vladimir M. Veliov. Second-order Runge-Kutta approximations in control constrained optimal control. *SIAM Journal on Numerical Analysis*, 38(1):202–226, 2000.

[36] Todd Dupont and Ridgway Scott. Polynomial approximation of functions in Sobolev spaces. *Mathematics of Computation*, 34(150):441–463, 1980.

[37] J. F. Eastham and J. S. Peterson. The finite element method in anisotropic Sobolev spaces. *Computers and Mathematics with Applications*, 47:1775–1786, 2004.

[38] Benno Eierle. *Berechnungsmodelle für rißgefährdete Betonbauteile unter frühem Temperaturzwang*. PhD thesis, TU München, 2000.

[39] Alexandre Ern and Jean-Luc Guermond. *Theory and Practice of Finite Elements*. Springer, Berlin, 2004.

[40] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. AMS, Providence, Rhode Island, 1998.

[41] Stefan Funken, Dirk Praetorius, and Philipp Wissgott. Efficient implementation of adaptive P1-FEM in Matlab. *Computational Methods in Applied Mathematics*, 11(4):460–490, 2011.

[42] Eric Gebhardt. Implementierung eines Netzmanagers für 3D Finite-Element-Netze. Bachelor's thesis, Universität der Bundeswehr München, Neubiberg, 2011.

[43] Bernhard Gehrmann. Berechnung von Temperaturspannungen während der Hydratation von Beton. Bachelor's thesis, Universität der Bundeswehr München, Neubiberg, 2009.

[44] Bernhard Gehrmann. Berechnungen zur Energieeffizienzbewertung von Gebäuden. Master's thesis, Universität der Bundeswehr München, Neubiberg, 2011.

[45] Christophe Geuzaine and Jean-François Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[46] Wei Gong, Michael Hinze, and Zhaojie Zhou. Space-time finite element approximation of parabolic optimal control problems. Preprint-Number SPP1253-126, DFG Priority Program 1253, Erlangen, `http://www.am.uni-erlangen.de/home/spp1253/wiki/images/2/2a/Preprint-SPP1253-126.pdf`, 2011.

[47] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman, Boston, 1985.

[48] Christian Großmann and Hans-Görg Roos. *Numerische Behandllung partieller Differentialgleichungen*. Teubner, Wiesbaden, 2005.

[49] Max D. Gunzburger and Houston G. Wood III. A finite element method for the Onsager pancake equation. *Computer Methods in Applied Mechanics and Engineering*, 31:43–59, 1982.

[50] Max D. Gunzburger and Angela Kunoth. Space-time adaptive wavelet method for optimal control problems constrained by parabolic evolution equations. *SIAM J. Control Optim.*, 49(3):1150–1170, 2011.

[51] Alex-Walter Gutsch. *Stoffeigenschaften jungen Betons – Versuche und Modelle*. Deutscher Ausschuss für Stahlbeton Heft 495. Beuth Verlag GmbH, Berlin, 1999.

[52] Wolfgang Hackbusch. A numerical method for solving parabolic equations with opposite orientations. *Computing*, 20:229–240, 1978.

[53] Wolfgang Hackbusch. *Elliptic Differential Equations - Theory and Numerical treatment*. Springer, Berlin, 1992.

[54] William W. Hager. Rates of convergence for discrete approximations to unconstrained control problems. *SIAM Journal on Numerical Analysis*, 13(4):449–472, 1976.

[55] William W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87:247–282, 2000.

[56] Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the Störmer-Verlet method. *Acta Numerica*, 12:399–450, 2003.

[57] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag, Berlin, second edition, 2006.

[58] Ernst Hairer, Syvert Paul Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer-Verlag, Berlin, second revised edition, 1993.

[59] G. N. Hile, C. P. Mawata, and Chiping Zhou. A priori bounds for semielliptic operators. *Journal of Differential Equations*, 176:29–64, 2001.

[60] Michael Hinze, Michael Köster, and Stefan Turek. A hierarchical space-time solver for distributed control of the Stokes equation. Preprint-Number SPP1253-16-01, DFG Priority Program 1253, Erlangen, `http://www.am.uni-erlangen.de/home/spp1253/wiki/images/6/63/Preprint-spp1253-16-01.pdf`, 2008.

[61] Michael Hinze, Michael Köster, and Stefan Turek. A space-time multigrid solver for distributed control of the time-dependent Navier-Stokes equation. Preprint-Number SPP1253-16-02, DFG Priority Program 1253, Erlangen, `http://www.am.uni-erlangen.de/home/spp1253/wiki/images/1/1a/Preprint-spp1253-16-02.pdf`, 2008.

[62] Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE Constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, Berlin, 2009.

[63] Karl-Heinz Hoffman and Lishang Jiang. Optimal control of phase field model for solidification. *Numerical functional analysis and optimization*, 13(1&2):11–27, 1992.

[64] L. Steven Hou, Oleg Imanuvilov, and Hee-Dae Kwon. Eigen series solutions to terminal state tracking optimal control problems and exact controllability problems constrained by linear parabolic PDEs. *Journal of Mathematical Analysis and Applications*, 313:284–310, 2006.

[65] Jens Huckfeldt. *Thermomechanik hydratisierenden Betons – Theorie, Numerik und Anwendung*. PhD thesis, TU Carolo – Wilhelmina, Braunschweig, 1993.

[66] Claes Johnson. *Numerical solution of partial differential equations by the finite element methode*. Cambridge University press, Cambridge, 1987.

[67] Jan-Erik Jonasson. *Modelling of Temperature, Moisture and Stresses in Young Concrete*. PhD thesis, Luelå University of Technology, 1994.

[68] Mark Kachanov, Boris Shafiro, and Igor Tsukrov. *Handbook of Elasticity Solutions*. Kluwer Academics Publishers, Dordrecht, 2003.

[69] Bernd Kalkowski. *Zur mathematischen Modellierung und optimalen Steuerung des Hydratationsprozesses in dünnwandigen Betonelementen*. PhD thesis, Ingenieurhochschule Cottbus, 1986.

[70] Michael Köster. *A Hierarchical Flow Solver for Optimisation with PDE Constraints*. PhD thesis, TU Dortmund, Lehrstuhl III für Angewandte Mathematik und Numerik, 2011. Slightly corrected version with an additional appendix concerning prolongation/restriction, `http://www.mathematik.tu-dortmund.de/lsiii/cms/papers/Koester2011a.pdf`.

[71] Matias Krauß. *Probabilistischer Nachweis der Wirksamkeit von Maßnahmen gegen frühe Trennrisse in massigen Betonbauteilen*. PhD thesis, Technische Universität Carolo-Wilhelmina zu Braunschweig, 2004.

[72] Erwin Kreyzig. *Introduction to functional analysis with applications*. John Willey & Sons, New York, 1989.

[73] Ernst Kunz. *Einführung in die algebraische Geometrie*. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig, 1997.

[74] O. A. Ladyzhenskaya, V. A. Solonnikov, and N.N. Ural'ceva. *Linear and quasilinear equations of parabolic type*, volume 23 of *Translations of mathematical monographs*. AMS, Providence, Rhode Island, 1968.

[75] Jens Lang. Adaptive computation for boundary control of radiative heat transfer in glass. *Jornal of Computational and Applied Mathematics*, 183:312–326, 2005.

[76] Stig Larsson and Vidar Thomée. *Partielle Differentialgleichungen und numerische Methoden*. Springer, Berlin, 2005.

[77] Hans-Gerd Leopold. On function spaces of variable order of differentiation. *Forum Mathematicum*, 3:1–21, 1991.

[78] Jacques Louis Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Springer, Berlin, 1971.

[79] Jacques Louis Lions and Enrico Magenes. *Non-Homogeneous Boundary Value Problems and Applications II*. Springer, Berlin, 1972.

[80] Dominik Meidner and Boris Vexler. Adaptive Space-Time Finite Element Methods for Parabolic Optimization Problems. *SIAM Journal on Control and Optimization*, 46(1):116 – 142, 2007.

[81] Dominik Meidner and Boris Vexler. A Priori Error Estimates for Space-Time Finite Element Discretization of Parabolic Optimal Control Problems. Part I: Problems without Control Constraints. *SIAM Journal on Control and Optimization*, 47(3):1150 – 1177, 2008.

[82] Dominik Meidner and Boris Vexler. A Priori Error Estimates for Space-Time Finite Element Discretization of Parabolic Optimal Control Problems. Part II: Problems with Control Constraints. *SIAM Journal on Control and Optimization*, 47(3):1301–1329, 2008.

[83] Dominik Meidner and Boris Vexler. A priori error analysis of the Petrov Galerkin Crank Nicolson scheme for parabolic optimal control problems. Preprint-Nunmber SPP1253-109, DFG Priority Program 1253, Erlangen `http://www.am.uni-erlangen.de/home/spp1253/wiki/images/b/b8/Preprint-SPP1253-109.pdf`, 2010.

[84] Dominik Meidner and Boris Vexler. A priori error analysis of the Petrov Galerkin Crank Nicolson scheme for parabolic optimal control problems. *SIAM Journal on Control and Optimization*, 49(5):2183 – 2211, 2011.

[85] Solomon G. Michlin. *Partielle Differentialgleichungen in der mathematischen Physik*. Verlag Harri Deutsch, Thun, 1978.

[86] Ira Neitzel, Uwe Prüfert, and Thomas Slawig. Strategies for time-dependent pde control using an integrated modeling and simulation environment. Part one: problems without inequality constraints. Matheon preprint no. 408, `http://nbn-resolving.de/urn:nbn:de:0296-matheon-4206`, 2007.

[87] Ira Neitzel, Uwe Prüfert, and Thomas Slawig. Strategies for time-dependent pde control with inequality constraints using an integrated modeling and simulation environment. *Numerical Algorithms*, 50(3):241–269, 2009.

[88] Ira Neitzel, Uwe Prüfert, and Thomas Slawig. On solving parabolic optimal control problems by using space-time discretization. Technical Report 05-2009, TU Berlin, 2009.

[89] Ira Neitzel, Uwe Prüfert, and Thomas Slawig. A smooth regularization of the projection formula for constrained parabolic optimal control problems. *Numerical Functional Analysis and Optimization*, 32(12), 2011.

[90] Lutz Nietner. Thermisch bedingte Risse. In Harald S. Müller, Ulrich Nolting, and Michael Haist, editors, *Beherrschung von Rissen in Beton: 7. Symposium Baustoffe und Bauwerkserhaltung, Karlsruher Institut für Technologie; Karlsruhe, 23. März 2010.*, Karlsruhe, 2010.

[91] Lutz Nietner and Detlef Schmidt. Temperatur- und Feuchtigkeitsmodellierung durch Praxiswerkzeuge – Grundlagen dauerhafter Betonteile. *Beton- und Stahlbetonbau*, 98(12):738–746, 2003.

[92] Sergeĭ Michaĭlovič Nikol'skiĭ. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, Berlin, 1975.

[93] N.V. Oganesyan. Solution of the first boundary value problem for model semielliptic equation by projection-grid method. *Soviet Journal of Contemporary Mathematical Analysis. Armenian Academy of Sciences.*, 4:82–90, 1989. translation from Izv. Akad. Nauk Arm. SSR. Matematika, Vol. 24, No.4, 393-402 (1989).

[94] Peter Onken and Ferdinand S. Rostásy. *Wirksame Betonzugfestigkeit im Bauwerk bei früh einsetzendem Temperaturzwang*. Beuth Verlag, Berlin, 1995.

[95] Ole Østerby. Five ways of reducing the Crank-Nicolson oscillations. *BIT Numerical Mathematics*, 43:811–822, 2003.

[96] Erik Steen Pedersen, Helle Spange, Erik Jørgen Pedersen, Henrik Elgaard Jensen, Mette Elbæk Andersen, Per Fogh Jensen, and Jan Graabek Knudsen. HETEK – control of early age in concrete – guidelines. Technical Report No. 120, Road Directorate, Denmark Ministry of Transport, `http://www.hetek.teknologisk.dk/_root/media/17069%5FHetek%2C%20Report%20No%20120%2C1997.pdf`, 1997.

[97] René Pinnau and Guido Thömmes. Optimal boundary control of glass cooling processes. *Mathematical methods in the applied sciences*, 27:1261–1281, 2004.

[98] Robert Pree. Temperatursteuerung von Beton in Theorie und Praxis. In *Expertenforum Beton 2005: Tempertaursteuerung von Beton*, pages 20–31. Vereinigung der Österreichischen Zementindustrie, Wien, `http://www.zement.at/service/literatur/detail.asp?wid=253`, 2005.

[99] Waldemar Rachowicz. An anisotropic *h*-type mesh-refinement strategy. *Computer Methods in Applied Mechanics and Engineering*, 109:169–181, 1993.

[100] Rolf Rannacher. Finite element solution of diffusion problems with irregular data. *Numerische Mathematik*, 43:309–327, 1984.

[101] Jean-Pierre Raymond. Optimal control of partial differential equations. Lecture notes of Université Paul Sabatier and French Indian Cyber-University in Science, `http://www.math.univ-toulouse.fr/~raymond/book-ficus.pdf`, 2009.

[102] Stefan Röhling. *Zwangsspannungen infolge Hydratationswärme*. Verlag Bau+Technik, Düsseldorf, 2005.

[103] Arnd Rösch. Error estimates for parabolic optimal control problems with control constraints. *Zeitschrift für Analysis und ihre Anwendungen*, 23(2):353–376, 2004.

[104] Ferdinand S. Rostásy and Matias Krauß. *Frühe Risse in massigen Betonbauteilen – Ingenieurmodelle für die Planung von Gegenmaßnahmen*, volume 520 of *Deutscher Ausschuss für Stahlbeton (DafStb) im DIN, Deutsches Institut für Normung e.V.* Beuth Verlag, Berlin, 2001.

[105] Ferdinand S. Rostásy, Matias Krauß, and Harald Budelmann. Planungswerkzeug zur Kontrolle der frühen Rißbildung in massigen Betonbauteilen. – Teil 1: Einführung, Bezeichnungen und Literatur. *Bautechnik*, 79(7):431–435, 2002.

[106] Ferdinand S. Rostásy, Matias Krauß, and Harald Budelmann. Planungswerkzeug zur Kontrolle der frühen Rißbildung in massigen Betonbauteilen. – Teil 2: Hydratation und Waermefreisetzung. *Bautechnik*, 79(8):431–435, 2002.

[107] Ferdinand S. Rostásy, Matias Krauß, and Harald Budelmann. Planungswerkzeug zur Kontrolle der frühen Rißbildung in massigen Betonbauteilen. – Teil 3: Eigenschaften und Stoffmodelle jungen Betons. *Bautechnik*, 79(9):641–647, 2002.

[108] Ferdinand S. Rostásy, Matias Krauß, and Harald Budelmann. Planungswerkzeug zur Kontrolle der frühen Rißbildung in massigen Betonbauteilen. – Teil 4: Felder der freien Verformungen und mechanische Betoneigenschaften im Bauteil. *Bautechnik*, 79(10):697–703, 2002.

[109] Ferdinand S. Rostásy, Matias Krauß, and Harald Budelmann. Planungswerkzeug zur Kontrolle der frühen Rißbildung in massigen Betonbauteilen. – Teil 5: Behinderung und Zwang. *Bautechnik*, 79(11):778–789, 2002.

[110] Ferdinand S. Rostásy, Matias Krauß, and Harald Budelmann. Planungswerkzeug zur Kontrolle der frühen Rißbildung in massigen Betonbauteilen. – Teil 6: Entscheidung über Rißbildung mit Rißkriterien – Teil 7: Zusammenfassung. *Bautechnik*, 79(12):869–874, 2002.

[111] Ferdinand S. Rostásy, Matias Krauß, and Alex-Walter Gutsch. *Spannungsberechnung und Risskriterien für jungen Beton – Methoden des IBMB*. Number 156 in Schriftenreihe des iBMB. Institut für Baustoffe, Massivbau und Brandschutz, TU Braunschweig, 2001.

[112] Walter Rudin. *Analysis*. Oldenburg Verlag, München, 2., korrigierte Auflage edition, 2002.

[113] Friedhelm Schieweck. *A*-stable discontinuous Galerkin-Petrov time discretization of higher order. *Journal of Numerical Mathematics*, 18(1):25–57, 2010.

[114] Karl Schikora and Benno Eierle. Berechnungsmodelle für Betonbauteile unter frühem Temperaturzwang. In Konstantin Meskouris, editor, *Baustatik – Baupraxis 7*, pages 423–430. A. A. Balkema, Rotterdam, 1999.

[115] Michael Schmich and Boris Vexler. Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. *SIAM Journal on Scientific Computing*, 30(1):369 – 393, 2008.

[116] Klaus Schöppel. *Entwicklung der Zwangsspannungen im Beton während der Hydratation*. PhD thesis, TU München, 1993.

[117] Dominik Schötzau and Christoph Schwab. Time discretisation of parabolic problems by the hp-version of the discontinuous Galerkin finite element method. *SIAM Journal on Numerical Analysis*, 38(3):837–875, 2000.

[118] Loren Shure. Memory management for functions and variables. In: Loren on the Art of MATLAB, `http://blogs.mathworks.com/loren/2006/05/10/memory-management-for-functions-and-variables/`, 2006.

[119] Wladimir Iwanowitsch Smirnow. *Lehrgang der höheren Mathematik, Teil 5*. VEB Deutscher Verlag der Wissenschaften, Berlin, 10. edition, 1988.

[120] Wolfram Sperber and Fredi Tröltzsch. Function spaces and Lagrange multipliers rule for a parabolic control problem arising from the hydratation of concrete. *Disscussiones Mathematicae*, 8:109–120, 1986.

[121] Jürgen Sprekels and Songmu Zheng. Optimal control problems for a thermodynamically consistent model of phase field type for phase transitions. *Schwerpunktprogramm der Deutschen Forschungsgemeinschaft Anwendungsbezogene Optimierung und Steuerung*, Repoert No. 277, 1991.

[122] Amy Henderson Squillacote. *The ParaView Guide*. Kitware Inc., Clifton Park, NY, 2007.

[123] Michael Staffa. Zur Vermeidung von hydratationsbedingten Rissen in Stahlbetonwänden. *Beton- und Stahlbetonbau*, 89(1):4–8, 1994.

[124] Olaf Steinbach. *Numerische Näherungsverfahren für elliptische Randwertprobleme*. Teubner, Stuttgart, 2003.

[125] Vidar Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer, Berlin, second edition, 2006.

[126] Guido Thömmes, René Pinnau, Mohammed Seaid, Thomas Götz., and Axel Klar. Numerical methods and optimal control for glass cooling processes. *Transport theory and statistical physics*, 31(4-6):513–529, 2002.

[127] Hans Triebel. A priori estimates and boundary value problems for semi-elliptic differential equations: A model case. *Communications in partial differential equations*, 8(15):1621–1664, 1983.

[128] Fredi Tröltzsch. *Optimality conditions for parabolic control problems and applications*. Teubner, Leipzig, 1984.

[129] Fredi Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen - Theorie, Verfahren und Anwendungen*. Vieweg, Wiesbaden, 2005.

[130] Karlhans Wesche. Baustoffkennwerte zur Berechnung von Temperaturfeldern in Betonbauteilen. In *Liber Amicorum opgedragen aan F.G. Riessauw ter gelegenheid van zijn zeventigste verjaardag 17 april 1982*. Gent, 1982.

[131] Joseph Wloka. *Partielle Differentialgleichungen - Sobolevräume und Randwertaufgaben*. Teubner, Stuttgart, 1982.