Space-Time Finite Element Methods for Parabolic Initial-Boundary Problems

Master Thesis
to obtain the academic degree of
Diplom-Ingenieur
in the Master’s Program
Industriemathematik
Abstract

In this thesis, we consider the numerical solution of parabolic initial-boundary value problems with variable in space and time, possibly discontinuous coefficients. Such problems typically arise in the simulation of heat conduction problems, diffusion problems, but also for two-dimensional eddy current problems in electromagnetics. Discontinuous coefficients allow the treatment of moving interfaces like the rotation of an electrical motor. We recall two different approaches to prove that the continuous problem is well-posed in different settings (spaces) under quite general (physical) assumptions. In order to solve parabolic problems numerically, a vertical or horizontal method of lines is traditionally applied. However, in this thesis, an alternative approach is chosen. We treat time just as another variable and derive a conforming space-time finite element method. This introduces some challenges, but enables us to apply results from the existing and well investigated theory on elliptic boundary value problems. We show stability of the method, and, additionally, an a priori error estimate is provided. The case of local stabilizations, which are important for adaptivity, is also investigated.

To study the method in practice, we introduced typical model problems in one, two, and three spatial dimensions. The implementation of our space-time finite element method is fully parallelized. The numerical studies were performed on the high performance computing cluster RADON1, and the outcomes verify the theoretical results.
Zusammenfassung


Für die praktische Anwendung definieren wir uns typische Modelprobleme in einer, zwei, und drei Raumdimensionen. Die Implementierung unserer Raum-Zeit Finiten Elemente Methode is komplett parallelisiert. Die numerischen Studien wurden am High-Performance-Computing-Cluster RADON1 durchgeführt und die Resultate bestätigen unsere theoretischen Aussagen.
Acknowledgments

I would like to express special thanks to my girlfriend Sarah for always pushing me forward to give my best, and my parents Stefan and Brigitte for enabling me to follow my studies.

I would like to thank my supervisors Prof Ulrich Langer and Dr Martin Neumüller for introducing me to the topic and giving me the opportunity to write this thesis. Moreover, I would like to thank them for the valuable discussions, for proofreading, and for organizing financial support.

I also thank the Institute of Computational Mathematics for providing me a place to write this thesis.

This work was supported by a DK-fellowship from the Doctoral Program ”Computational Mathematics” (W1214).

Andreas Schafelner
Linz, November 2017
## Contents

1 Introduction ............................................. 1

2 Preliminaries ........................................... 4
   2.1 Basic Notation ........................................ 4
   2.2 Function spaces ....................................... 7
   2.3 Additional Theorems ................................... 10

3 Space-time variational formulations .................. 11
   3.1 Formulation in Sobolev spaces on the space-time cylinder ........ 11
   3.2 Alternative approach in Bochner spaces of abstract functions .... 16

4 Space-time Finite Element Methods .................. 18
   4.1 A Space-time FEM based on time upwind stabilization ............ 18
   4.2 A Space-time FEM based on stabilizing bubble functions ........ 37

5 Numerical Results ....................................... 39
   5.1 1D Examples .......................................... 41
   5.2 2D Examples .......................................... 49
   5.3 3D Examples .......................................... 57

6 Conclusions & Outlook .................................. 59
Chapter 1

Introduction

When we deal with physical problems, for instance, diffusion problems, heat-conduction problems, or simulations of electrical machines, the governing partial differential equations (PDEs) are often of parabolic type. Thus, the development of numerical schemes to solve parabolic equations is of great importance. The standard approach for solving parabolic PDEs is usually some kind of time-stepping method, with semi-discretization in the spatial variables. Another approach would be to first discretize with respect to time and then perform a discretization in the spatial variables. This approach is called Rothe’s method. A more recent and alternative approach consists in a full space-time discretization at once by treating time just as another space variable, i.e., we solve a problem with one dimension more. The basic steps for these methods can be summarized in the following way:

1. Line Variational Formulation and Vertical Method of Lines:
   - multiply the PDE by an appropriate test-function $v(x)$,
   - integrate over the spatial computational domain $\Omega$,
   - use integration by parts on the highest order spatial derivative,
   - discretize first in space by some spatial discretization like finite element method (FEM), and then solve the resulting first-order system of ordinary differential equations in time with an appropriate time-stepping method, e.g., a Runge-Kutta method.

2. Line Variational Formulation and Horizontal Method of Lines (Rothe’s method):
   - multiply by the PDE an appropriate test-function $v(x)$,
   - integrate over the spatial computational domain $\Omega$,
   - use integration by parts on the highest order spatial derivative,
   - discretize first in time by some time-stepping method like the implicit Euler scheme, and then discretize the resulting sequence of elliptic problems by means of an appropriate discretization method like the FEM.
CHAPTER 1. INTRODUCTION

3. Space-time Variational Formulation:

- multiply the PDE by an appropriate test-function \( v(x,t) \),
- integrate over the space-time domain (cylinder) \( Q = \Omega \times (0,T) \),
- use integration by parts, e.g. on the highest order spatial derivative and/or the temporal derivative,
- discretize in space and time simultaneously, e.g., by space-time FEM or Isogeometric Analysis (IgA), and solve the resulting linear system by an efficient solver.

In this master thesis, we will focus on the latter approach. The motivation behind this is that, for elliptic problems, there exist plenty of efficient and, most important, parallel solving methods. If we would be able to derive a stable discrete bilinear form, for which we can prove coercivity (ellipticity) in some mesh-dependent norm in the space-time FE-space, then we can solve the space-time problem fully in parallel. Another reason for the space-time approach is that we are not restricted to a special structure of the mesh. This means that we can apply adaptive mesh refinement both in space and time simultaneously. Last but not least, we can easily deal with moving interfaces and domains, where the coefficients of the PDE and/or the spatial domain \( \Omega_t \) depend on the time as well. Under certain assumptions imposed on the movement, we can transform the time dependent spatial domain to a fixed spatial domain via a change of variables (see [14], Chapter III, §1).

The standard discretization techniques, namely the vertical method of lines and Rothe’s method, and their properties are well investigated, see [31] and [15], respectively. However, their sequential structure complicates the parallel solution of the resulting discretized problem, the development of efficient space-time adaptive methods, as well as the treatment of moving interfaces and spatial domains. The application of a space-time finite element scheme has already a long history, see e.g. [9], [12]. However, the analysis of the equivalent operator equations was done more recently, see, e.g. [26], [18], [33]. Another popular approach are time-parallel multigrid methods [10]. Most of the more recent space-time finite element methods use discontinuous Galerkin methods, at least in time, see, e.g., [19], [20], [21], [30], and the references given therein. But also conforming space-time methods have been developed, e.g., Steinbach introduced a stable Petrov-Galerkin method [28], and Toulopoulos used bubble functions to stabilise a Galerkin method [32]. In the context of using Isogeometric Analysis as space-time discretization method, Langer, Moore and Neumüller [16] proposed a space-time method for parabolic evolution equations.

The main aims of this thesis are first to generalize the results for a space-time scheme proposed by Langer, Moore and Neumüller in [16], where the authors use IgA for the discretization, to the case of moving interfaces, i.e., \( t \)-dependent, discontinuous coefficients, and, second, to allow local, i.e., element-wise, stabilization. The last issue is important for space-time adaptivity. Instead of IgA, we will use a conforming finite element method (FEM) to discretize the parabolic initial-boundary value problem,
which we specify in the following. Then we will shortly describe a second space-time scheme, introduced by Toulopoulos in [32]. Let $Q = Q_T := \Omega \times (0, T)$ be the space-time cylinder, with $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, being a sufficiently smooth and bounded spatial domain, and $T > 0$ being the final time. Furthermore, let $\Sigma := \partial \Omega \times (0, T)$, $\Sigma_0 := \Omega \times \{0\}$ and $\Sigma_T := \Omega \times \{T\}$ such that $\partial Q = \Sigma \cup \Sigma_0 \cup \Sigma_T$. Then we consider the following model problem that can formally be written as follows: Given $f$, $g$, $\nu$ and $u_0$, find $u$ such that (s.t.)

$$\frac{\partial u}{\partial t}(x, t) - \text{div}_x(\nu(x, t)\nabla_x u(x, t)) = f(x, t), \quad (x, t) \in Q,$$  

$$u(x, t) = g(x, t) = 0, \quad (x, t) \in \Sigma,$$  

$$u(x, 0) = u_0(x) = 0, \quad x \in \Omega.$$

where the diffusion coefficient (reluctivity in electromagnetics) $\nu$ is a given uniformly positive and bounded coefficient. The dependence of $\nu$ not only on space but also on time enables us to model moving interfaces. Note that we do not require $\nu$ to be smooth. In fact, we will admit discontinuities for $\nu$. For simplicity, we assume homogeneous Dirichlet boundary and initial conditions.

The thesis is structured in the following way: In Chapter 2 we will introduce some basic notation and basic mathematical concepts used throughout the thesis. In Chapter 3 we will consider the existence and uniqueness of a weak solution to the parabolic initial boundary value problem (1.1) – (1.3). In Chapter 4 we will derive a stable discrete variational formulation and the space-time finite element scheme. Moreover, we will prove an a priori error estimate. Additionally, we shortly describe the space-time finite element scheme with stabilizing bubble functions by Toulopoulos. In Chapter 5 we discuss the numerical results. In Chapter 6 we draw some conclusions, and provide an outlook to future work.
Chapter 2

Preliminaries

In this chapter, we introduce some definitions and notations, which we will use throughout this thesis. We start with the basic notation.

2.1 Basic Notation

Notation. We denote the Euclidean inner product on $\mathbb{R}^d$ by

$$x \cdot y := \sum_{i=1}^{d} x_i y_i,$$

for $x, y \in \mathbb{R}^d$. We define the Euclidean standard norm by

$$|x| = \left(\sum_{i=1}^{d} |x_i|^2\right)^{1/2},$$

for $x \in \mathbb{R}^d$.

Notation (Kronecker symbol).

$$\delta_{ij} = \begin{cases} 1, & \text{for } i = j, \\ 0, & \text{else.} \end{cases}$$

Notation (Multiindex). Let $\alpha = (\alpha_1, \ldots, \alpha_{d+1})$ be a vector. If all components $\alpha_i$ are non-negative integers, we call $\alpha$ a multiindex of order

$$|\alpha| = \alpha_1 + \cdots + \alpha_{d+1}.$$

Notation (Partial derivatives). Let $u : \mathbb{R}^{d+1} \to \mathbb{R}$, and $\alpha$ be a multiindex. Then we define

$$D^\alpha u = \frac{\partial^{\alpha_1} u}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_{d+1}} u}{\partial x_{d+1}^{\alpha_{d+1}}}.$$
We denote the full gradient of a sufficiently smooth scalar function $u : \mathbb{R}^{d+1} \to \mathbb{R}$ by
\[
\nabla u := (\nabla_x u, \partial_t u)^\top,
\]
where
\[
\nabla_x u := \left( \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d} \right)^\top, \quad (2.2)
\]
\[
\partial_t u := \frac{\partial u}{\partial x_{d+1}}. \quad (2.3)
\]
Moreover, we denote the (spatial) divergence of a sufficiently smooth vector field $u : \mathbb{R}^{d+1} \to \mathbb{R}^d$ by
\[
\text{div } u = \frac{\partial u_1}{\partial x_1} + \cdots + \frac{\partial u_d}{\partial x_d}, \quad (2.4)
\]
\[
\text{div}_x u = \frac{\partial u_1}{\partial x_1} + \cdots + \frac{\partial u_d}{\partial x_d}. \quad (2.5)
\]
Note that we treat the time $t$ as just another variable $x_{d+1}$.

**Remark 2.1.** We denote the derivative of a one-dimensional function $u : \mathbb{R} \to \mathbb{R}$ by
\[
u' : x \mapsto \frac{d}{dx} u(x).
\]

**Notation.** We will use the following spaces of (scalar) continuous functions:
\[
C(\Omega) := \{v : \Omega \to \mathbb{R} : v \text{ is continuous on } \Omega\},
\]
\[
C^k(\Omega) := \{v \in C(\Omega) : D^\alpha v \in C(\Omega), \text{ for all } |\alpha| \leq k\},
\]
\[
C^\infty(\Omega) := \{v \in C(\Omega) : v \text{ is infinitely differentiable }\},
\]
\[
C_c^\infty(\Omega) := \{v \in C^\infty(\Omega) : v \text{ has compact support}\},
\]
\[
C^{k,m}(\mathcal{Q}) := \{v \in C(\Omega) : D^{(\alpha_1, \ldots, \alpha_d, 0)} v \in C(\Omega), |(\alpha_1, \ldots, \alpha_d, 0)| \leq k \land \partial_j^l v \in C(\Omega), j \leq m\}
\]
for $k, m \in \mathbb{N}_0$. We identify $C^0(\Omega) = C(\Omega)$. For vector fields, e.g., $u : \mathbb{R}^{d+1} \to \mathbb{R}^d$, we write
\[
[C(\Omega)]^d := \{v : \Omega \to \mathbb{R}^d : v_i \in C(\Omega), i = 1, \ldots, d\}.
\]
With this notations, we can now formulate smoothness conditions for a solution to the model problem (1.1) – (1.3). If we require the right hand side $f$ to be continuous and the diffusion coefficient $\nu$ to be continuously differentiable in $\mathcal{Q}$, then the exact solution
\[

u \in C^{2,1}(\mathcal{Q}) \cap C(\mathcal{Q} \cup \Sigma_0 \cup \Sigma),
\]
is called a classical solution. However, these are very strong restrictions on both the right hand side and the diffusion coefficient. So it is naturally to ask if we can relax the smoothness conditions on the given functions $f$ and $\nu$, and how does this affect the solution. This motivates the introduction of weak (or generalized) derivatives, which are formally defined as follows.

**Definition 2.2** (Weak (or generalized) derivative). Let $U \subset \mathbb{R}$ open and $u, w \in L_{1,loc}(U)$. If they satisfy the integral identity

$$
\int_U w(x) \varphi(x) \, dx = - \int_U u \varphi'(x) \, dx, \quad \text{for all } \varphi \in C^\infty_c(U),
$$

(2.6)

then we call $w$ the weak (or generalized) derivative of $u$. Here,

$$L_{1,loc} := \{ u : U \to \mathbb{R} : v \in L_1(V) \text{ for each } V \subset \subset U \}.$$

One can show that this weak derivative is uniquely defined up to a set of measure zero (see [8]).

**Remark 2.3.** For functions $u : \mathbb{R}^{d+1} \to \mathbb{R}$, we can define the weak gradient $w : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ in a similar manner to Definition 2.2, i.e., if

$$
\int_U w(x,t) \cdot \varphi(x,t) \, d(x,t) = - \int_U u(x,t) \text{div} \varphi(x,t) \, d(x,t), \quad \forall \varphi \in [C^\infty_c(U)]^{d+1}
$$

(2.7)

then we call $w$ the weak gradient of $u$. Similarly, if $w : \mathbb{R}^{d+1} \to \mathbb{R}^d$ satisfies

$$
\int_U w(x,t) \cdot \psi(x,t) \, d(x,t) = - \int_U u(x,t) \text{div}_x \psi(x,t) \, d(x,t), \quad \forall \psi \in [C^\infty_c(U)]^d,
$$

(2.8)

then we call $w$ the weak spatial gradient and if $w : \mathbb{R}^{d+1} \to \mathbb{R}$

$$
\int_U w(x,t) \eta(x,t) \, d(x,t) = - \int_U u(x,t) \partial_t \eta(x,t) \, d(x,t), \quad \forall \eta \in C^\infty_c(U),
$$

(2.9)

then $w$ is called the weak temporal derivative.

**Remark 2.4.** If $w$ is the weak derivative of $u$ in the sense of Definition 2.2 or Remark 2.3 then we use the notation

$$u' = w, \quad \text{or} \quad \nabla u = w,$$

respectively.
2.2 Function spaces

In this section, we will briefly present Banach and Hilbert spaces, which we need throughout the thesis. For further analysis of such spaces and their properties, we refer to e.g. [6], and the references given therein. We start with the Lebesgue spaces of order $p$, which we introduce in the following definition. In the following and throughout this thesis, $\Omega \subset \mathbb{R}^n$ is always a bounded Lipschitz domain.

**Definition 2.5.** Let $\Omega \subset \mathbb{R}^n$. For $1 \leq p < \infty$, the space $L^p(\Omega)$ consists of all measurable functions $u$, such that
\[
\int_{\Omega} |u|^p \, dx < \infty.
\]
A measurable function $u$ belongs to the space $L^\infty(\Omega)$, if it is essentially bounded, i.e.,
\[
\text{ess sup}_{x \in \Omega} |u(x)| < \infty.
\]

If we equip the Lebesgue spaces $L^p(\Omega)$ with the norm
\[
\|u\|_{L^p(\Omega)} := \left\{ \left( \int_{\Omega} |u|^p \, dx \right)^{1/p} \right. \quad \text{for } 1 \leq p < \infty,
\]
\[
\text{ess sup}_{x \in \Omega} |u(x)| < \infty, \quad \text{for } p = \infty,
\]
one can show the following result.

**Theorem 2.6.** For $1 \leq p \leq \infty$, the Lebesgue spaces $L^p(\Omega)$ are Banach spaces with the norm $\| \cdot \|_{L^p(\Omega)}$. Moreover, the space $L^2(\Omega)$ is a Hilbert space equipped with the inner product
\[
(u, v) = (u, v)_{L^2(\Omega)} := \int_{\Omega} u(x)\overline{v(x)} \, dx, \forall u, v \in L^2(\Omega). \quad (2.10)
\]

**Proof.** See [6].

**Remark 2.7.** Two functions $f$ and $g$ are said to be identical in $L^p(\Omega)$, if $\|f - g\|_{L^p(\Omega)} = 0$. This means that if two functions $f$ and $g$ differ on a zero set $N \subset \Omega$, we still consider them identical in $L^p(\Omega)$, as they coincide almost everywhere (a.e.). Moreover, we deduce that a function in $L^p(\Omega)$ must only be defined a.e., i.e., in general, point evaluation is not well defined.

We can now combine the weak derivatives with these Lebesgue spaces, leading us to Sobolev spaces of weak derivatives. We again start with a formal definition:

**Definition 2.8 (W^k_p(\Omega)).** The Sobolev space $W^k_p(\Omega)$ consists of all functions $u$, whose weak derivatives up to index $k$ belong to the space $L^p(\Omega)$, i.e.,
\[
W^k_p(\Omega) := \{ u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega), \text{ for all } |\alpha| \leq k \},
\]
for $1 \leq p \leq \infty$. 

CHAPTER 2. PRELIMINARIES

On $W^k_p(\Omega)$, we can define the norm

$$\|u\|_{W^k_p(\Omega)} := \left( \sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p(\Omega)}^2 \right)^{1/2}.$$ 

**Remark 2.9.** There is an alternative way to introduce Sobolev spaces by using dense subspaces. Let us denote the closure of $C^\infty(\Omega) \cap W^k_p(\Omega)$ wrt the norm $\|\cdot\|_{W^k_p(\Omega)}$ by $H^k_p(\Omega)$, i.e.,

$$H^k_p(\Omega) := \overline{C^\infty(\Omega) \cap W^k_p(\Omega)} \|\cdot\|_{W^k_p(\Omega)}.$$ 

Meyers and Serrin showed in [17], that

$$W^k_p(\Omega) = H^k_p(\Omega),$$

i.e., the space $C^\infty(\Omega) \cap W^k_p(\Omega)$ is dense in $W^k_p(\Omega)$.

**Theorem 2.10.** For each $k = 1, \ldots$ and $1 \leq p \leq \infty$, the Sobolev space $W^k_p(\Omega)$ is a Banach space.

**Proof.** [1, 6, 8].

**Remark 2.11.** For the special case $p = 2$, we will use the notation

$$H^k(\Omega) := W^k_2(\Omega),$$

as these spaces are Hilbert spaces [1, 8], equipped with the inner product

$$(u, v)_{H^k(\Omega)} := \sum_{|\alpha| \leq k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)},$$

and the norm

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

Moreover, we will use the semi-norm

$$|u|_{H^k(\Omega)} := \sqrt{\sum_{|\alpha| = k} (D^\alpha u, D^\alpha u)_{L^2(\Omega)}}.$$ 

**Definition 2.12.** We denote the closure of $C^\infty_c(\Omega)$ in $W^k_p(\Omega)$ by $W^k_{p,0}(\Omega)$, i.e.,

$$W^k_{p,0}(\Omega) := \overline{C^\infty_c(\Omega)} \|\cdot\|_{W^k_p(\Omega)},$$

for $1 \leq p < \infty$ and $k \in \mathbb{N}$. Moreover, for $p = 2$, we use the notation

$$H^k_0(\Omega) := W^k_{2,0}(\Omega).$$
CHAPTER 2. PRELIMINARIES

This definition may seem rather abstract, however, we will give an equivalent characterisation of a function \( u \in W_{p,0}^k(\Omega) \) in the following.

If we want to apply the theory above in the context of boundary value problems, we have to consider expressions on the boundary \( \partial \Omega \) of \( \Omega \). But in general, as already mentioned, functions in Sobolev spaces only have to be defined a.e.. Moreover, as the boundary \( \partial \Omega \) is a zero set, two functions would be considered equal even for different boundary data. However, using the concept of traces, we can again prescribe values on the boundary \( \partial \Omega \).

**Theorem 2.13** (Trace theorem [8]). Let \( \Omega \) be bounded with sufficiently smooth boundary \( \partial \Omega \). Then there exists a bounded linear operator

\[
T : W_1^p(\Omega) \rightarrow L_p(\partial \Omega)
\]

such that

(i) \( Tu = u|_{\partial \Omega} \), for \( u \in W_1^1(\Omega) \cap C(\overline{\Omega}) \), and

(ii) \[
\|Tu\|_{L_p(\partial \Omega)} \leq C\|u\|_{W_1^1(\Omega)};
\]

for all \( u \in W_1^1(\Omega) \), where \( C \) does not depend on \( u \).

**Proof.** For the existence, see [6, 8], for the norm inequality see [8]. \( \square \)

**Remark 2.14.** For \( p = 2 \), we denote \( T(H^1(\Omega)) = H^{1/2}(\partial \Omega) \subset L_2(\partial \Omega) \).

For details on Sobolev spaces \( W_{k}^p(\Omega) \) with non-integer \( k \), we refer to [1, 6, 8].

**Definition 2.15.** We call \( Tu \) the trace of \( u \) on \( \partial \Omega \).

The above theorem gives only information about the traces of functions belonging to \( W_1^1(\Omega) \). However, we might be interested in traces of functions with higher order weak derivatives. The following theorem gives information about traces of functions belonging to \( W_p^k(\Omega) \).

**Theorem 2.16** ([6]). Let \( \Omega \) be a Lipschitz domain and \( m \in \mathbb{N} \), \( 1 \leq p \leq \infty \). Then \( W_{p,0}^k(\Omega) \) consists of all functions \( u \in W_p^k(\Omega) \) with \( TD^\alpha u = 0 \), for \( |\alpha| \leq k - 1 \).

**Proof.** See [6]. \( \square \)

**Remark 2.17.** For simplicity, we will not always write the trace operator for expressions on the boundary but instead use the same notation as for smooth functions, i.e., \( u|_{\partial \Omega} = Tu \) for \( u \in W_1^1(\Omega) \).

With these spaces and the definition of traces, we can now formally define special function spaces on the whole space-time cylinder \( Q \) (see [14]), which we will need in the analysis of our model problem (1.1) – (1.3).
Definition 2.18. We define the following Sobolev (Hilbert) spaces

\[ H^1_0(Q_T) = W^{1,0}_2(Q_T) := \{ u \in L^2(Q_T) : \nabla u \in L^2(Q_T) \land u|_\Sigma = 0 \}, \]
\[ H^{1,0}(Q_T) = W^{1,0}_2(Q_T) := \{ u \in L^2(Q_T) : \nabla_x u \in L^2(Q_T) \}, \]
\[ \hat{H}^{1,0}(Q_T) = \hat{W}^{1,0}_2(Q_T) := \{ u \in H^{1,0}(Q_T) : u|_\Sigma = 0 \}, \]

equipped with the usual inner products and norms, as well as the Banach space

\[ V_2(Q_T) := \{ u \in H^{1,0}(Q_T) : |u|_{Q_T} < \infty \}, \]

with subspaces

\[ \hat{V}_2(Q_T) := \{ u \in \hat{H}^{1,0}(Q_T) : |u|_{Q_T} < \infty \}, \]
\[ V_2^1(Q_T) := \{ u \in V_2(Q_T) : \lim_{\Delta t \to 0} \| u(\cdot, t + \Delta t) - u(\cdot, t) \|_{L^2(\Omega)} = 0, \text{uniformly on } [0, T] \}, \]
\[ \hat{V}_2^1(Q_T) := V_2^{1,0}(Q_T) \cap \hat{H}^{1,0}(Q_T), \]

where the norm |.|_{Q_T} is defined by

\[ |u|_{Q_T} := \max_{0 \leq \tau \leq t} \| u(\cdot, \tau) \|_{L^2(\Omega)} + \| \nabla_x u \|_{Q_T}. \] (2.11)

We will make use of the above theory in Chapter 3 where we will introduce a new class of solutions to the model problem (1.1) – (1.3), the so called weak (or generalized) solutions.

2.3 Additional Theorems

When we consider operator equations, e.g.,

\[ Bu = g, \]

we want to know whether this equation has a unique solution. The following theorem provides conditions under which there is indeed a unique solution.

Theorem 2.19 ([27, Thm. 3.7]). Let \( X \) and \( \Pi \) be Hilbert spaces and let \( B : X \to \Pi' \) be a bounded and linear operator. Further we assume the stability condition

\[ c_S \| v \|_X \leq \sup_{\theta \neq q \in \Pi} \frac{\langle Bv, q \rangle}{\| q \|_\Pi} \text{ for all } v \in (\ker B)\perp. \]

For a given \( g \in \text{Im}_X(B) \) there exists a unique solution \( u \in (\ker B)\perp \) of the operator equation \( Bu = g \) satisfying

\[ \| u \|_X \leq \frac{1}{c_S} \| g \|_{\Pi'}. \] (2.12)

Proof. See [27].
Chapter 3

Space-time variational formulations

In this chapter, we will focus on deriving a variational formulations for our model problem (1.1) – (1.3). We want to show the existence and uniqueness of a solution, and, moreover, to what class of solutions it belongs. There are different ways to obtain existence and uniqueness results, see e.g. [14, 28]. We will focus on the techniques presented by Ladyzhenskaya in [14].

3.1 Formulation in Sobolev spaces on the space-time cylinder

Now let us reconsider the model problem (1.1) – (1.3) : Find $u$ s.t.

$$
\mathcal{M}u \equiv \partial_t u - \text{div}(\nu \nabla_x u) = f \text{ in } Q_T,
$$

$$
u \leq \nu(x,t) \leq \overline{\nu}, \quad \text{for almost all } (x,t) \in Q_T,
$$

with given data

$$
\varphi \in L_2(\Omega) \quad \text{and} \quad f \in L_{2,1}(Q_T) := \{ v : \int_0^T \| f(\cdot, t) \|_{L_2(\Omega)} \, dt < \infty \},
$$

and a uniformly bounded coefficient

where $\underline{\nu}, \overline{\nu} = \text{const.} > 0$. To show now the existence of a weak solution in an appropriate function space, we use Galerkin’s method. We formally start with multiplying the PDE by the solution $u$ and integrate over the truncated space-time domain $Q_t = \Omega \times (0,t)$, $t \in (0,T)$, i.e.,

$$
\int_{Q_t} \mathcal{M}u \cdot u \, dx \, dt = \int_{Q_t} f u \, dx \, dt.
$$
CHAPTER 3. SPACE-TIME VARIATIONAL FORMULATIONS

Using integration by parts, the homogeneous boundary condition on the lateral boundary $\Sigma$ and $n_x = 0$ on $\Sigma_0 \cup \Sigma_t$, where $\Sigma_t := \Omega \times \{t\}$, we obtain

\[
\int_{Q_t} \mathcal{M} u \cdot u \, dx \, dt = \int_{Q_t} \partial_t uu - \text{div}_x (\nu(x, t) \nabla_x u) \, dx \, dt = \int_{Q_t} \frac{1}{2} \partial_t (u^2) \, dx \, dt + \int_{Q_t} \nu(x, t) |\nabla_x u|^2 \, dx \, dt,
\]

for the left hand side of (3.5). Now we use Gauss’ theorem and the fact that $n_t \equiv 0$ on $\Sigma$ to get rid of the time derivative and we obtain the following identity

\[
\frac{1}{2} \|u(\cdot, t)\|_{L^2(\Omega)}^2 + \int_{Q_t} \nu(x, t) |\nabla_x u|^2 \, dx \, dt = \frac{1}{2} \|u(\cdot, 0)\|_{L^2(\Omega)}^2 + \int_{Q_t} f \cdot u \, dx \, dt. \quad (3.6)
\]

We call (3.6) energy balance equation. From this equation, we will derive a bound for $u$ in some specific norm $|\cdot|_{Q_t}$. First, we estimate the left hand side (lhs) of (3.6) from below by

\[
\frac{1}{2} \|u(\cdot, t)\|_{L^2(\Omega)}^2 + \int_{Q_t} \nu(x, t) |\nabla_x u|^2 \, dx \, dt \geq \frac{1}{2} \|u(\cdot, t)\|_{L^2(\Omega)}^2 + \nu \int_{Q_t} |\nabla_x u|^2 \, dx \, dt,
\]

and the right hand side (rhs) of (3.6) from above by

\[
\frac{1}{2} \|u(\cdot, 0)\|_{L^2(\Omega)}^2 + \int_{Q_t} f \cdot u \, dx \, dt \leq \frac{1}{2} \|u(\cdot, 0)\|_{L^2(\Omega)}^2 + \int_0^t \int_{\Omega} f(x, \tau) u(x, \tau) \, dx \, d\tau + \int_0^t \int_{\Omega} f(x, \tau) u(x, \tau) \, dx \, d\tau \leq \frac{1}{2} \|u(\cdot, 0)\|_{L^2(\Omega)}^2 + \int_{\Omega} f \cdot u \, dx \, dt \leq \frac{1}{2} \|u(\cdot, 0)\|_{L^2(\Omega)}^2 + \nu \int_{Q_t} \max_{\tau \in [0, t]} |u(\cdot, \tau)|_{L^2(\Omega)} \, dx \, dt.
\]

Combining these two estimates gives us

\[
\frac{1}{2} \|u(\cdot, t)\|_{L^2(\Omega)}^2 + \nu \|
abla_x u\|_{Q_t}^2 \leq \|u(\cdot, 0)\|_{L^2(\Omega)}^2 + \max_{0 \leq \tau \leq t} \|u(\cdot, \tau)\|_{L^2(\Omega)} \|f\|_{L^1(0, t)} \cdot (3.7)
\]

Denoting $\max_{0 \leq \tau \leq t} \|u(\cdot, \tau)\|_{L^2(\Omega)}$ by $y(t)$ and multiplying (3.7) by 2, we obtain

\[
\|u(\cdot, t)\|_{L^2(\Omega)}^2 + 2\nu \|
abla_x u\|_{Q_t}^2 \leq 2y(t)\|u(\cdot, 0)\|_{L^2(\Omega)} + 2y(t)\|f\|_{L^1(0, t)} = j(t),
\]

where we used the estimate $\|u(\cdot, 0)\|_{L^2(\Omega)} \leq \max_{0 \leq \tau \leq t} \|u(\cdot, \tau)\|_{L^2(\Omega)} \|u(\cdot, 0)\|_{L^2(\Omega)}$. From this, we deduce two inequalities, i.e,

\[
y(t)^2 \leq j(t) \quad \text{and} \quad \|
abla_x u\|_{Q_t}^2 \leq (2\nu)^{-1} j(t). \quad (3.8)
\]
The second estimate can easily be verified, whereas the first one is obtained by estimating the lhs from below by \( \|u(., t)\|_{L^2(\Omega)} \). This expression holds for any \( \tau \in [0, t] \), hence it holds also for the maximum. However, the only terms in \( j(t) \) depending on \( t \) are \( y(t) \), where we already take a maximum over \([0, t]\). Thus, the first expression of (3.8) follows. We take the square-root of both expressions and add them up to obtain

\[
|u|_{\mathcal{Q}_t} = y(t) + \|\nabla_x u\|_{\mathcal{Q}_t} \leq (1 + \frac{1}{2L^2})^{-1/2} |u|_{\mathcal{Q}_t}^{1/2} (\|u(., 0)\|_{L^2(\Omega)} + 2 \|f\|_{2,1,\mathcal{Q}_t})^{1/2}. \tag{3.9}
\]

We bring similar terms on the same side and take the square on each side of the inequality. Thus we have obtained an upper bound for \( |u|_{\mathcal{Q}_t} \) in the form

\[
|u|_{\mathcal{Q}_t} \leq (1 + \frac{1}{2L^2})^{-1} (\|u(., 0)\|_{L^2(\Omega)} + 2 \|f\|_{2,1,\mathcal{Q}_t}) =: c\mathcal{F}(t), \tag{3.10}
\]

which holds for any \( t \in [0, T] \). However, this bound requires a solution where point evaluation with respect to (wrt) time is well defined. Before we proof that our problem (3.1) has such a weak solution, we have to introduce a suitable definition of the weak solution.

**Definition 3.1.** A function \( u \in \tilde{H}^{1,0}(\mathcal{Q}_T) \) is called a generalized (weak) solution in \( H^{1,0}(\mathcal{Q}_T) \) of the parabolic initial-boundary value problem (3.1) – (3.2) if it satisfies the identity

\[
\mathcal{M}(u, v) \equiv \int_{\mathcal{Q}_T} -u \partial_t v + \nu(x, t) \nabla_x u \nabla_x v \, dx \, dt
= \int_{\Omega} \varphi v(., 0) \, dx + \int_{\mathcal{Q}_T} f v \, dx \, dt, \tag{3.11}
\]

for all \( v \in \tilde{H}_0^1(\mathcal{Q}_T) := \{ v \in H_0^1(\mathcal{Q}_T) : v = 0 \text{ on } \Sigma_T \} \).

To proof solvability of (3.1) in this class, i.e. solvability of (3.11), we will use Galerkin’s method. Let \( \{ \varphi_j \} \) be a \( L_2 \)-orthonormal fundamental system in \( \mathcal{W}_2^1(\Omega) \). In (3.1), we substitute \( u \) with an appropriate test function \( u^N \), multiply the obtained equation by each \( \varphi_j \) for \( j = 1, \ldots, N \) and integrate wrt \( x \) over \( \Omega \). We use integration by parts in the principle term, and obtain a system of \( N \) equations

\[
(\partial_t u^N, \varphi_j) + (\nu(., t) \nabla_x u, \nabla_x \varphi_j) = (f, \varphi_j), \tag{3.12}
\]

where \( (., .) = (., .)_{L_2(\Omega)} \) is the standard \( L_2(\Omega) \) inner product. In (3.12), we express \( u^N \) with the fundamental system \( \{ \varphi_j \} \), i.e., \( u^N(x, t) := \sum_{j=1}^N c_j^N(t) \varphi_j(x) \). We can rewrite (3.12) wrt the coefficient functions \( c_i(t) = c_i^N(t) \),

\[
\sum_{j=1}^N \frac{d}{dt} c_j(t) (\varphi_j, \varphi_i) + \sum_{j=1}^N c_j(t) (\nu(., t) \nabla_x \varphi_j, \nabla_x \varphi_i) = (f(., t), \varphi_i), \text{ for } i = 1, \ldots, N.
\tag{3.13}
\]
CHAPTER 3. SPACE-TIME VARIATIONAL FORMULATIONS

with the initial condition

$$\sum_{j=1}^{N} c_j(0)(\varphi_j, \varphi_i) = (\varphi, \varphi_i),$$

(3.14)

that is nothing but the $L_2$-projection to span$\{\varphi_1, \ldots, \varphi_N\}$. This system is a system of $N$ linear ordinary differential equations with principal terms $\frac{d}{dt} c_i(t)$ and bounded coefficient functions in front of the zero-order terms $c_i(t)$. This system has a unique solution of absolutely continuous functions $c_i^N(t)$, $l = 1, \ldots, N$ and sum these equations up from 1 to $N$. We proceed by integrating this sum over $(0,t)$ and obtain an equation of the form (3.6) for $u^N$, i.e.,

$$\frac{1}{2}\|u^N(.,t)\|_{L_2(\Omega)}^2 + \int_{Q_t} \nu(x,t)|\nabla_x u^N| \, dx \, dt = \frac{1}{2}\|u^N(.,0)\|_{L_2(\Omega)}^2 + \int_{Q_t} f \, dx \, dt.$$  

(3.15)

We can derive a bound for $|u^N|_{Q_T}$ from (3.15) in the same manner as we did for $|u|_{Q_T}$ from (3.6), i.e.,

$$|u^N|_{Q_t} \leq \tilde{c}\left(\|u^N(.,0)\|_{L_2(\Omega)} + 2\|f\|_{L_1(Q_t)}\right).$$

Furthermore, we know the upper bound $\|u^N(.,0)\|_{L_2(\Omega)} \leq \|\varphi\|_{L_2(\Omega)}$. Therefore, we obtain the bound

$$|u^N|_{Q_T} \leq \tilde{c},$$

(3.16)

where $\tilde{c}$ is a constant independent of $N$. Hence the sequence $\{u^N\}$ is a bounded sequence in the Hilbert space $L_2(Q_T)$. This can easily be deduced from (3.16) and Definition 2.18 Hilbert spaces are reflexive spaces, see e.g. [6]. Thus, $\{u^N\}$ has a weakly convergent subsequence $\{u^{N_k}\}$. The same holds true for its derivatives $\{\nabla_x u^{N_k}\}$ with $\{\nabla_x u^{N_k}\}$. Therefore, $\{u^{N_k}\}$ and $\{\nabla_x u^{N_k}\}$ converge weakly to some unique element $u \in H^{1,0}(Q_T)$. Is this $u$ the desired generalized (weak) solution of our model problem (3.1) - (3.2)? Let us again multiply (3.12) by some arbitrary absolutely continuous functions $d_i(t)$ with $\frac{d}{dt}d_i \in L_2(0,T)$, with $d_i(T) = 0$. We sum the obtained equations up from 1 to $N$, integrate over the interval $(0,T)$ and perform integration by parts with respect to time. The resulting equation is

$$\int_{Q_T} -u^N \partial_t \Phi + \nu(x,t)\nabla_x u^N \nabla_x \Phi \, dx \, dt = \int_{Q_T} u^N \Phi |_{t=0} \, dx + \int_{Q_T} f \, \Phi \, dx \, dt,$$  

(3.17)

for $\Phi(x,t) = \sum_{k=1}^{N} d_k(t)\varphi_k(x)$. The set of all such functions $\Phi$ with the desired properties of $d_i$ is denoted by $\mathcal{M}_N$. The superset $\bigcup_{p=1}^{\infty} \mathcal{M}_p$ is dense in $\hat{H}_0^1(Q_T)$ (see [14]). We fix a $\Phi \in \mathcal{M}_p$ and take the limit of (3.17) for $N_k \geq p$, i.e.,

$$\int_{Q_T} -u^{N_k} \partial_t \Phi + \nu(x,t)\nabla_x u^{N_k} \nabla_x \Phi \, dx \, dt = \int_{Q_T} u^{N_k} \Phi |_{t=0} \, dx + \int_{Q_T} f \, \Phi \, dx \, dt.$$  

(3.18)
We obtain exactly the definition of a generalized \((\text{weak})\) solution \((3.11)\), with \(v = \Phi \in \mathcal{M}\). As these union of all such spaces is dense in \(H^1_0(Q_T)\), the equation \((3.18)\) holds for any \(v \in \hat{H}^1_0(Q_T)\). Thus \(u\) is indeed a generalized solution of our model problem \((3.1) - (3.2)\). We gather these results in the following theorem.

**Theorem 3.2** ([14, Chapter III, Thm. 3.1]). Under the conditions \((3.3)\) and \((3.4)\), the problem \((3.1) - (3.2)\) has at least one generalized \((\text{weak})\) solution in \(\mathring{\mathcal{H}} V^1_2(Q_T)\), as defined in Definition 3.1.

**Proof.** Follows from the derivation above.

We know now that at least one solution \(u\) exists, but is this solution unique? To prove this, we will make again use of the results presented by Ladyženskaya in [14, Chapter III, §2]. First, we consider our generalized solution \(u\) as a generalized solution in \(L^2(Q_T)\) of the problem

\[
\partial_t u - \Delta u = \tilde{f} + \text{div}_x(F) \quad \text{in } Q_T, \tag{3.19}
\]

\[
u(\nu \cdot \nabla_x u) - \nabla_x u \quad \text{in } Q_T, \tag{3.20}
\]

with \(\tilde{f} \equiv f\) and \(F_i = \nu(x, t)\nabla_x u - \nabla_x u\). Hence, by [14] Chapter III, Thm. 2.2 & Thm. 2.3, it follows that \(u(x, t)\) is a generalized solution of \((3.19) - (3.20)\) in \(V^1_2(Q_T)\). By this, we can define a new class of generalized solutions.

**Definition 3.3** ([14, Chapter III]). A generalized solution \(u \in \hat{H}^{1,0}\) of \((3.1) - (3.2)\) is a called a generalized solution of \((3.1) - (3.2)\) in \(V^1_2(Q_T)\), if \(u \in V^1_2(Q_T)\) and it fulfils the energy-balance equation \((3.6)\) and the identity

\[
\int_{\Omega} u(x, t)v(x, t) \, dx - \int_{\Omega} \varphi v(x, 0) \, dx
+ \int_{Q_t} -u \partial_t v + \nu \nabla_x u \nabla_x v \, dxdt = \int_{Q_t} f v \, dxdt, \tag{3.21}
\]

for all \(v \in \hat{H}^1_0(Q_T)\) and any \(t \in (0, T)\).

We will show uniqueness of the problem \((3.1) - (3.2)\) in \(H^{1,0}(Q_T)\) as usual by contradiction. Let \(u_1 \neq u_2 \in H^{1,0}(Q_T)\) be two generalized solutions of \((3.1) - (3.2)\), then the difference \(u := u_1 - u_2\) is also a generalized solution of \((3.1) - (3.2)\), but with homogeneous initial data and zero right hand side. Moreover, by what we have shown above, it is also a generalized solution in \(V^1_2(Q_T)\), so it satisfies \((3.6)\) with zero right hand side. If it satisfies \((3.6)\), we have shown that its norm \(|u|_{Q_T}\) is subject to the bound \((3.10)\), but also with zero right hand side. We obtain \(u = u_1 - u_2 \equiv 0\), which is a contradiction to our assumption \(u_1 \neq u_2\). Moreover, the operator \(B\), which assigns each tuple \((f, \varphi)\) its generalized solution in \(V^1_2(Q_T)\) is linear and the energy balance equation \((3.6)\) can be obtained from the identity \((3.21)\) (see [14]). We can summarise the results in the following theorem.
Theorem 3.4 (Chapter III, Thm. 3.2). If the assumptions (3.3) and (3.4) are fulfilled, then any generalized solution of (3.1) – (3.2) in $\dot{H}^{1,0}(Q_T)$ is the generalized solution in $\mathcal{V}_2^{1,0}(Q_T)$ and it is unique in $\dot{H}^{1,0}(Q_T)$.

Corollary 3.5. If the assumptions (3.3) and (3.4) hold, then there exists a unique generalized solution $u \in \dot{H}^{1,0}(Q_T) \cap V^{1,0}_2(Q_T)$ to the problem (3.1) – (3.2).

3.2 Alternative approach in Bochner spaces of abstract functions

In the previous section we ensure the existence and uniqueness of a solution of (1.1) – (1.3) using Sobolev spaces defined on the whole space-time cylinder $Q$. An alternative approach makes use of Bochner spaces of abstract functions, see, e.g. [18, 26, 28]. In this section, we will briefly summarise the results of Steinbach in [28]. In order to do that, we first need to recall the definition of Bochner spaces.

Definition 3.6 (Bochner space). Let $X$ be a Banach space with norm $\| \cdot \|$. Then we define the spaces

\[
C(0, T; X) := \{ u : [0, T] \to X : u \text{ is continuous and } \| u \|_{C(0, T; X)} < \infty \},
\]

\[
L_p(0, T; X) := \{ u : (0, T) \to X : u \text{ is measurable and } \| u \|_{L_p(0, T; X)} < \infty \},
\]

of abstract functions mapping $[0, T]$ into some Banach space $X$, with the respective norms

\[
\| u \|_{C(0, T; X)} := \max_{0 \leq t \leq T} \| u(t) \|,
\]

\[
\| u \|_{L_p(0, T; X)} := \left( \int_0^T \| u(\tau) \|^p \, d\tau \right)^{1/p}.
\]

For further details on Bochner spaces, see e.g. [8, 34].

Let us consider once more the model problem (1.1) – (1.3), i.e. find $u$ s.t.

\[
\partial_t u - \text{div}_x(\nu \nabla_x u) = f \text{ in } Q,
\]

\[
u u = 0 \text{ on } \Sigma,
\]

\[
u u = u_0 \text{ on } \Sigma_0
\]

for given data $f$ and $u_0$. We derive the weak formulation in the usual manner and obtain the variational problem: find $u \in L_2(0, T; H^1_0(\Omega)) \cap H^1(0, T; H^{-1}(\Omega))$, with $u(\cdot, 0) = u_0(\cdot)$, s.t.

\[
a(u, v) := \int_0^T \int_\Omega \partial_t u v + \nu \nabla_x u \cdot \nabla_x v \, dx \, dt = \int_0^T \int_\Omega f \, v \, dx \, dt =: l(v),
\]
for all $v \in L_2(0, T; H^1_0(\Omega))$, where $\nu$ is as before, $f \in L_2(0, T; H^{-1}(\Omega))$ and $u(\cdot, 0) \in H^1_0(\Omega)$. We can not show any form of ellipticity for this Petrov-Galerkin formulation. Hence, Steinbach used a stability or inf-sup-condition to guarantee existence and uniqueness of an weak solution. The stability condition uses the corresponding energy norm

$$
\|u\|_{L_2(0,T;H^1_0(\Omega))\cap H^1(0,T;H^{-1}(\Omega))}^2 = \|\partial_t u\|^2_{L_2(0,T;H^{-1}(\Omega))} + \|u\|^2_{L_2(0,T; H^1_0(\Omega))}.
$$

**Theorem 3.7** ([28] Thm. 2.1). For all $u \in L_2(0, T; H^1_0(\Omega)) \cap H^1(0, T; H^{-1}(\Omega))$, with $u(\cdot, 0) = 0$ in $\Omega$, there holds the stability condition

$$
\frac{1}{2\sqrt{2}}\|u\|_{L_2(0,T;H^1_0(\Omega))\cap H^1(0,T;H^{-1}(\Omega))} \leq \sup_{0 \neq v \in L_2(0,T;H^1_0(\Omega))} \frac{a(u,v)}{\|v\|_{L_2(0,T;H^1_0(\Omega))}}. \quad (3.27)
$$

**Proof.** The proof makes use of the so called quasi-static elliptic Dirichlet boundary value problem, for further details see [28].

The initial condition can be seen as an Dirichlet datum for the space-time domain $Q$. Hence, we can introduce the splitting

$$
u(\cdot, \cdot) = \bar{\nu}(\cdot, \cdot) + \bar{u}_0(\cdot, \cdot) \text{ in } Q, \quad (3.28)
$$

where $\bar{u}_0 \in L_2(0, T; H^1_0(\Omega)) \cap H^1(0, T; H^{-1}(\Omega))$ is an extension of the initial datum $u_0 \in H^1_0(\Omega)$. Now we homogenise the variational problem (3.26), i.e. find $\bar{\nu} \in X$ s.t.

$$
a(\bar{\nu}, v) = (f, v)_Q - a(\bar{\nu}_0, v) \quad \text{for all } v \in Y, \quad (3.29)
$$

where $\bar{\nu}_0$ is a given extension of the initial datum $u \in H^1_0(\Omega)$ and $f \in L_2(0, T; H^{-1}(\Omega))$ is a given right hand side. We use the spaces

$$
X := \{u \in L_2(0, T; H^1_0(\Omega)) \cap H^1(0, T; H^{-1}(\Omega)) : u(\cdot, 0) = 0 \text{ in } \Omega\},
$$

$$
Y := L_2(0, T; H^1_0(\Omega)).
$$

The unique solvability of (3.29) follows now from Theorem 2.19.

**Corollary 3.8** ([28] Col. 2.2]). Let $\bar{u}_0 \in L_2(0, T; H^1_0(\Omega)) \cap H^1(0, T; H^{-1}(\Omega))$ be some given extension of the initial datum $u_0 \in H^1_0(\Omega)$ and assume $f \in L_2(0, T; H^{-1}(\Omega))$. Then the bilinear form $a(\cdot, \cdot)$ as given in (3.26) is bounded and satisfies the stability condition (3.27). Hence there exists a unique solution $\bar{u} \in X$ of the variational problem (3.29) satisfying the a priori estimate

$$
\|\bar{u}\|_{L_2(0,T;H^1_0(\Omega))\cap H^1(0,T;H^{-1}(\Omega))} \leq 2\sqrt{2}[\|f\|_{L_2(0,T;H^{-1}(\Omega))} + \sqrt{2}\|u_0\|_{L_2(0,T;H^1_0(\Omega))\cap H^1(0,T;H^{-1}(\Omega))}]. \quad (3.30)
$$
Chapter 4

Space-time Finite Element Methods

4.1 A Space-time Finite Element Method based on time upwind stabilization

From the previous chapter, we know that there exists a unique generalized solution of the initial-boundary value problem (1.1) – (1.3) in $H^{1,0}(\mathcal{Q}) \cap \dot{V}_2^{1,0}(\mathcal{Q})$. The goal of this chapter is to derive a stable space-time finite element scheme with a coercive (elliptic) discrete bilinear form, and, therefore, to ensure existence and uniqueness of a finite element solution. Similar to Langer et.al. in [16], we use special time-upwind test functions that are locally scaled in our case. First, we need a regular triangulation $\mathcal{T}_h$ of our space-time domain $\mathcal{Q}$ (for details, see e.g. [2, 4]). We formally define this triangulation as

$$\mathcal{T}_h := \{ E : E \subset \mathcal{Q} \text{ and } E \text{ open} \}$$

with the properties

$$\mathcal{Q} = \bigcup_{E \in \mathcal{T}_h} \bar{E} \quad \text{and} \quad E \cap E' = \emptyset \text{ for all } E' \in \mathcal{T}_h \text{ with } E \neq E'. \quad (4.2)$$

On each of these elements $E$, we define individual time upwind test functions

$$v_{h,t}(x,t) := v_h(x,t) + \theta_E h_E \partial_t v_h(x,t), \text{ for all } (x,t) \in E, \quad (4.3)$$

where $\theta_E$ is a positive parameter that will be defined later, and $h_E := \text{diam}(E)$. Here, $v_h$ is some test function from a standard conform finite element space $V_{0h}$, e.g., $V_{0h} = \{ v \in C(\mathcal{Q}) : v|_E \in \mathbb{P}_p \subset Q_p \}$ that is considered in this thesis. From now on, unless specified otherwise, all functions depend on both space and time variables. For simplicity, we can omit the arguments. In this chapter, we will make use of the
following spaces:

\[ V_0 = H^{1,0}_0(Q) := \{ u \in L_2(Q) : \nabla_x u \in L_2(Q), \partial_t u \in L_2(Q) \text{ and } u|_{\Sigma_0} = 0 \}, \]  

\[ H^{2,1}_0(T_h) := \{ v \in H^{1,1}_0(Q) : v|_E \in H^{2,1}(E), \forall E \in T_h \}, \]  

\[ W^{1}_{\infty}(T_h) := \{ v \in L_{\infty}(Q) : v|_E \in W^{1}_{\infty}(E), \forall E \in T_h \}. \]

We assume that \( \nu \in W^{1}_{\infty}(T_h) \) and that the PDE has a sufficiently smooth solution \( u \), e.g., \( u \in H^{2,1}_0(T_h) \). Then we proceed in the usual manner, i.e., we first multiply the PDE (1.1) by our space-time test function \( v_{h,t} \), and then integrate over a single element \( E \), obtaining

\[
\int_E (\partial_t u - \text{div}_x(\nu \nabla_x u)) v_{h,t} \, d(x,t) = \int_E f v_{h,t} \, d(x,t).
\]

Summing up over all elements and applying integration by parts on the principle term, we obtain

\[
\sum_{E \in T_h} \int_E \partial_t u \, v_{h,t} + \nu \nabla_x u \cdot \nabla_x v_{h,t} \, d(x,t) - \int_{\partial E} \nu \nabla_x u \cdot n_x v_{h,t} \, ds_{(x,t)} = \sum_{E \in T_h} \int_E \partial_t v_{h,t} + \theta_E h_E \partial_t u \partial_t v_{h} + \nu \nabla_x u \cdot \nabla_x v_{h} + \theta_E h_E \nu \nabla_x u \cdot \nabla_x (\partial_t v_h) \, d(x,t) \\
- \int_{\partial E} \nu \nabla_x u \cdot n_x v_{h} + \theta_E h_E \nu \nabla_x u \cdot n_x \partial_t v_{h} \, ds_{(x,t)},
\]

for the left hand side, while the right hand side remains unchanged. For the exact solution \( u \) of (1.1) - (1.3), we know that the fluxes have to be continuous, i.e., let \( E \) and \( E' \) be two adjacent elements, then

\[
(\nu \nabla_x u \cdot n_x)|_E = (\nu \nabla_x u \cdot n_x)|_{E'}.
\]  

(4.7)

From this, we know that one part of the boundary terms vanishes from all inner edges, i.e. we obtain

\[
\sum_{E \in T_h} \int_E \partial_t u \, v_{h} + \theta_E h_E \partial_t u \partial_t v_{h} + \nu \nabla_x u \cdot \nabla_x v_{h} + \theta_E h_E \nu \nabla_x u \cdot \nabla_x (\partial_t v_h) \, d(x,t) \\
- \sum_{E \in T_h} \int_{\partial E} \nu \nabla_x u \cdot n_x v_{h} \, ds_{(x,t)} \\
- \sum_{E \in T_h} \int_{\partial E} \nu \theta_E h_E \nabla_x u \cdot n_x \partial_t v_{h} \, ds_{(x,t)} = \sum_{E \in T_h} \int_E f v_{h,t} \, d(x,t).
\]

We require \( v_{h} \) to be zero on \( \Sigma \), and know that \( n_x \) vanishes on \( \Sigma_0 \) and \( \Sigma_T \). Therefore, the first boundary term completely disappears from our equation, and we obtain

\[
\sum_{E \in T_h} \int_E [\partial_t u v_{h} + \theta_E h_E \partial_t u \partial_t v_{h} + \nu \nabla_x u \cdot \nabla_x v_{h} + \theta_E h_E \nu \nabla_x u \cdot \nabla_x (\partial_t v_h)] \, d(x,t) \\
- \sum_{E \in T_h} \int_{\partial E} \nu \theta_E h_E \nabla_x u \cdot n_x \partial_t v_{h} \, ds_{(x,t)} = \sum_{E \in T_h} \int_E f v_{h} + \theta_E h_E \partial_t v_{h} \, d(x,t).
\]
We now arrived at the consistency identity for (1.1)

\[ a_h(u, v_h) = l_h(v_h), \quad \forall v_h \in V_{0h}, \quad (4.8) \]

that holds for a sufficiently smooth solution \( u \), e.g., \( u \in H^2_{0, \mathcal{Q}}(T_h) \), where

\[
a_h(u, v_h) := \sum_{E \in T_h} \int_E \partial_t u \, v_h + \theta_E h_E \partial_t u \, \partial_t v_h \, d(x, t)
+ \int_E \nu \nabla_x u \cdot \nabla_x v_h + \theta_E h_E \nu \nabla_x u \cdot \nabla_x (\partial_t v_h) \, d(x, t)
- \int_{\partial E} \theta_E h_E \nu \nabla_x u \cdot n_x \partial_t v_h \, ds(x, t), \quad (4.9)
\]

\[
l_h(v_h) := \sum_{E \in T_h} \int_E f(v_h + \theta_E h_E \partial_t v_h) \, d(x, t), \quad (4.10)
\]

with given \( \nu \in W^1_{\infty}(T_h) \) and \( f \in L^2(Q) \).

**Remark 4.1.** We can derive an equivalent scheme to (4.9). In particular, we perform the same steps as above, but instead of applying integration by parts on both principal terms, we only apply it to the first principal term and keep the second. Hence we obtain another consistency identity for (1.1)

\[ \tilde{a}_h(u, v_h) = l_h(v_h), \quad \forall v_h \in V_{0h}, \]

that holds for a sufficiently smooth solution \( u \), e.g., \( u \in H^2_{0, \mathcal{Q}}(T_h) \), where

\[
\tilde{a}_h(u, v_h) := \sum_{E \in T_h} \int_E \partial_t u \, v_h + \theta_E h_E \partial_t u \, \partial_t v_h \, d(x, t)
+ \int_E \nu \nabla_x u \cdot \nabla_x v_h + \theta_E h_E \nu \nabla_x u \cdot \nabla_x (\partial_t v_h) \, d(x, t)
+ \int E \nabla_x u \cdot \nabla_x (\nu \nabla_x u) \partial_t v_h \, d(x, t)
\]

with \( \nu \in W^1_{\infty}(T_h) \) and \( f \in L^2(Q) \), and \( l_h \) as in (4.10).

**Remark 4.2.** If the test functions \( v_h \in V_{0h} \) are continuous and piecewise linear \((p = 1)\), then the term in (4.9) containing \( \nabla_x (\partial_t v_h) \) vanishes in all elements \( E \in T_h \), since it only contains mixed second order derivatives.

Now we look for a Galerkin approximation \( u_h \in V_{0h} \) to the generalized solution \( u \) of our initial boundary value problem (1.1) – (1.3) using the variational identity (4.8), i.e., find \( u_h \in V_{0h} \) such that

\[ a_h(u_h, v_h) = l_h(v_h), \quad \forall v_h \in V_{0h}, \quad (4.11) \]

with \( a_h \) and \( l_h \) as defined above by (4.9) and (4.10), respectively. In Chapter 3, we already showed existence and uniqueness of a weak solution to the initial-boundary
value problem (1.1) – (1.3). However, our discrete variational problem (4.11) is of a different form. Thus, we have to investigate the stability of the space-time finite element scheme. More precisely, we will even show ellipticity of the bilinear form $a_h(\cdot,\cdot) : V_{0h} \times V_{0h} \to \mathbb{R}$ wrt the mesh-dependent norm
\[
\|v_h\|_{h}^2 := \sum_{E \in T_h} \left[\|\sqrt{\nu} v_h\|_{L^2(E)}^2 + \theta_E h_E \|\partial_t v_h\|_{L^2(E)}^2\right] + \frac{1}{2} \|v_h\|_{L^2(\Sigma_T)}^2. \tag{4.12}
\]
For the following derivations, we assume that our triangulation $T_h$ of $Q$ is shape regular such that the local approximation error estimates are available, [2, 4]. The triangulation $T_h$ of $Q$ is called quasi-uniform, if there exists a constant $c_u$ such that
\[
h_E \leq h \leq c_u h_E, \quad \text{for all } E \in T_h, \tag{4.13}
\]
where $h = \max_{E \in T_h} h_E$. Moreover, we introduce localised bounds for our coefficient function $\nu$, i.e.,
\[
\underline{\nu}_E \leq \nu(x,t) \leq \overline{\nu}_E, \quad \text{for almost all } (x,t) \in E \text{ and for all } E \in T_h, \tag{4.14}
\]
where $\underline{\nu}_E$ and $\overline{\nu}_E = \text{const.} > 0$. In the following, we need some inverse inequalities for functions from finite element spaces.

**Lemma 4.3.** There exist generic positive constants $c_{I,1}$ and $c_{I,2}$, such that
\[
\|v_h\|_{L^2(\partial E)} \leq c_{I,1} h_E^{-1/2} \|v_h\|_{L^2(E)}, \tag{4.15}
\]
\[
\|\nabla v_h\|_{L^2(E)} \leq c_{I,2} h_E^{-1} \|v_h\|_{L^2(E)} \tag{4.16}
\]
for all $v_h \in V_{0h}$ and for all $E \in T_h$.

**Proof.** For (4.15), see e.g. [22, 3], and for (4.16) see e.g. [2, 4, 5].

From $\nabla = (\nabla_x, \partial_t)^\top$ and (4.16), we can immediately deduce
\[
\|\partial_t v_h\|_{L^2(E)} \leq c_{I,2} h_E^{-1} \|v_h\|_{L^2(E)}. \tag{4.17}
\]
The above inequalities hold for the standard norms. However, we will also need such a result in some scaled norm.

**Lemma 4.4.** Let $\nu \in W_\infty^1(T_h)$ be a given uniformly positive function. Then
\[
\|v\|_{L^2_\nu(E)}^2 = \int_E \nu(x,t) |v(x,t)|^2 \, d(x,t)
\]
is a norm and there holds the inverse estimate
\[
\|\partial_t v_h\|_{L^2_\nu(E)} \leq \|\nabla v_h\|_{L^2_\nu(E)} \leq c_{I,\nu} h_E^{-1} \|v_h\|_{L^2_\nu(E)}, \tag{4.18}
\]
for all $v_h \in V_{0h}$ and for all $E \in T_h$. 

Proof. If \( \nu = \nu_E = const > 0 \) on \( E \), then (4.18) is nothing else than the classical inverse inequality (4.16). In general, we can at least assume that (4.14) holds. Using (4.14) and (4.16), we obtain
\[
\|\nabla v_h\|_{L^p_2(E)} \leq \sqrt{\nu_E} \|\nabla v_h\|_{L^2(E)} \leq \sqrt{\nu_E} c_{I,2} h_{E}^{-1} \|v_h\|_{L^2(E)} \\
\leq \left( \frac{\nu_E}{\nu_E} \right)^{1/2} c_{I,2} h_{E}^{-1} \|v_h\|_{L^2_2(E)} =: \text{const}
\]
It is clear that in \( 1 \leq \nu_E/\nu_E \) is close to 1 in practical applications.

Below, we will need the estimate
\[
\|\partial_t \partial_z v_h\|_{L^p_2(E)} \leq c_{I,\nu} h_{E}^{-1} \|\partial_z v_h\|_{L^2_2(E)}, \tag{4.19}
\]
which obviously holds for all \( v_h \in V_{0h} \) and for all \( E \in \mathcal{T}_h \). Moreover, we need the following inverse inequality.

**Lemma 4.5.** Let \( \nu \in W^1_\infty (\mathcal{T}_h) \) be a given uniformly positive function. Let \( W_h|_E := \{w_h : w_h = \nabla x v_h, v_h \in V_{0h}|_E\} \). Then there holds the inverse estimate
\[
\|\text{div}_x (\nu w_h)\|_{L^2(E)} \leq c_{I,3} h_{E}^{-1} \|\nu w_h\|_{L^2(E)}, \forall w_h \in W_h|_E, \tag{4.20}
\]
where \( c_{I,3} \) is a positive constant, independent of \( h_E \).

**Proof.** First, we know that \( V_{0h}|_E \) is a finite space spanned by the local shape functions \( \{p_0^{(i)}\}_{i \in \omega_E} \). Hence the space \( W_h|_E \) is also finite and spanned by the generating system \( \{\nabla_x p_0^{(i)}\}_{i \in \omega_E} \). Moreover, for a fixed \( \nu \), each product \( z_h := \nu w_h \) can be represented by means of a non-necessary unique linear combination \( \{\nu \nabla x p_0^{(i)}\}_{i \in \omega_E} \) on \( E \). We denote this space by \( Z_h(E) := \text{span}_{i \in \omega_E} \{\nu \nabla x p_0^{(i)}\} \). Using Cauchy’s inequality, we obtain
\[
\|\text{div}_x z_h\|^2_{L^2(E)} = \int_E |\text{div}_x z_h|^2 \, d(x, t) = \int_E \left| \sum_{i=1}^d \partial_{x_i} z_{h,i} \right|^2 \, d(x, t) \\
\leq d \int_E \left| \sum_{i=1}^d \partial_{x_i} z_{h,i} \right|^2 \, d(x, t) = d \sum_{i=1}^d \|\partial_{x_i} z_{h,i}\|_{L^2(E)},
\]
for all \( z_h \in Z_h(E) \). Now, by a simple scaling argument, we can estimate each element in the sum and obtain
\[
d \sum_{i=1}^d \|\partial_{x_i} z_{h,i}\|^2_{L^2(E)} \leq d \sum_{i=1}^d C^2 h_{E}^{-2} \|z_{h,i}\|^2_{L^2(E)} \\
= d C^2 h_{E}^{-2} \|z_h\|^2_{L^2(E)}.
\]
Indeed, transforming to the reference triangle, using the norm equivalence on finite

dimensional spaces, and transforming back to $E$, we obtain

$$
\| \partial_x z_{h,i} \|_{L^2(E)}^2 \leq \| \nabla z_{h,i} \|_{L^2(E)}^2 = \int_E |\nabla z_{h,i}|^2 \, d(x,t)
$$

$$
\leq c h_E^{d+1} \int_\Delta |\nabla \hat{z}_{h,i}|^2 d(\xi,\tau) \leq c h_E^{d+1} h_E^{-2} \int_\Delta |\nabla \hat{z}_{h,i}|^2 d(\xi,\tau)
$$

$$
\leq C h_E^{-2} \int_E |z_{h,i}|^2 \, d(x,t) = C h_E^{-2} \| z_{h,i} \|_{L^2(E)}.
$$

Taking the square root and setting $c\, I, 3 := C \sqrt{d}$ closes the proof.

Lemma 4.5 gives information how the two norms involved scale wrt the mesh-size $h_E$. However, the estimate (4.20) is not sharp wrt the constant.

**Lemma 4.6.** Let the assumptions of Lemma 4.5 hold. Then

$$
\| \text{div}_x (\nu w_h) \|_{L^2(E)} \leq c_{opt} \| \nu w_h \|_{L^2(E)}, \forall w_h \in W_h|_E
$$

(4.21)

with $c_{opt}^2 = \sup_{0 \neq z_h \in Z_h(E)} \frac{\| \text{div}_x (z_h) \|^2_{L^2(E)}}{\| z_h \|^2_{L^2(E)}}$.

**Proof.** From Lemma 4.5 we know that there must be a constant $c$ such that

$$
\| \text{div}_x (z_h) \|_{L^2(E)} \leq c \| z_h \|_{L^2(E)} \quad \forall z_h \in Z_h(E).
$$

With the assumption $z_h \neq 0$ we can rewrite the inequality above as

$$
\frac{\| \text{div}_x (z_h) \|^2_{L^2(E)}}{\| z_h \|^2_{L^2(E)}} \leq c^2.
$$

Now we immediately see that the optimal value for $c$ is nothing else than the supremum of the expression on left hand side, i.e.,

$$
c_{opt}^2 := \sup_{0 \neq z_h \in Z_h(E)} \frac{\| \text{div}_x (z_h) \|^2_{L^2(E)}}{\| z_h \|^2_{L^2(E)}}.
$$

What remains is to ensure that this supremum is finite. We start by identifying the kernel of $\| \nu \nabla \cdot \cdot \|_{L^2(E)}$. Using the notation of the proof of Lemma 4.5, we know

$$
0 = \| z_h \|_{L^2(E)} = \| \sum_{i \in \omega_E} z_i \phi^{(i)} \|_{L^2(E)}
$$

$$
= \| \sum_{i \in \omega_E} z_i \nu \nabla_x \phi^{(i)} \|_{L^2(E)} = \| \nu \nabla_x \sum_{i \in \omega_E} z_i \phi^{(i)} \|_{L^2(E)}.
$$
This identity holds if and only if $\nabla \varphi \equiv 0$, i.e., if $\varphi = \varphi(t)$. Now let $\varphi \in \ker \|\varphi\|_{L^2(E)}$. Then we immediately deduce that
\[
\|\ \nabla \varphi \|_{L^2(E)} = \|\varphi\|_{H^1_0(E)} = \|\varphi\|_{H^1_0(E)} = 0,
\]
i.e., $\ker \|\nabla \varphi \|_{L^2(E)} \subset \ker \|\varphi\|_{L^2(E)}$.

**Remark 4.7.** Note that the constant $c_{\text{opt}}$ in Lemma 4.6 is not only optimal but also computable. Let $\varphi_h \in Z_h(E)$, then by definition we have
\[
\varphi_h(x, t) = \sum_{j \in \mathcal{Q}_E} \tilde{z}_j \tilde{q}^{(j)}.
\]
Here we assume that the $\{\tilde{q}^{(j)}\}_{j \in \mathcal{Q}_E}$ form a basis of $Z_h(E)$. Moreover, we know
\[
\|\varphi_h\|_{L^2(E)}^2 = \langle \varphi_h, \varphi_h \rangle_{H^1_0(E)} \quad \text{and} \quad \|\nabla \varphi_h\|_{L^2(E)}^2 = \langle \nabla \varphi_h, \nabla \varphi_h \rangle_{H^1_0(E)}.
\]
As our space $Z_h(E)$ is finite, we can further rewrite these inner products. Let $y_h, \varphi_h \in Z_h(E)$, then
\[
\langle y_h, \varphi_h \rangle_{L^2(E)} = \sum_{i \in \mathcal{Q}_E} \langle y_h, \tilde{q}^{(i)} \rangle_{L^2(E)} \tilde{z}_i = \sum_{i, j \in \mathcal{Q}_E} \tilde{y}_j \langle \tilde{q}^{(j)}, \tilde{q}^{(i)} \rangle_{L^2(E)} \tilde{z}_i.
\]
This can be interpreted as
\[
\langle y_h, \varphi_h \rangle_{L^2(E)} = \langle M_h \tilde{y}, \tilde{z} \rangle_{\mathcal{Q}_E}, \quad \text{with} \quad (M_h)_{ij} = \langle \tilde{q}^{(j)}, \tilde{q}^{(i)} \rangle_{L^2(E)},
\]
where $\tilde{y}$ and $\tilde{z}$ are the vector of coefficients wrt the basis. By the same argument we obtain
\[
b(y_h, \varphi_h) = \langle B_h \tilde{y}, \tilde{z} \rangle_{\mathcal{Q}_E}, \quad \text{with} \quad (B_h)_{ij} = b(\tilde{q}^{(j)}, \tilde{q}^{(i)})_{L^2(E)}.
\]
Combining the above identities, we get with $N_E = |\mathcal{Q}_E|
\[
c_{\text{opt}}^2 = \sup_{0 \neq \varphi \in Z_h(E)} \|\nabla \varphi\|_{L^2(E)}^2 = \sup_{\tilde{z} \in \mathcal{Q}_E} \frac{(B_h \tilde{z}, \tilde{z})_{\mathcal{Q}_E}}{(M_h \tilde{z}, \tilde{z})_{\mathcal{Q}_E}}.
\]
Hence, $c_{\text{opt}}^2$ is the largest eigenvalue of the generalized eigenvalue problem
\[
B_h \tilde{z} = \lambda M_h \tilde{z}.
\]
Now, we are able to proof the following lemma.

**Lemma 4.8.** There exits a constant $\mu_a$ such that
\[
a_h(v_h, v_h) \geq \mu_a \|v_h\|_{L^2(E)}^2, \quad \forall v_h \in V_{0h},
\]
with $\mu_a = \min_{E \in \mathcal{T}_h} \left\{ 1 - c_{1, 3} \sqrt{\frac{\pi_E}{4 \theta_E}} \right\} \geq \frac{1}{2}$ for $\theta_E \leq \frac{h_E}{c_{1, 3} \theta_E}$, i.e., $\mu_a = \frac{1}{2}$ for $\theta_E = \frac{h_E}{c_{1, 3} \theta_E}$. 

Proof. We first do integration by parts at the last term, obtaining

\[ a_h(v_h, v_h) = \sum_{E \in \mathcal{T}_h} \int_{\partial E} \frac{1}{2} \partial_t (v_h^2) + \theta_E h_E (\partial_t v_h)^2 + \nu |\nabla v_h|^2 \, d(x,t) \]

\[ + \int_E \theta_E h_E \nu |\nabla v_h|^2 \, d(x,t) - \int_{\partial E} \theta_E h_E \nu \nabla v_h \cdot \partial_t v_h \, ds(x,t) \]

\[ = \sum_{E \in \mathcal{T}_h} \int_{\partial E} \frac{1}{2} \partial_t (v_h^2) \, d(x,t) + \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 + \int_E \nu |\nabla v_h|^2 \, d(x,t) \]

\[ - \int_E \theta_E h_E \, \text{div}_x (\nu \nabla v_h) \partial_t v_h \, d(x,t) \]

Now using Gauss’ theorem and the facts that \( v_h \) is continuous across the element boundary and that \( n_t = 0 \) on \( \Sigma \), we obtain

\[ a_h(v_h, v_h) = \sum_{E \in \mathcal{T}_h} \int_{\partial E} \frac{1}{2} v_h^2 n_t \, ds(x,t) + \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \]

\[ + \int_E \nu |\nabla v_h|^2 - \theta_E h_E \, \text{div}_x (\nu \nabla v_h) \partial_t v_h \, d(x,t) \]

\[ = \frac{1}{2} \left( \| v_h \|_{L^2(\Sigma_T)}^2 - \| v_h \|_{L^2(\Sigma_0)}^2 \right) + \sum_{E \in \mathcal{T}_h} \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \]

\[ + \int_E \nu |\nabla v_h|^2 - \theta_E h_E \, \text{div}_x (\nu \nabla v_h) \partial_t v_h \, d(x,t) \]

The first, second and third term already appear in the definition of our discrete norm (4.12). What remains is to estimate the last term. Using the Cauchy-Schwarz inequality, Lemma 4.5 and a scaled Young’s inequality, we arrive at the estimates

\[ |\theta_E h_E \int_E \text{div}_x (\nu \nabla v_h) \partial_t v_h \, d(x,t)| \leq \theta_E h_E \| \partial_t v_h \|_{L^2(E)} \| \text{div}_x (\nu \nabla v_h) \|_{L^2(E)} \]

\[ \leq \theta_E h_E c_{1,3} h_E^{-1} \| \nabla v_h \|_{L^2(E)} h_E^{-1/2} \| \partial_t v_h \|_{L^2(E)} \]

\[ \leq c_{1,3} \left( \frac{\varepsilon \nu_E \theta_E}{2 h_E} \| \nabla v_h \|_{L^2(E)}^2 + \frac{1}{2 \varepsilon} \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \right). \]

Using this estimate in the equality above and the fact that \( v_h = 0 \) on \( \Sigma_0 \), we get

\[ a_h(v_h, v_h) \geq \frac{1}{2} \| v_h \|_{L^2(\Sigma_T)}^2 \sum_{E \in \mathcal{T}_h} \left[ \left( 1 - \frac{c_{1,3} \varepsilon}{2 h_E} \right) \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \right] \]

\[ + \left( 1 - \frac{c_{1,3} \varepsilon}{2 h_E} \right) \| \nabla v_h \|_{L^2(E)}^2. \]
Now we choose $\varepsilon = \sqrt{h_E/(\theta_E v_E)}$ and obtain

$$
a_h(v_h, v_h) \geq \min_{E \in T_h} \left( 1 - c_{I,3} \sqrt{\frac{\theta_E v_E}{4 h_E}} \right)
\times \left( \sum_{E \in T_h} \left[ \|\nabla_x v_h\|_{L^2(E)}^2 + \theta_E h_E \|\partial_t v_h\|_{L^2(E)}^2 \right] + \frac{1}{2} \|v_h\|_{L^2(\Sigma_T)}^2 \right)
\geq \mu_a \|v_h\|_{L^2(T_h)}^2,
$$

which concludes the first part of the proof. The second assertion can be shown by a simple calculation, i.e.,

$$
1 - c_{I,3} \sqrt{\frac{\theta_E v_E}{4 h_E}} \geq \frac{1}{2} \Leftrightarrow c_{I,3} \sqrt{\frac{\theta_E v_E}{4 h_E}} \leq \frac{1}{2}
\Leftrightarrow c_{I,3}^2 \frac{\theta_E v_E}{h_E} \leq 1
\Leftrightarrow \theta_E \leq \frac{h_E}{v_E c_{I,3}^2}.
$$

\[\square\]

**Remark 4.9.** The above proof does hold for any polynomial degree $p \geq 1$ of $v_h$ and any fixed, uniformly positive $\nu \in L_\infty(Q)$. However, for the special case $p = 1$ and $\nu|_E = \text{const}$, the above proof is trivial, since

$$
\partial_t(\nabla_x v_h) \equiv 0 \quad \text{and} \quad \nu|_E \Delta_x v_h \equiv 0.
$$

Hence, there holds the identity

$$
a_h(v_h, v_h) = \sum_{E \in T_h} \int_E \partial_t v_h \nu \partial_t v_h + \theta_E h_E (\partial_t v_h)^2 + \nu |\nabla_x v_h|^2 \, d(x,t)
= \sum_{E \in T_h} \frac{1}{2} \int_{\partial E} \nu n_t \, ds(x,t) + \theta_E h_E \int_{L_2(E)} \|\partial_t v_h\|_{L^2(E)}^2 \|\nabla_x v_h\|_{L^2(E)}^2
= \|v_h\|_{L^2(T_h)}^2,
$$

i.e., $\mu_a = 1$. Moreover, we immediately deduce that for this special case, the choice of $\theta_E$ has no influence on the ellipticity of the space-time finite element method.

**Remark 4.10.** An alternative approach to the proof of Lemma 4.8 consists of not applying integration by parts on the last two terms of (4.9), but instead estimate

$$
\theta_E h_E \int_E \nu \nabla_x v_h \nabla_x (\partial_t v_h) \, d(x,t) \quad \text{and} \quad \theta_E h_E \int_{\partial E} \nu \nabla_x v_h n_x \partial_t v_h \, ds(x,t)
$$

separately.
Lemma 4.8 already ensures uniqueness of the finite element solution \( u_h \in V_{0h} \). Furthermore, since we use the same trial- and test-space \( V_{0h} \), and this space is finite dimensional, uniqueness implies existence of finite element solution \( u_h \in V_{0h} \) of (4.8).

For the special case of uniform meshes and uniform \( \theta \), i.e., \( h_E = h \) and \( \theta_E = \theta \) for all \( E \in T_h \), and and \( \nu \equiv 1 \), a proof for ellipticity with a mesh-independent constant was done by Langer et.al. [16]. For a second special case, where \( \theta_E \) vanishes, i.e., \( \theta_E = \theta = 0 \) for all \( E \in T_h \), Steinbach in [28] has shown existence and uniqueness of discrete version of (4.8). In addition, both papers include also a priori error estimates, where Steinbach’s estimate is based on a discrete inf-sup condition.

To show an a priori error estimate wrt the mesh dependent norm (4.12), we need where Steinbach’s estimate is based on a discrete inf-sup condition. Furthermore, since we use the same trial- and test-space

\[
\text{Lemma 4.8 already ensures uniqueness of the finite element solution } u_h \in V_{0h}. \quad \text{For the special case of uniform meshes and uniform } \theta, \text{ i.e., } h_E = h \text{ and } \theta_E = \theta \text{ for all } E \in T_h, \text{ and } \nu \equiv 1, \text{ a proof for ellipticity with a mesh-independent constant was done by Langer et.al. [16]. For a second special case, where } \theta_E \text{ vanishes, i.e., } \theta_E = \theta = 0 \text{ for all } E \in T_h, \text{ Steinbach in [28] has shown existence and uniqueness of discrete version of (4.8). In addition, both papers include also a priori error estimates, where Steinbach’s estimate is based on a discrete inf-sup condition.}
\]

To show an a priori error estimate wrt the mesh dependent norm (4.12), we need to show that our bilinear form \( a_h(\cdot, \cdot) \) is uniformly bounded on \( V_{0h,*} \times V_{0h} \), where \( V_{0h,*} = H_0^{1,0}(\Omega) \cap H^2(T_h) + V_{0h} \) with the norm

\[
\|v\|_{h,*}^2 = \|v\|_h^2 + \sum_{E \in T_h} \left( (\theta_E h_E)^{-1}\|v\|_{L^2(E)}^2 + \theta_E h_E \|v\|_{H^2(E)}^2 \right) \\
= \frac{1}{2}\|v\|_{L^2(\Sigma_T)}^2 + \sum_{E \in T_h} \left( \theta_E h_E \|\partial_t v\|_{L^2(E)}^2 + \|\nabla x v\|_{L^2(E)}^2 \right) + (\theta_E h_E)^{-1}\|v\|_{L^2(E)}^2 + \theta_E h_E \|v\|_{H^2(E)}^2 \tag{4.24}
\]

Moreover, we will make use of the following scaled trace inequality.

\[
\text{Lemma 4.11. There exists a positive constants } c_{T_{r}} > 0 \text{ such that } \\
\|v\|_{L^2(\partial E)}^2 \leq 2c_{T_{r}} h_E^{-1}\left(\|v\|_{L^2(E)}^2 + h_E^2 \|\nabla v\|_{L^2(E)}^2\right) \tag{4.25}\]

for all \( v \in H^1(E), \forall E \in T_h \).

\text{Proof. See e.g. [22].} \quad \square

\text{Lemma 4.12. The discrete bilinear form } a_h(\cdot, \cdot) \text{ is uniformly bounded on } V_{0h,*} \times V_{0h}, \text{ i.e., } \\
|a_h(u, v_h)| \leq \mu_h \|u\|_{h,*} \|v_h\|_h. \tag{4.26}
\]

where \( \mu_h = \max_{E \in T_h} \left\{ 2\left(1 + \theta_E h_E^{-1} c_{T}^2 \frac{\nu}{\nu_E}\right), 2c_{T}^2 \nu_E^2, 2 + c_{T,1}^2, 1 + (c_{T,1} \theta_E)^2 \right\}^{1/2} \text{ that is bounded provided that } \theta_E = O(h_E) \).

\text{Proof. We will estimate the bilinear form (4.9) term by term. For the first term, since } V_{0h} \subset H_{0,0}^{1,1}(\Omega), \text{ we can apply integration by parts and the Cauchy-Schwarz inequality, and obtain} \\
\sum_{E \in T_h} \int_E \partial_t u v_h \, dx \, dt = \sum_{E \in T_h} \left[ - \int_E u \partial_t v_h \, dx \, dt + \int_{\partial E} u n_t v_h \, ds(x,t) \right] \\
\leq \sum_{E \in T_h} \left[ ((\theta_E h_E)^{-1}\|u\|_{L^2(E)}^2)^{1/2} ((\theta_E h_E \|\partial_t v_h\|_{L^2(E)}^2)^{1/2} \right] + \left(\|u\|_{L^2(\Sigma_T)}^2\right)^{1/2} \left(\|v_h\|_{L^2(\Sigma_T)}^2\right)^{1/2}. \]
For the second and third term, applying the Cauchy-Schwarz inequality for each term of the sum yields

\[ \theta_E h_E \int_E \partial_t u \partial_t v_h \, d(x, t) \leq (\theta_E h_E \| \partial_t u \|_{L^2(E)}^2)^{1/2} \left( \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \right)^{1/2}, \]

\[ \int_E \nu \nabla_x u \nabla_x v_h \, d(x, t) \leq \left( \| \nabla_x u \|_{L^2(E)}^2 \right)^{1/2} \left( \| \nabla_x v_h \|_{L^2(E)}^2 \right)^{1/2}, \]

respectively. For the fourth term, we use again Cauchy-Schwarz' inequality, the inverse estimate (4.19), and obtain

\[ \theta_E h_E \int_E \nu \nabla_x u \nabla_x (\partial_t v_h) \, d(x, t) \leq \left( \| \nabla_x u \|_{L^2(E)}^2 \right)^{1/2} \left( \theta_E h_E \| \partial_t \nabla_x v_h \|_{L^2(E)}^2 \right)^{1/2} \]

\[ = \left( \| \nabla_x u \|_{L^2(E)}^2 \right)^{1/2} \left( \theta_E h_E \| \partial_t \nabla_x v_h \|_{L^2(E)}^2 \right)^{1/2} \sum_{i=1}^{d} \| \partial_t (\partial_x v_h) \|_{L^2(E)}^2 \]

\[ \leq \left( \| \nabla_x u \|_{L^2(E)}^2 \right)^{1/2} \left( \theta_E h_E \| \partial_t \nabla_x v_h \|_{L^2(E)}^2 \right)^{1/2} \sum_{i=1}^{d} c_{I, \nu} h_E^2 \| \partial_x v_h \|_{L^2(E)}^2 \]

\[ = \left( \| \nabla_x u \|_{L^2(E)}^2 \right)^{1/2} \left( \theta_E h_E \| \partial_t \nabla_x v_h \|_{L^2(E)}^2 \right)^{1/2} \sum_{i=1}^{d} c_{I, \nu} h_E^2 \| \partial_x v_h \|_{L^2(E)}^2 \]

For the last term, we apply Cauchy-Schwarz and the trace inequalities (4.17) and (4.25), and get

\[ \theta_E h_E \int_{\partial E} \nu \nabla_x u \nu \partial_t v_h \, ds(x, t) \leq \left( \theta_E h_E^2 \| \nabla_x u \|_{L^2(\partial E)}^2 \right)^{1/2} \left( \theta_E h_E^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \]

\[ \leq \left( \theta_E h_E^2 c_{2,1}^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \left( \theta_E h_E^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \]

\[ \leq \left( \theta_E h_E^2 c_{2,1}^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \left( \theta_E h_E^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \]

\[ \times \left( \theta_E h_E^2 c_{2,1}^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \left( \theta_E h_E^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \]

\[ \times \left( \theta_E h_E^2 c_{2,1}^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \left( \theta_E h_E^2 \| \partial_t v_h \|_{L^2(\partial E)}^2 \right)^{1/2} \]

Now we combine the above terms, apply Cauchy’s inequality and gather all similar
\[ |a_h(u, v_h)| \leq \left( \| u \|_{L^2(\Sigma_T)}^2 \right)^{1/2} \left( \| v_h \|_{L^2(\Sigma_T)}^2 \right)^{1/2} \]
\[ + \sum_{E \in T_h} \left[ ((\theta_E h_E)^{-1}\| u \|_{L^2(E)}^2)^{1/2} \left( \theta_E h_E \| \partial_t u \|_{L^2(E)} \right)^{1/2} \right. \]
\[ + (\theta_E h_E \| \partial_t u \|_{L^2(E)}^2)^{1/2} \left( (\theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2)^{1/2} + (\| \nabla_x u \|_{L^2(E)}^2)^{1/2} \right. \]
\[ + (\| \nabla_x u \|_{L^2(E)}^2)^{1/2} \left( (c_{I,1} \theta_E)^2 \| \nabla_x v_h \|_{L^2(E)}^2)^{1/2} \right. \]
\[ + \left. (2\theta_E c_{I,1} \frac{\nabla_E}{\nabla} h_E^{-1} \| \nabla_x u \|_{L^2(E)}^2 + 2 \nabla_E \theta_E h_E \| u \|_{H^2(E)}^2)^{1/2} \right] \]
\[ \times \left( c_{I,1} \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \right)^{1/2} \]
\[ \leq \left( \| u \|_{L^2(\Sigma_T)}^2 + \sum_{E \in T_h} \left[ \theta_E h_E \| \partial_t u \|_{L^2(E)}^2 + 2(1 + \theta_E c_{I,1} \frac{\nabla_E}{\nabla} h_E^{-1}) \| \nabla_x u \|_{L^2(E)}^2 \right. \]
\[ + (\theta_E h_E)^{-1} \| u \|_{L^2(E)}^2 + 2c_{I,1}^2 \| \nabla_E \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \right)^{1/2} \]
\[ \times \left( \| v_h \|_{L^2(\Sigma_T)}^2 + \sum_{E \in T_h} \left[ (2 + c_{I,1}^2) \theta_E h_E \| \partial_t v_h \|_{L^2(E)}^2 \right. \]
\[ + \left. (1 + (c_{I,1} \theta_E)^2) \| \nabla_x v_h \|_{L^2(E)}^2 \right)^{1/2} \]
\[ \leq \max_{E \in T_h} \left\{ 2(1 + \theta_E h_E^{-1} \frac{\nabla_E}{\nabla} c_{I,1}^2), 2c_{I,1}^2 \frac{\nabla_E}{\nabla} h_E, 2 + c_{I,1}^2, 1 + (c_{I,1} \theta_E)^2 \right\}^{1/2} \left\| u \right\|_{H^s} \left\| v_h \right\|_{H^s}. \]

Choosing now \( \theta_E = \mathcal{O}(h_E) \) ensures the boundedness of the constant \( \mu_h \). \( \square \)

**Remark 4.13.** Choosing \( \theta_E \) as in Lemma 4.8, i.e., \( \theta_E = h_E / (c_{I,3} \nabla_E) \), we obtain \( \mu_a = 1/2 \) and \( \mu_b = \max_{E \in T_h} \left\{ 2(1 + \frac{\nabla_E}{\nabla} c_{I,3}^2), 2c_{I,3}^2 \frac{\nabla_E}{\nabla} h_E, 2 + c_{I,3}^2, 1 + (c_{I,3} \theta_E)^2 \right\}^{1/2} \).

**Remark 4.14.** As in Remark 4.9, we can provide a simplified estimate for the special case \( p = 1 \) and \( \nu_E = \nu_E = \text{const.} \) The first three terms can be estimated as in the above proof. The fourth term completely vanishes, since \( \nabla_x(\partial_t v_h) = 0 \). For the fifth term, we use the fact that \( \partial_t v_h = \text{const.} \), Gauss’ theorem and the Cauchy-Schwarz
We immediately deduce that this new constant \( \tilde{\mu}_b \) is some positive integer. For further details on such spaces, we refer to [5, 22].
Lemma 4.15. Let \( s \) and \( k \) be positive integers with \( s \in [2, p + 1] \) and \( k > (d + 1)/2 \), respectively. Let \( v \in V_0 \cap H^k(Q) \cap H^s(T_h) \). Then there exists an interpolation operator \( \Pi_h \), mapping from \( V_0 \cap H^k(Q) \) to \( V_0h \), such that

\[
\|v - \Pi_h v\|_{L^2(E)} \leq C h_{E}^{s+1} |v|_{H^s(E)}, \tag{4.29}
\]

\[
\|\nabla (v - \Pi_h v)\|_{L^2(E)} \leq C h_{E}^{s} |v|_{H^s(E)}, \tag{4.30}
\]

\[
|v - \Pi_h v|_{H^2(E)} \leq C h_{E}^{s-1} |v|_{H^s(E)}, \tag{4.31}
\]

where \( C \) is some generic constant independent of \( v \). Here, \( p \) denotes the polynomial degree of the finite element basis functions.

Proof. See e.g. [3, Theorem 4.4.4] or [4, Theorem 3.1.6].

Lemma 4.16. Let the assumptions of Lemma 4.15 hold. Then the following interpolation error estimates hold:

\[
\|v - \Pi_h v\|_{L^2(\Sigma_T)} \leq c_1 \left( \sum_{E \in T_h, \partial E \cap \Sigma_T \neq \emptyset} h_{E}^{2s-1} |v|_{H^s(E)}^2 \right)^{1/2}, \tag{4.32}
\]

\[
\|v - \Pi_h v\|_{H^1(E)} \leq c_2 \left( \sum_{E \in T_h} h_{E}^{2(s-1)} |v|_{H^s(E)}^2 \right)^{1/2}, \tag{4.33}
\]

\[
|v - \Pi_h v|_{H^2(E)} \leq c_3 \left( \sum_{E \in T_h} h_{E}^{2(s-1)} |v|_{H^s(E)}^2 \right)^{1/2}. \tag{4.34}
\]

The constants \( c_1, c_2, c_3 \) do not depend on \( h_E \) or \( v \), provided that \( \theta_E = O(h_E) \).

Proof. We start with the first estimate (4.32). We use the scaled trace inequality (4.25), and the interpolation error estimates (4.29) and (4.30), obtaining

\[
\|v - \Pi_h v\|^2_{L^2(\Sigma_T)} = \sum_{E \in T_h, \partial E \cap \Sigma_T \neq \emptyset} \|v - \Pi_h v\|^2_{L^2(\partial E \cap \Sigma_T)} \leq \sum_{E \in T_h, \partial E \cap \Sigma_T \neq \emptyset} \|v - \Pi_h v\|^2_{L^2(\partial E)}
\]

\[
\leq \sum_{E \in T_h, \partial E \cap \Sigma_T \neq \emptyset} \left[ 2c_T^2 h_{E}^{-1} \|v - \Pi_h v\|^2_{L^2(E)} + h_{E}^2 \|\nabla(v - \Pi_h v)\|^2_{L^2(E)} \right]
\]

\[
\leq c_T^2 \sum_{E \in T_h, \partial E \cap \Sigma_T \neq \emptyset} \left[ C h_{E}^{2s-1} |v|_{H^s(E)}^2 + C h_{E}^{2s-1} |v|_{H^s(E)}^2 \right]
\]

\[
\leq c_T^2 C \sum_{E \in T_h, \partial E \cap \Sigma_T \neq \emptyset} [h_{E}^{2s-1} |v|_{H^s(E)}].
\]

For (4.33), we use definition (4.12), assumption (4.14), the interpolation error estimate
Remark 4.17. The strong assumption $v \in H^k(\mathcal{Q})$ with $k > (d+1)/2$ is needed for the interpolation error estimates for the Lagrange interpolation operator. However, in practical application this requirement is most likely never met. However, in such a practical application, the space-time cylinder $\mathcal{Q} = \bigcup_{i=1}^{M} \mathcal{Q}_i$ can be split into subdomains $\mathcal{Q}_i$, which correspond e.g. to different materials. On each such subdomain $\mathcal{Q}_i$, we can assume some regularity for $v \in H^s(\mathcal{T}(\mathcal{Q}_i)) := \{ v \in L^2(\mathcal{Q}_i) : v|_{\mathcal{Q}_i} \in H^s(\mathcal{Q}_i), \text{ for all } i = 1, \ldots, M \}$ and the diffusion coefficient $\nu \in W^1_\infty(\mathcal{T}(\mathcal{Q}_i)) := \{ v \in L^\infty(\mathcal{Q}_i) : v|_{\mathcal{Q}_i} \in W^1_\infty(\mathcal{Q}_i), \text{ for all } i = 1, \ldots, M \}$. For a similar case, Duan et.al. [7] have shown an interpolation error estimate of the form

$$\|\nabla(v - I_h v)\|_{L^2(\mathcal{T})} \leq C h^{s-1} \sum_{i=1}^{M} \|v\|_{H^s(\mathcal{Q}_i)}.$$  

Now we can formulate the following a priori estimate for the error.

Theorem 4.18. Let $s$ and $k$ be positive integers with $s \in [2, p + 1]$ and $k > (d+1)/2$. Furthermore, let $u \in V_0 \cap H^k(\mathcal{Q}) \cap H^s(\mathcal{T}_h)$ be the exact solution, and $u_h \in V_0 h$ the solution of the finite element scheme (4.11). Then there holds the a priori error estimate

$$\|u - u_h\|_h \leq c \left( \sum_{E \in \mathcal{T}_h} h_E^{2(s-1)} \|u\|^2_{H^s(E)} \right)^{1/2}.$$  

(4.35)
Proof. First, we know from the consistency identity (4.8) that \( a_h(u, v_h) = l_h(v_h) \), and, since \( u_h \) is the approximate solution of (4.11), that \( a_h(u_h, v_h) = l_h(v_h) \). Hence we have Galerkin orthogonality for our bilinear form \( a_h(\cdot, \cdot) \), i.e.,

\[
a_h(u - u_h, v_h) = 0, \quad \forall v_h \in V_0h. \tag{4.36}
\]

We start with the triangle inequality for the discretization error, i.e.,

\[
\|u - u_h\|_h \leq \|u - \Pi_h u\|_h + \|\Pi_h u - u_h\|_h.
\]

We continue by estimating the second term. Using the ellipticity proved in Lemma 4.8, the Galerkin orthogonality and the generalized boundedness from Lemma 4.12, we obtain

\[
\mu_a \|\Pi_h u - u_h\|_h^2 \leq a_h(\Pi_h u - u_h, \Pi_h u - u_h) = a_h(\Pi_h u - u, \Pi_h u - u_h)
\]

\[
\leq \mu_b \|\Pi_h u - u\|_{h, * \Pi_h u - u_h}.
\]

We insert this estimate in the triangle inequality above, use the interpolation error estimates (4.33) and (4.34), and obtain

\[
\|u - u_h\|_h \leq \|u - \Pi_h u\|_h + \frac{\mu_b}{\mu_a} \|\Pi_h u - u\|_{h, *}
\]

\[
\leq c_2 \left( \sum_{E \in T_h} h_E^{2(s-1)} |u|_{H^s(E)}^2 \right)^{1/2} + c_3 \frac{\mu_b}{\mu_a} \left( \sum_{E \in T_h} h_E^{2(s-1)} |u|_{H^s(E)}^2 \right)^{1/2}
\]

\[
\leq \left( c_2 + c_3 \frac{\mu_b}{\mu_a} \right) \left( \sum_{E \in T_h} h_E^{2(s-1)} |u|_{H^s(E)}^2 \right)^{1/2},
\]

which proves the estimate (4.35) with \( c = c_2 + c_3 (\mu_b / \mu_a) \).

The Finite Element Method

Now we proceed with solving the discrete variational problem (4.11) that is nothing but a huge system of linear algebraic equations. Indeed, let \( \{ p^{(i)} : i \in \mathcal{I}_h \} \) be some basis of \( V_0h \), where \( \mathcal{I}_h \) is some index set, which we will specify later. Then we can express the approximate solution \( u_h \) in terms of this basis, i.e. \( u_h(x, t) = \sum_{i \in \mathcal{I}_h} u_i p^{(i)}(x, t) \). Furthermore, each basis function is a valid test function. Thus, we obtain \( \mathcal{N}_h \) equations from (4.11),

\[
a_h(u_h, p^{(i)}) = l_h(p^{(i)}), \quad \text{for all } i \in \mathcal{I}_h, \tag{4.37}
\]

where \( N_h = |\mathcal{I}_h| \) is the dimension of \( V_0h \). Now we replace \( u_h \) by its basis representation, which yields

\[
\sum_{j \in \mathcal{I}_h} u_j a_h(p^{(j)}, p^{(i)}) = l_h(p^{(i)}), \quad \text{for all } i \in \mathcal{I}_h. \tag{4.38}
\]

We can rewrite this system in terms of a system of linear algebraic equations

\[
\textbf{K}_h \textbf{u}_h = \textbf{f}_h, \tag{4.39}
\]
where \( K_h = (K_{ij}) \), \( u_h = (u_i) \) and \( f_h = (f_i) \). The system matrix is non-symmetric, but positive definite due to Lemma 4.8. Indeed,

\[
(K_h v_h, v_h) = a_h(v_h, v_h) \geq \mu a \|v_h\|_h^2 > 0
\]

for all \( V_{0h} \ni v_h \leftrightarrow v_h \in \mathbb{R}^{N_h} : v_h \neq 0 \). The linear system (4.39) can be solved efficiently and most important in parallel by either a sparse direct solver (e.g. sparse LU-factorisation) or an iterative solver (e.g. preconditioned GMRES).

But how to construct such a basis \( \{p^{(i)}\} \) of \( \bar{V}_{0h} \)? We need again the regular triangulation \( T_h \) of our space-time domain \( Q \), which we already introduced in (4.1). We now define \( \text{shape functions} \) and the corresponding function set \( \mathcal{F}(E) \). This function set is either a subspace or equal to the following function spaces

\[
P_k := \left\{ \sum_{|\alpha| \leq k} c_{\alpha} x^{\alpha} : c_{\alpha} \in \mathbb{R} \right\},
\]

\[
Q_k := \left\{ \sum_{\alpha_i \leq k} c_{\alpha} x^{\alpha} : i = 1, \ldots, d + 1, c_{\alpha} \in \mathbb{R} \right\}.
\]

Furthermore, we need some degrees of freedom, denoted by \( l^{(E,\alpha)} \), which are functionals in the dual space \( \mathcal{F}(E)^* \) of \( \mathcal{F}(E) \). If these functionals span the whole dual space, i.e., they are a basis, or, equivalently, have the interpolation property

\[
l^{(E,\alpha)}(v) = c_{\alpha}, \text{ for } v \in \mathcal{F}(E),
\]

then we can uniquely determine all coefficients of \( v \in \mathcal{F}(E) \) on an element \( E \). A particular example for a degree of freedom is the point evaluation, i.e., let \( v \in \mathcal{F}(E) \), then

\[
l^{(E,\alpha)}(v) = v(x^{(E,\alpha)}), \alpha \in \mathcal{A}_E,
\]

where \( x^{(E,\alpha)} \in E \) is called node and \( \mathcal{A}_E \) is a set of local indices. Combining all of the above, we can now introduce a formal definition for a finite element.

**Definition 4.19 (Finite Element [2]).** Let \( T_h \) be an admissible triangulation, i.e., for \( E, E' \in T_h \),

\[
E \cap E' = \begin{cases} 
\emptyset, & \text{common vertex}, \\
\text{common edge}, & \text{common face (3D)}. 
\end{cases}
\]

A triple \( (E, \mathcal{F}(E), \{l^{(E,\alpha)}\}) \) is called a finite element with the following properties:

(i) \( E \) is a polyhedron in \( \mathbb{R}^{d+1} \).

(ii) \( \mathcal{F}(E) \) is a subspace of \( C(E) \) with finite dimension \( p \).

(iii) \( \{l^{(E,\alpha)}\} \) is a set of \( p \) linear independent functionals over \( \mathcal{F}(E) \) which uniquely determine any \( \phi \in \mathcal{F}(E) \).
Hence, we have local nodal basis of shape functions, i.e.,
\[
\{ p^{(E,\alpha)} : p^{(E,\alpha)} \in \mathcal{F}(E), \alpha \in \mathcal{A}_E \},
\]  
with the property \( l^{(E,\alpha)}(p^{(E,\beta)}) = \delta_{\alpha\beta} \).

Now we define the global set of nodes \( \{ x^{(i)} : i \in \mathcal{I}_h \} \), and if \( x^{(i)} \in E \), then \( x^{(i)} = x^{(E,\alpha)} \) for some \( \alpha \). Furthermore, we need a global set of degrees of freedom \( \{ l^{(i)} : i \in \mathcal{I}_h \} \), where \( l^{(i)} = l^{(E,\alpha)} \) on \( E \), and a global nodal basis \( \{ p^{(i)} : i \in \mathcal{I}_h \} \), with \( l^{(i)}(p^{(j)}) = \delta_{ij} \) and \( p^{(i)} = p^{(E,\alpha)} \) on \( E \). Then this global nodal basis spans our discrete function space \( V_{0h} = \text{span}\{ p^{(i)} : i \in \mathcal{I}_h \} \).

For the rest of this thesis, we will restrict ourselves to the polynomial space \( \mathbb{P}_p \) on \( d + 1 \)-simplices, e.g., triangles in 2D and tetrahedrons in 3D. The degree of freedom is the point evaluation. This class of finite elements belongs to the class of Lagrange finite elements.

(a) 2D
(b) 3D

Figure 4.1: The linear and quadratic Lagrange finite element for different dimensions. The black dot denotes the point evaluation in the vertex.

To efficiently compute the entries in \( K_h \) and \( f_h \), we observe that the nodal basis functions \( p^{(i)} \) have only local support, which will result in a sparse matrix \( K_h \). Therefore,
we can write each entry as

\[
(K_h)_{ij} = \begin{cases} 
0, & \text{if } B_{ij} = B_i \cap B_j = \emptyset, \\
\sum_{E \in B_{ij}} a_{h,E}(p^{(j)}, p^{(i)}), & \text{else}
\end{cases},
\]

\[
(f_h)_i = \sum_{E \in B_i} l_{h,E}(p^{(i)}),
\]

where \( B_i = \{ E \in T_h : x^{(i)} \in \overline{E} \} \) is the neighbourhood of a node \( x^{(i)} \) and

\[
a_{h,E}(u_h, v_h) := \int_E \partial_t u_h v_h + \theta_{E} h_E \partial_t u_h \partial_t v_h \, d(x, t) \\
+ \int_E \nu \nabla_x u_h \cdot \nabla_x v_h + \theta_{E} h_E \nu \nabla_x u_h \cdot \nabla_x (\partial_t v_h) \, d(x, t) \\
- \int_{\partial E} \theta_{E} h_E \nu \nabla_x u_h \cdot n_x v_h \, ds(x, t),
\]

\[
l_{h,E}(v_h) := \int_E f(v_h + \theta_{E} h_E \partial_t v_h) \, d(x, t),
\]

are the integrals over one element.

In order to compute the entries of \( K_h \), we will assemble the stiffness matrix \( K_h \) and the load vector \( f_h \) element-wise, i.e., on each element \( E \), we have to identify \( i \leftrightarrow \alpha \) and \( j \leftrightarrow \beta \), so we obtain the local element matrix \( K^{(E)}_h \) and the element load vector \( f^{(E)}_h \), with

\[
(K^{(E)}_h)_{\alpha \beta} = a_{h,E}(p^{(E,\beta)}, p^{(E,\alpha)}), \quad \text{and} \quad (f^{(E)}_h)_\alpha = l_{h,E}(p^{(E,\alpha)}).
\]

In order to avoid the computation of the coefficients of \( p^{(E,\alpha)} \) on each element, we will instead transform the arbitrary element \( E \) to a canonical element, the so-called reference element \( \Delta \), with

\[
\Delta := \{ (\xi_1, \ldots, \xi_{d+1}) : \sum_{i=1}^{d+1} \xi_i \leq 1 \land \xi_i \geq 0, \text{ for } i = 1, \ldots, d + 1 \}.
\]

For our finite elements, it is sufficient to do this transformation via an affine mapping \( X_E \), which is defined as

\[
\xi \mapsto X_E(\xi) := x^{(1)} - x^{(E,1)} + \left( \underbrace{x^{(E,2)} - x^{(E,1)}}_{j=2} + \ldots + \underbrace{x^{(E,d+1)} - x^{(E,1)}}_{j=d+1} \right) \xi,
\]

where \( x^{(E,\alpha)} \) are the vertices of the element \( E \in T_h \). For instance, for \( d = 1 \), the shape functions \( p^{(\alpha)} \) can be easily computed on the reference element \( \Delta \), i.e.,

\[
p^{(1)}(\xi) = 1 - \xi_1 - \xi_2, \quad p^{(2)} = \xi_1, \quad \text{and} \quad p^{(3)} = \xi_2.
\]
The shape functions on an arbitrary element $E$ are obtained via the inverse mapping $X_{E}^{-1}$, i.e.,

$$p^{(E,a)} = p^{(a)} \circ X_{E}^{-1}. \quad (4.46)$$

Hence, we can compute the element matrices $K_{h}^{(E)}$ and element load vectors $f_{h}^{(E)}$ by transforming the integrals to the reference element $\Delta$. The transformed integrals can now be approximated by some quadrature rule. Let $f : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$, then

$$\int_{\Delta} f(\xi) \, d\xi \approx \sum_{i=1}^{M} \omega_{i} f(\xi_{i}),$$

where $\omega_{i}$ are the quadrature weights and $\xi_{i}$ are the quadrature points, for $i = 1, \ldots, M$. For some quadrature rules, see e.g. [29]. If we now perform these calculations for each element $E$, and add the entries of $K_{h}^{(E)}$ and $f_{h}^{(E)}$ to the corresponding entries of $K_{h}$ and $f_{h}$, respectively, we have fully assembled our linear system (4.39). Note that as our bilinear form $a_{h}(. , .)$ is non-symmetric, the stiffness matrix $K_{h}$ is also non-symmetric. However, so far, we do not have incorporated the initial- and boundary-conditions. First of all, we deduce that the initial condition (1.3) can be seen as a Dirichlet boundary condition for the space time cylinder $Q$. For homogeneous initial and boundary values, this incorporation can be easily achieved. We first identify all vertices which are on the Dirichlet boundary. Let us denote the set of indices of such vertices by $\bar{I}_{h}$. Then, for each $i \in \bar{I}_{h}$, we set

$$f_{i} = 0, \quad K_{ij} = 0 \text{ for } j \in \bar{I}_{h} \setminus \{i\}, \quad \text{and} \quad K_{ii} = 1.$$

For further details on the incorporation of other types of boundary conditions, we refer to e.g. [13].

### 4.2 A Space-time Finite Element Method based on stabilizing bubble functions

In this section, we will briefly describe an alternative space-time finite element scheme, which was first introduced by Touloupoulos in [32]. For this section, we will restrict ourselves to a constant reluctivity $\nu = \text{const} > 0$. Moreover, let $V_{0}^{1}_{\nu}$ be the conforming finite element space of piecewise linear functions (c.f. Section 4.1). The starting point for this scheme is the following space-time finite element scheme for the model problem (1.1) – (1.3): find $u_{h} \in V_{0}^{1}_{\nu}$, s.t.

$$\int_{Q} \partial_{t} u_{h} v_{h} + \nu \nabla_{x} u_{h} \nabla_{x} v_{h} \, dxdt = \int_{Q} f_{h} v_{h} \, dxdt, \quad \text{for all } v_{h} \in V_{0}^{1}_{\nu}. \quad (4.47)$$

However, for small values of $\nu$, this scheme does not perform well [32]. To regain stability, we enrich the space $V_{0}^{1}_{\nu}$ by so-called bubble functions, and obtain the larger space-time finite element space

$$V_{h,b} := \{v_{h} \in H_{0,\nu}^{1}(Q) : v_{h}|_{E} \in P_{1}(E) \oplus V_{b}(E), \quad \text{for all } E \in T_{h}\}, \quad (4.48)$$
where \( V_b(E) = V_b|_E \) for all \( E \in \mathcal{T}_h \). Here, \( V_b \) denotes the space of bubble functions, which vanish entirely on the boundary of an element \( E \in \mathcal{T}_h \) and have exactly one degree of freedom for each element. For our case of piecewise linear basis functions, this space consists of cubic functions, i.e., \( V_b := \{ v^b : v^b|_E = c_E p^1(E) p^2(E) p^2(E) \} \), for all \( E \in \mathcal{T}_h \). Based on (4.48), we can split every function \( v_h \in V_{h,b} \) into a linear part \( v^1_h \in V^1_h \) and a bubble function \( v^b_h \in V_b(E) \), i.e., \( v_h = v^1_h + v^b_h \). Now we can introduce a space-time finite element method, where the discrete problem reads as follows: find \( u_h \in V_{h,b} \), s.t.

\[
\hat{a}_h(u_h, v_h) = l(v_h), \quad \text{for all } v_h \in V_{h,b},
\]

with

\[
\hat{a}_h(u_h, v_h) := \int_Q \partial_t u_h v_h + \nu \nabla_x u_h \nabla_x v_h \, dx \, dt + \theta_h \int_Q \partial_t v^b_h \partial_t u^b_h \, dx \, dt,
\]

\[
l_h(v_h) = \int_Q f v_h \, dx \, dt.
\]

The bilinear form \( \hat{a}_h \) is bounded and coercive in some norm, see [32] for further details. Moreover, Touloupoulos was able to derive a priori error estimates in the norm

\[
\| u - u_h \|^2_{h,b} := \nu \| \nabla_x u \|_{L^2(Q)} + \theta_h \| \partial_t u \|_{L^2(Q)} + \frac{1}{2} \| u \|_{L^2(\Sigma_T)},
\]

which we summarize in the next theorem.

**Theorem 4.20.** Let \( u \in H^{1,1}_w(Q) \cap H^s(Q) \), with \( s \geq 2 \), be the exact solution and \( u_h \in V_{h,b} \) be the solution of the finite element scheme (4.49). Then there holds the a priori estimate

\[
\| u - u_h \|_{h,b} \leq c h \| u \|_{H^s(Q)}, \quad \text{for } \theta = \mathcal{O}(h) \text{ and } h < \nu.
\]
Chapter 5

Numerical Results

In this chapter, we present and analyze the results of the numerical tests we performed. We implemented the FEM with MFEM\(^1\), a C++ library for finite elements. The resulting linear systems were solved by means of the solver library HYPRE\(^2\), where we used the GMRES method, with BoomerAMG as preconditioner. This library is already fully parallelized. Hence, we were able to reach a high number of dofs. However, we observed very poor performance of the iterative solver for the 1+1-dimensional case. Therefore, we decided to solve the linear system by means of a direct solver. In particular, we used the MUMPS direct solver, which we included via the PETSc suite\(^3\).

Any visualization of the approximate solution is done via GLVis\(^4\). The initial meshing for \(d = 1\) and \(d = 2\) was done via NETGEN \([25]\), whereas the initial mesh for \(d = 3\) was included by MFEM. The finer meshes were then obtained by subsequent uniform refinements.

To obtain a convergence result, we computed the error in the \(L^2\)- and the \(\| \cdot \|_h\)-norm, i.e., let \(u\) be the exact solution of (4.8) and let \(u_h\) be the approximate solution of (4.11), then we computed \(\|u - u_h\|_{L^2(Q)}\) and \(\|u - u_h\|_h\), respectively.

**Example 1.** For this example we consider the \(d\)-dimensional unit cube as our spatial domain, and \((0, 1)\) as our time interval, i.e., \(Q = (0, 1)^{d+1}\). We require \(\nu \equiv 1\) and choose the function

\[
u(x, t) = \sin(x_1 \pi) \cdots \sin(x_d \pi) \sin(t \pi)
\]

as our exact solution of the model problem (1.1) – (1.3), where the right hand side is computed accordingly. As the exact solution is very smooth, we expect optimal convergence rates for this example.

**Example 2.** For this example, we consider only \(d = 1\) and \(d = 2\). Our space-time cylinder is again the unit-cube, i.e., \(Q = (0, 1)^2\) and \(Q = (0, 1)^3\), respectively. We

\(^{1}\)http://www.mfem.org


\(^{3}\)https://www.mcs.anl.gov/petsc/

\(^{4}\)http://glvis.org/
choose the right hand side as in Example 1. However, in contrast to Example 1, we now allow discontinuities for $\nu$. The finite element mesh is generated such that the discontinuities are only on the element boundary, i.e., there are no jumps in $\nu$ in the interior of an element $E$. We treated three cases of discontinuities:

(a) jumps only in the spatial directions, i.e., $\nu(x,t) = \nu(x)$ for $(x,t) \in Q$, see Figure 5.2a and Figure 5.4a

(b) jumps in space and time, see Figure 5.2b and Figure 5.4b

and

(c) jumps in space and time, resulting in a non-smooth space-time interface, see Figure 5.2c and Figure 5.4c.

For this problem, we do not know the exact solution. Hence, we have to work around this fact. An alternative way to obtain convergence rates without the knowledge of the exact solution is to perform as many uniform refinements as possible and solve the problem on this very fine mesh. Then we consider this solution $u^{(fms)}$ on the finest mesh as the "exact" solution and use it to compute error rates. However, we have to treat the obtained rates with care, as the approximate solutions tend to naturally converge to the finest mesh solution.

Another challenge that arises for this example is the loss of regularity in the exact solution, which we are not able to quantify. This means we are not able to guarantee that the a priori error estimate (4.35) holds. We therefore do not expect to obtain the optimal rates.

**Example 3.** Let once more $Q = (0,1)^{d+1}$. As exact solution, we choose the function

$$u(x,t) = |t - 0.5|^\lambda \sin(x_1\pi) \cdots \sin(x_d\pi),$$

with $\lambda > 0$, and we require $\nu \equiv 1$. The choice of $\lambda$ now determines the regularity of the solution. If we choose $\lambda$ small enough, e.g., $\lambda = 1.45$, then $u \notin H^2(Q)$, but $u \in H^{1.95-\varepsilon}(Q)$, with $\varepsilon > 0$.

As for the choice of the parameter $\theta_E$, we have seen in Remark 4.9 and Remark 4.14 that for $p = 1$, the choice of $\theta$ does not influence the stability of the method. It still influences the consistency error. For higher polynomial degree, e.g., $p = 2$ or $p = 3$, the choice $\theta_E = O(h_E)$ is essential to ensure the stability (c.f. e.g. Table 5.3 or Table 5.20). For the time being, we do not compute the optimal constant given in Lemma 4.8 but instead use $\theta_E = c h_E$, which turned out to be sufficient for most of our experiments, provided that the constant $c$ is sufficiently small. This $c$ can be seen as an approximation to $1/(c_1^2 \nu_E)$ from the inverse inequality (4.20).
## 5.1 1D Examples

We start with the simplest case, i.e., \(d = 1\), which results in a two-dimensional space-time finite element method. As mentioned in the beginning of this chapter, the performance of the preconditioned GMRES method was really poor. Thus, all examples in this section are solved by means of the direct solver MUMPS. For this case, we did not include any comparison in solving time, as the sparsity pattern of the system matrices for different \(\theta_E\) does not change, hence the solving time is almost the same for each case.

### Example 1

For the problem with a high regularity solution, we deduce that the error rates in the \(\| \cdot \|_h\)-norm are indeed optimal. We also see that for \(p = 1\), the stability of the method is not affected by choosing \(\theta_E = 1 = \mathcal{O}(1)\). However, the absolute value of the error is influenced by the choice of \(\theta_E\). As for the choice of \(c\), for \(p = 1, 2, c = 1\) was sufficient, whereas for piecewise cubic functions, we chose \(c = 10^{-5}\). For higher polynomial degree, i.e., \(p = 2\) or \(p = 3\), we note that the rate of convergence is slowed down or no convergence is attained, respectively (c.f. Table 5.2 and Table 5.3). This confirms the theoretical results of Chapter 4.

<table>
<thead>
<tr>
<th>(d=5)</th>
<th>(\theta_E=0)</th>
<th>(\theta_E=h_E)</th>
<th>(\theta=1)</th>
<th>(|u−u_0|_h)</th>
<th>(|u−u_0|_h)</th>
<th>(|u−u_0|_h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1.13 × 10^{-4}</td>
<td>9.42 × 10^{-4}</td>
<td>289</td>
<td>1.05 × 10^{-1}</td>
<td>1.56 × 10^{-1}</td>
<td>3.75 × 10^{-1}</td>
</tr>
<tr>
<td>1.0</td>
<td>2.95 × 10^{-3}</td>
<td>5.12 × 10^{-2}</td>
<td>1909</td>
<td>7.70 × 10^{-2}</td>
<td>7.74 × 10^{-2}</td>
<td>1.96 × 10^{-1}</td>
</tr>
<tr>
<td>2.0</td>
<td>7.49 × 10^{-4}</td>
<td>2.67 × 10^{-3}</td>
<td>4225</td>
<td>3.85 × 10^{-3}</td>
<td>3.96 × 10^{-3}</td>
<td>9.95 × 10^{-2}</td>
</tr>
<tr>
<td>4.0</td>
<td>1.76 × 10^{-4}</td>
<td>1.37 × 10^{-3}</td>
<td>16641</td>
<td>1.93 × 10^{-3}</td>
<td>1.93 × 10^{-3}</td>
<td>5.02 × 10^{-2}</td>
</tr>
<tr>
<td>8.0</td>
<td>4.49 × 10^{-5}</td>
<td>6.43 × 10^{-4}</td>
<td>60649</td>
<td>9.64 × 10^{-4}</td>
<td>9.64 × 10^{-4}</td>
<td>2.52 × 10^{-2}</td>
</tr>
<tr>
<td>16</td>
<td>1.11 × 10^{-5}</td>
<td>3.49 × 10^{-4}</td>
<td>265169</td>
<td>4.82 × 10^{-4}</td>
<td>4.82 × 10^{-4}</td>
<td>1.26 × 10^{-3}</td>
</tr>
<tr>
<td>32</td>
<td>1.99 × 10^{-6}</td>
<td>1.75 × 10^{-5}</td>
<td>1050625</td>
<td>2.41 × 10^{-5}</td>
<td>2.41 × 10^{-5}</td>
<td>6.33 × 10^{-2}</td>
</tr>
<tr>
<td>64</td>
<td>6.94 × 10^{-7}</td>
<td>8.78 × 10^{-6}</td>
<td>4158401</td>
<td>1.20 × 10^{-5}</td>
<td>1.20 × 10^{-5}</td>
<td>3.17 × 10^{-2}</td>
</tr>
<tr>
<td>128</td>
<td>1.74 × 10^{-7}</td>
<td>4.39 × 10^{-6}</td>
<td>16758499</td>
<td>6.02 × 10^{-6}</td>
<td>6.02 × 10^{-6}</td>
<td>1.58 × 10^{-2}</td>
</tr>
<tr>
<td>256</td>
<td>4.31 × 10^{-8}</td>
<td>2.20 × 10^{-6}</td>
<td>67125249</td>
<td>3.01 × 10^{-6}</td>
<td>3.01 × 10^{-6}</td>
<td>7.93 × 10^{-3}</td>
</tr>
</tbody>
</table>

### Error rates in the \(L_2\)-norm.

**Table 5.1:** Error rates for Example 1 with \(d = 1\) and \(p = 1\).

<table>
<thead>
<tr>
<th>(d=5)</th>
<th>(\theta_E=0)</th>
<th>(\theta_E=h_E)</th>
<th>(\theta=1)</th>
<th>(|u−u_0|_h)</th>
<th>(|u−u_0|_h)</th>
<th>(|u−u_0|_h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>3.60 × 10^{-4}</td>
<td>4.40 × 10^{-4}</td>
<td>0.27</td>
<td>3.27 × 10^{-4}</td>
<td>0.14</td>
<td>1.56 × 10^{-4}</td>
</tr>
<tr>
<td>1.0</td>
<td>4.73 × 10^{-4}</td>
<td>5.08 × 10^{-4}</td>
<td>0.32</td>
<td>1.41 × 10^{-4}</td>
<td>1.09</td>
<td>3.17 × 10^{-4}</td>
</tr>
<tr>
<td>2.0</td>
<td>4.91 × 10^{-5}</td>
<td>9.82 × 10^{-5}</td>
<td>0.27</td>
<td>6.20 × 10^{-5}</td>
<td>2.88</td>
<td>1.73 × 10^{-5}</td>
</tr>
<tr>
<td>4.0</td>
<td>5.58 × 10^{-6}</td>
<td>2.13 × 10^{-5}</td>
<td>0.20</td>
<td>3.94 × 10^{-5}</td>
<td>2.71</td>
<td>4.71 × 10^{-5}</td>
</tr>
<tr>
<td>8.0</td>
<td>6.31 × 10^{-7}</td>
<td>4.75 × 10^{-6}</td>
<td>0.17</td>
<td>5.13 × 10^{-6}</td>
<td>2.65</td>
<td>1.29 × 10^{-6}</td>
</tr>
<tr>
<td>16</td>
<td>8.03 × 10^{-8}</td>
<td>1.11 × 10^{-6}</td>
<td>0.10</td>
<td>6.20 × 10^{-7}</td>
<td>2.54</td>
<td>3.73 × 10^{-6}</td>
</tr>
<tr>
<td>32</td>
<td>8.04 × 10^{-9}</td>
<td>2.49 × 10^{-7}</td>
<td>0.05</td>
<td>5.67 × 10^{-8}</td>
<td>2.18</td>
<td>1.06 × 10^{-6}</td>
</tr>
<tr>
<td>64</td>
<td>8.02 × 10^{-9}</td>
<td>1.03 × 10^{-7}</td>
<td>0.03</td>
<td>4.41 × 10^{-8}</td>
<td>1.76</td>
<td>3.12 × 10^{-6}</td>
</tr>
</tbody>
</table>

### Error rates in the \(L_2\)-norm.

**Table 5.2:** Error rates for Example 1 with \(d = 1\) and \(p = 2\).
The strange behaviour of the error rates for piecewise cubic basis functions, i.e. \( p = 3 \) is most likely due to our direct solver. As the system matrix \( K_h \) has a high condition number for 37,761,025 dofs, this can lead to a loss in precision.

<table>
<thead>
<tr>
<th>dofs</th>
<th>( \theta_E = 0 )</th>
<th>( \theta_E = \varepsilon h )</th>
<th>( \theta = 1 )</th>
<th>( \theta_E = 0 )</th>
<th>( \theta_E = \varepsilon h )</th>
<th>( \theta = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( |u - u_h|_h )</td>
<td>( |u - u_h|_h )</td>
<td>( |u - u_h|_h )</td>
<td>( |u - u_h|_h )</td>
<td>( |u - u_h|_h )</td>
<td>( |u - u_h|_h )</td>
</tr>
<tr>
<td>2401</td>
<td>2.68 \times 10^{-3}</td>
<td>2.68 \times 10^{-3}</td>
<td>2.55 \times 10^{2}</td>
<td>1.45 \times 10^{-4}</td>
<td>1.45 \times 10^{-4}</td>
<td>5.40 \times 10^{2}</td>
</tr>
<tr>
<td>9439</td>
<td>1.09 \times 10^{-3}</td>
<td>3.99</td>
<td>1.08 \times 10^{7}</td>
<td>3.85</td>
<td>1.81 \times 10^{-1}</td>
<td>3.00</td>
</tr>
<tr>
<td>37,749</td>
<td>1.06 \times 10^{-3}</td>
<td>3.99</td>
<td>1.12 \times 10^{4}</td>
<td>4.05</td>
<td>2.26 \times 10^{-3}</td>
<td>3.00</td>
</tr>
<tr>
<td>148,225</td>
<td>6.65 \times 10^{-10}</td>
<td>3.99</td>
<td>1 \times 10^{-6}</td>
<td>3.53</td>
<td>-</td>
<td>148,225</td>
</tr>
<tr>
<td>594,361</td>
<td>4.16 \times 10^{-10}</td>
<td>4.00</td>
<td>7.04 \times 10^{-10}</td>
<td>3.79</td>
<td>-</td>
<td>594,361</td>
</tr>
<tr>
<td>2,362,869</td>
<td>1.95 \times 10^{-11}</td>
<td>1.10</td>
<td>1.67 \times 10^{-11}</td>
<td>2.08</td>
<td>-</td>
<td>2,362,869</td>
</tr>
<tr>
<td>9,443,329</td>
<td>1.46 \times 10^{-11}</td>
<td>0.41</td>
<td>5.81 \times 10^{-11}</td>
<td>1.80</td>
<td>-</td>
<td>9,443,329</td>
</tr>
<tr>
<td>37,761,025</td>
<td>8.64 \times 10^{-13}</td>
<td>-2.56</td>
<td>1.93 \times 10^{-11}</td>
<td>1.59</td>
<td>-</td>
<td>37,761,025</td>
</tr>
</tbody>
</table>

Error rates in the \( L_2 \)-norm.  

Error rates in the \( \| \cdot \|_h \)-norm.

Table 5.3: Error rates for Example 1 with \( d = 1 \) and \( p = 3 \).
Example 2

For this example, we considered a piece-wise constant, but globally discontinuous \( \nu \), with

\[
\nu(x,t) = \begin{cases} 
\nu_1, & \text{for } (x,t) \in Q_1 \cup Q_3; \\
\nu_2, & \text{for } (x,t) \in Q_2.
\end{cases}
\]

The jump in \( \nu \) is ranging from a moderate height, i.e., \( \nu_1 = 10 \) and \( \nu_2 = 1 \), to big jumps, i.e., \( \nu_1 = 795.774 \) and \( \nu_2 = 165 \). The latter resembles an example where \( \nu_1 \) around the reluctivity of air, whereas \( \nu_2 \) is close to the reluctivity of iron. Moreover, we also tested for switched \( \nu_1 \) and \( \nu_2 \), where we deduced a different behaviour. Since we only considered \( p = 1 \), the choice \( c = 1 \) was sufficient. However, we consider now a coefficient \( \nu \) different from 1. Thus, we chose \( \theta_E = h_E/\nu_E \) (c.f. (4.20)).

![Figure 5.2: Plots of the space time cylinder \( Q \) used for Example 2 for \( d = 1 \).](image)

We start with Example 2(a) and the case \( \nu_1 = 10 \) and \( \nu_2 = 1 \). Here, the reference solution \( u^{(ref)} \) is calculated on a mesh with 25,176,065 dofs. We deduce that the error rates in the \( L_2 \)- and the \( \| \cdot \|_h \)-norm are very similar for all choices of \( \theta_E \). Moreover, for \( \theta_E \in \{0, h_E\} \), the convergence rates in the \( \| \cdot \|_1 \)-norm are better than expected, where we almost reach an order of \( O(h^2) \). This might be because we do not compute the error against the exact solution, but only a solution on a finer mesh. For constant \( \theta_E = 1 \), the convergence rates are almost cut in half.

<table>
<thead>
<tr>
<th>dofs</th>
<th>( \theta_E = 0 )</th>
<th>( \theta_E = h_E/\nu_E )</th>
<th>( \theta = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( |u - u_h|_2 )</td>
<td>Rate</td>
<td>( |u - u_h|_2 )</td>
<td>Rate</td>
</tr>
<tr>
<td>425</td>
<td>7.79 x 10^{-4}</td>
<td>0</td>
<td>9.41 x 10^{-4}</td>
</tr>
<tr>
<td>1617</td>
<td>4.36 x 10^{-4}</td>
<td>0.84</td>
<td>4.78 x 10^{-4}</td>
</tr>
<tr>
<td>6305</td>
<td>2.16 x 10^{-4}</td>
<td>1.01</td>
<td>2.31 x 10^{-4}</td>
</tr>
<tr>
<td>24897</td>
<td>7.52 x 10^{-5}</td>
<td>1.52</td>
<td>7.99 x 10^{-5}</td>
</tr>
<tr>
<td>98945</td>
<td>2.14 x 10^{-5}</td>
<td>1.81</td>
<td>2.27 x 10^{-5}</td>
</tr>
<tr>
<td>394497</td>
<td>5.62 x 10^{-6}</td>
<td>1.93</td>
<td>5.92 x 10^{-6}</td>
</tr>
<tr>
<td>1,575,425</td>
<td>1.37 x 10^{-6}</td>
<td>2.04</td>
<td>1.44 x 10^{-6}</td>
</tr>
<tr>
<td>6,296,577</td>
<td>2.77 x 10^{-7}</td>
<td>2.31</td>
<td>2.91 x 10^{-7}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dofs</th>
<th>( \theta_E = 0 )</th>
<th>( \theta_E = h_E/\nu_E )</th>
<th>( \theta = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( |u - u_h|_h )</td>
<td>Rate</td>
<td>( |u - u_h|_h )</td>
<td>Rate</td>
</tr>
<tr>
<td>425</td>
<td>9.32 x 10^{-5}</td>
<td>0</td>
<td>1.01 x 10^{-4}</td>
</tr>
<tr>
<td>1617</td>
<td>4.93 x 10^{-5}</td>
<td>0.92</td>
<td>5.16 x 10^{-5}</td>
</tr>
<tr>
<td>6305</td>
<td>2.26 x 10^{-5}</td>
<td>1.13</td>
<td>2.36 x 10^{-5}</td>
</tr>
<tr>
<td>24897</td>
<td>7.66 x 10^{-6}</td>
<td>1.56</td>
<td>7.97 x 10^{-6}</td>
</tr>
<tr>
<td>98945</td>
<td>2.30 x 10^{-6}</td>
<td>1.74</td>
<td>2.37 x 10^{-6}</td>
</tr>
<tr>
<td>394497</td>
<td>6.35 x 10^{-7}</td>
<td>1.85</td>
<td>6.51 x 10^{-7}</td>
</tr>
<tr>
<td>1,575,425</td>
<td>1.61 x 10^{-7}</td>
<td>1.98</td>
<td>1.64 x 10^{-7}</td>
</tr>
<tr>
<td>6,296,577</td>
<td>3.37 x 10^{-8}</td>
<td>2.25</td>
<td>3.43 x 10^{-8}</td>
</tr>
</tbody>
</table>

Table 5.4: Error rates for Example 2(a) with \( d = 1, p = 1 \), and \( \nu_1 = 10 \) and \( \nu_2 = 1 \).

Now we increase the height of the discontinuity in \( \nu \) and consider \( \nu_1 = 100 \), while \( \nu_2 \).
is the same as before. The error rates in both norms are again almost identical for \( \theta_E = 0 \) and \( \theta_E = h_E/\nu_E \) (c.f. Table 5.5).

### Table 5.5: Error rates for Example 2(a) with \( d = 1 \), \( p = 1 \), and \( \nu_1 = 100 \) and \( \nu_2 = 1 \).

By interchanging \( \nu_1 \) and \( \nu_2 \), i.e., \( \nu_1 = 1 \) and \( \nu_2 = 100 \), we notice that the error rates are now much better as before. For \( \theta_E = 0 \) and \( \theta_E = h_E \), we immediately reach a convergence order of 2, which is again better than expected. For \( \theta_E = 1 \), we deduce only reduced convergence (c.f. Table 5.6).

### Table 5.6: Error rates for Example 2(a) with \( d = 1 \), \( p = 1 \), and \( \nu_1 = 1 \) and \( \nu_2 = 100 \).

At last, we consider the most difficult case, with \( \nu_1 = 795774 \) and \( \nu_2 = 165 \). Here, we notice a different behaviour. In the beginning, the convergence rates for \( \theta_E = 0 \) and \( \theta_E = h_E \) behave very good. However, once we reach a certain number of dofs, in our case at 394497, the error starts growing again. This behaviour might be due to the bad condition number of the matrix \( K_h \). The height of the discontinuity has direct influence on the condition number, increasing it further by a factor of \( \sim 8000 \). But for \( \theta_E = 1 \), the situation is different. Although the quantity of the absolute errors is worse than for the other choices of \( \theta_E \), the overall convergence rates remain stable (c.f. Table 5.7).
CHAPTER 5. NUMERICAL RESULTS

Error rates in the $L_2$-norm.

Table 5.7: Error rates for Example 2(a) with $d = 1$, $p = 1$, and $\nu_1 = 795774$ and $\nu_2 = 1$.

We continue with Example 2(b), where the discontinuity now depends on space and time. The finest mesh solution $u^{(fns)}$ is computed on a mesh with 25176065 dofs. For the simplest case, i.e. $\nu_1 = 10$ and $\nu_2 = 1$, we observe again better than expected convergence rates, but slightly worse than for Example 2(a).

Error rates in the $L_2$-norm.

Table 5.8: Error rates for Example 2(b) with $d = 1$, $p = 1$, and $\nu_1 = 10$ and $\nu_2 = 1$.

The choice $\nu_1 = 100$ and $\nu_2 = 1$ behaves similarly as in Example 2(a).

Error rates in the $L_2$-norm.

Table 5.9: Error rates for Example 2(b) with $d = 1$, $p = 1$, and $\nu_1 = 100$ and $\nu_2 = 1$.

For the converse choice, i.e., $\nu_1 = 1$ and $\nu_2 = 100$, we note slightly worse convergence rates compared to Example 2(a), but still better than expected.
CHAPTER 5. NUMERICAL RESULTS

Finally, we consider Example 2(c), where we have a non-smooth space-time interface \( E \).

Error rates in the \( L_2 \)-norm. Error rates in the \( \| \cdot \|_h \)-norm.

Table 5.10: Error rates for Example 2(b) with \( d = 1, p = 1 \) and \( \nu_1 = 1 \) and \( \nu_2 = 100 \).

The rates for the last case, \( \nu_1 = 795774 \) and \( \nu_2 = 165 \), we have again a certain number of dofs where the error rate stagnates, however it seems that for this example, the rates start improving again, at least in the \( \| \cdot \|_h \)-norm.

Error rates in the \( L_2 \)-norm. Error rates in the \( \| \cdot \|_h \)-norm.

Table 5.11: Error rates for Example 2(b) with \( d = 1, p = 1 \) and \( \nu_1 = 795774 \) and \( \nu_2 = 165 \).

Finally, we consider Example 2(c), where we have a non-smooth space-time interface for the discontinuity. Here we expect the most troubles. The reference solution \( u^{(fms)} \) has 33,564,673 dofs. For the first two choices for \( \nu \), i.e., \( \nu_1 \in \{10, 100\} \) and \( \nu_2 = 1 \), the observed convergence rates are similar, but slightly reduced, as in the previous two examples.

Error rates in the \( L_2 \)-norm. Error rates in the \( \| \cdot \|_h \)-norm.

Table 5.12: Error rates for Example 2(c) with \( d = 1, p = 1 \), and \( \nu_1 = 10 \) and \( \nu_2 = 1 \).
CHAPTER 5. NUMERICAL RESULTS

Table 5.13: Error rates for Example 2(c) with $d = 1$, $p = 1$, and $\nu_1 = 100$ and $\nu_2 = 1$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta = 0$</th>
<th>$\theta = h_0/\nu_0$</th>
<th>$\theta = 1$</th>
<th>$\theta = h_0/\nu_0$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|$</td>
<td>Rate</td>
<td>$|u - u_h|$</td>
<td>Rate</td>
<td>$|u - u_h|$</td>
</tr>
<tr>
<td>553</td>
<td>$2.22 \times 10^{-3}$</td>
<td>0</td>
<td>$3.16 \times 10^{-4}$</td>
<td>0</td>
<td>$3.29 \times 10^{-4}$</td>
</tr>
<tr>
<td>2129</td>
<td>$1.06 \times 10^{-3}$</td>
<td>1.07</td>
<td>$1.25 \times 10^{-4}$</td>
<td>1.34</td>
<td>$1.90 \times 10^{-5}$</td>
</tr>
<tr>
<td>8523</td>
<td>$3.83 \times 10^{-4}$</td>
<td>0.33</td>
<td>$3.99 \times 10^{-5}$</td>
<td>0.49</td>
<td>$1.87 \times 10^{-5}$</td>
</tr>
<tr>
<td>31009</td>
<td>$5.71 \times 10^{-5}$</td>
<td>0.55</td>
<td>$5.90 \times 10^{-6}$</td>
<td>0.59</td>
<td>$5.81 \times 10^{-6}$</td>
</tr>
<tr>
<td>131713</td>
<td>$2.92 \times 10^{-6}$</td>
<td>0.97</td>
<td>$2.97 \times 10^{-7}$</td>
<td>0.99</td>
<td>$3.03 \times 10^{-7}$</td>
</tr>
<tr>
<td>525669</td>
<td>$1.31 \times 10^{-7}$</td>
<td>1.15</td>
<td>$1.33 \times 10^{-8}$</td>
<td>1.16</td>
<td>$1.58 \times 10^{-8}$</td>
</tr>
<tr>
<td>2099713</td>
<td>$5.40 \times 10^{-8}$</td>
<td>1.28</td>
<td>$5.42 \times 10^{-9}$</td>
<td>1.29</td>
<td>$6.68 \times 10^{-9}$</td>
</tr>
<tr>
<td>83931729</td>
<td>$1.90 \times 10^{-10}$</td>
<td>1.51</td>
<td>$1.90 \times 10^{-10}$</td>
<td>1.51</td>
<td>$2.29 \times 10^{-9}$</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm. Error rates in the $\| \cdot \|_{h}$-norm.

If we interchange the values of $\nu_1$ and $\nu_2$, we reach now at most linear convergence rates in the $\| \cdot \|_{h}$-norm.

Table 5.14: Error rates for Example 2(c) with $d = 1$, $p = 1$, and $\nu_1 = 100$ and $\nu_2 = 100$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta = 0$</th>
<th>$\theta = h_0/\nu_0$</th>
<th>$\theta = 1$</th>
<th>$\theta = h_0/\nu_0$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|$</td>
<td>Rate</td>
<td>$|u - u_h|$</td>
<td>Rate</td>
<td>$|u - u_h|$</td>
</tr>
<tr>
<td>553</td>
<td>$3.05 \times 10^{-3}$</td>
<td>0</td>
<td>$4.25 \times 10^{-4}$</td>
<td>0</td>
<td>$7.37 \times 10^{-5}$</td>
</tr>
<tr>
<td>2129</td>
<td>$9.33 \times 10^{-4}$</td>
<td>1.71</td>
<td>$1.19 \times 10^{-4}$</td>
<td>1.83</td>
<td>$3.99 \times 10^{-5}$</td>
</tr>
<tr>
<td>8523</td>
<td>$3.08 \times 10^{-4}$</td>
<td>1.60</td>
<td>$3.62 \times 10^{-5}$</td>
<td>1.72</td>
<td>$2.08 \times 10^{-5}$</td>
</tr>
<tr>
<td>31009</td>
<td>$1.14 \times 10^{-5}$</td>
<td>1.43</td>
<td>$1.23 \times 10^{-6}$</td>
<td>1.56</td>
<td>$1.05 \times 10^{-6}$</td>
</tr>
<tr>
<td>131713</td>
<td>$4.63 \times 10^{-7}$</td>
<td>1.30</td>
<td>$5.78 \times 10^{-8}$</td>
<td>1.36</td>
<td>$5.16 \times 10^{-8}$</td>
</tr>
<tr>
<td>525669</td>
<td>$1.88 \times 10^{-8}$</td>
<td>1.30</td>
<td>$1.91 \times 10^{-9}$</td>
<td>1.31</td>
<td>$2.43 \times 10^{-9}$</td>
</tr>
<tr>
<td>2099713</td>
<td>$7.29 \times 10^{-9}$</td>
<td>1.37</td>
<td>$7.32 \times 10^{-10}$</td>
<td>1.38</td>
<td>$1.04 \times 10^{-9}$</td>
</tr>
<tr>
<td>83931729</td>
<td>$2.27 \times 10^{-10}$</td>
<td>1.68</td>
<td>$2.27 \times 10^{-10}$</td>
<td>1.69</td>
<td>$4.14 \times 10^{-9}$</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm. Error rates in the $\| \cdot \|_{h}$-norm.

For the highest jump in the coefficient, the situation is now different than before. Here, we do not see any stagnation in the $\| \cdot \|_{h}$-norm and also have almost linear convergence. However, in the $L_2$-norm, the rates are as in the previous examples, i.e., stagnating and diverging.

Table 5.15: Error rates for Example 2(c) with $d = 1$, $p = 1$, and $\nu_1 = 795774$ and $\nu_2 = 165$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta = 0$</th>
<th>$\theta = h_0/\nu_0$</th>
<th>$\theta = 1$</th>
<th>$\theta = h_0/\nu_0$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|$</td>
<td>Rate</td>
<td>$|u - u_h|$</td>
<td>Rate</td>
<td>$|u - u_h|$</td>
</tr>
<tr>
<td>553</td>
<td>$8.17 \times 10^{-2}$</td>
<td>0</td>
<td>$8.15 \times 10^{-3}$</td>
<td>0</td>
<td>$1.58 \times 10^{-4}$</td>
</tr>
<tr>
<td>2129</td>
<td>$2.84 \times 10^{-2}$</td>
<td>1.52</td>
<td>$2.83 \times 10^{-3}$</td>
<td>1.53</td>
<td>$8.14 \times 10^{-5}$</td>
</tr>
<tr>
<td>8523</td>
<td>$9.82 \times 10^{-3}$</td>
<td>1.53</td>
<td>$9.78 \times 10^{-4}$</td>
<td>1.53</td>
<td>$4.11 \times 10^{-5}$</td>
</tr>
<tr>
<td>31009</td>
<td>$3.40 \times 10^{-3}$</td>
<td>1.53</td>
<td>$3.39 \times 10^{-4}$</td>
<td>1.53</td>
<td>$2.13 \times 10^{-5}$</td>
</tr>
<tr>
<td>131713</td>
<td>$1.24 \times 10^{-4}$</td>
<td>1.45</td>
<td>$1.24 \times 10^{-5}$</td>
<td>1.45</td>
<td>$1.18 \times 10^{-6}$</td>
</tr>
<tr>
<td>525669</td>
<td>$7.24 \times 10^{-5}$</td>
<td>0.78</td>
<td>$7.24 \times 10^{-6}$</td>
<td>0.77</td>
<td>$5.68 \times 10^{-7}$</td>
</tr>
<tr>
<td>2099713</td>
<td>$8.38 \times 10^{-6}$</td>
<td>0.21</td>
<td>$8.38 \times 10^{-7}$</td>
<td>0.21</td>
<td>$2.85 \times 10^{-7}$</td>
</tr>
<tr>
<td>83931729</td>
<td>$1.04 \times 10^{-8}$</td>
<td>0.31</td>
<td>$1.04 \times 10^{-8}$</td>
<td>0.31</td>
<td>$1.22 \times 10^{-7}$</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm. Error rates in the $\| \cdot \|_{h}$-norm.

Example 3

In this example, we consider again a solution with reduced regularity. However, here we know exactly to which space the exact solution belongs. We choose $\lambda = 1.45$ from which we deduce that the analytical solution $u$ belongs to the space $H^{1.95-\epsilon}(Q)$. 


For piecewise linear basis functions, we deduce no change in any of the convergence rates.

For increased polynomial degree, i.e., $p = 2$, we indeed note a difference. For all three values of $\theta_E$, the convergence rates in both norms are reduced, and, for $\theta_E = 0$ and $\theta_E = h_E$, are of order $O(h^{1.95})$. This directly resembles the regularity of our exact solution, which belongs to the space $H^{1.95-\epsilon}(Q)$.

For the highest polynomial degree $p = 3$, we observe the same rates as for $p = 2$, i.e., the convergence order in both norms is $O(h^{1.95})$, provided that we choose $\theta_E = 0$ or $\theta_E = h_E$. For $\theta_E = 1$, the error diverges.
5.2 2D Examples

In this section, we analyze the numerical results for \( d = 2 \). This results in an three dimensional finite element method, where the number of dofs is higher and also the sparsity pattern of the system matrix \( K_h \) is more dense. Hence, we switch from a direct solver to an iterative solving method. As already mentioned, our bilinear form \( a_h(\cdot, \cdot) \) is non symmetric and thus also the system matrix \( K_h \) has a non-symmetric, but positive definite, structure. This prevents us from using the powerful conjugate gradient (CG) method \( [21] \). For our non-symmetric linear problem (4.39), we could either transform it to the normal equation

\[
K_h^T K_h u_h = K_h^T f_h,
\]

and use here the CG-method. This is called the CG for the normal equation (CGNE). Or we could try to minimize the residual

\[
\| K_h u_h - f_h \|,
\]

over a special subspace of \( \mathbb{R}^{N_h} \). This is called generalized minimal residual method (GMRES). We decided to use the latter. As a preconditioner, we used the algebraic multigrid method described in \( [23] \), which is included in the HYPRE solver package.

**Example 1**

We begin with the highly smooth example. The convergence rates are again optimal for each polynomial degree, if we choose \( \theta_E = 0 \) or \( \theta_E = c \varepsilon h_E \). The necessity to choose \( \theta_E \) in such a way is now more apparent than for \( d = 1 \), as there is no convergence for \( \theta_E = 1 \) and \( p \geq 2 \) (compare Table 5.2 and Table 5.20). For \( p = 2 \), we chose \( c = 0.001 \), while for \( p = 3 \), the choice \( c = 10^{-5} \) was made.

<table>
<thead>
<tr>
<th>dofs</th>
<th>( \theta_E = 0 )</th>
<th>( \theta_E = h_E )</th>
<th>( \theta_E = 0 )</th>
<th>( \theta_E = h_E )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u - u_h )</td>
<td>( u - u_h )</td>
<td>( u - u_h )</td>
<td>( u - u_h )</td>
<td>( u - u_h )</td>
</tr>
<tr>
<td>Rate</td>
<td>Rate</td>
<td>Rate</td>
<td>Rate</td>
<td>Rate</td>
</tr>
<tr>
<td>125</td>
<td>( 3.11 \times 10^{-3} )</td>
<td>( 1.14 \times 10^{-3} )</td>
<td>( 1.38 \times 10^{-3} )</td>
<td>( 1.13 \times 10^{-3} )</td>
</tr>
<tr>
<td>729</td>
<td>( 3.25 \times 10^{-3} )</td>
<td>( 1.77 \times 10^{-3} )</td>
<td>( 1.73 \times 10^{-3} )</td>
<td>( 1.70 \times 10^{-3} )</td>
</tr>
<tr>
<td>4913</td>
<td>( 8.83 \times 10^{-3} )</td>
<td>( 1.88 \times 10^{-3} )</td>
<td>( 1.81 \times 10^{-3} )</td>
<td>( 1.79 \times 10^{-3} )</td>
</tr>
<tr>
<td>35937</td>
<td>( 2.25 \times 10^{-2} )</td>
<td>( 1.97 \times 10^{-2} )</td>
<td>( 1.95 \times 10^{-2} )</td>
<td>( 1.93 \times 10^{-2} )</td>
</tr>
<tr>
<td>274625</td>
<td>( 5.60 \times 10^{-3} )</td>
<td>( 2.01 \times 10^{-3} )</td>
<td>( 2.00 \times 10^{-3} )</td>
<td>( 1.99 \times 10^{-3} )</td>
</tr>
<tr>
<td>2146689</td>
<td>( 1.39 \times 10^{-3} )</td>
<td>( 2.01 \times 10^{-4} )</td>
<td>( 2.00 \times 10^{-4} )</td>
<td>( 1.99 \times 10^{-4} )</td>
</tr>
<tr>
<td>16974593</td>
<td>( 3.45 \times 10^{-3} )</td>
<td>( 2.01 \times 10^{-3} )</td>
<td>( 2.00 \times 10^{-3} )</td>
<td>( 1.99 \times 10^{-3} )</td>
</tr>
</tbody>
</table>

Error rates in the \( L_2 \)-norm.

<table>
<thead>
<tr>
<th>Error rates in the ( | \cdot |_h )-norm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 5.19: Error rates for Example 1 with ( d = 2 ) and ( p = 1 ).</td>
</tr>
</tbody>
</table>
CHAPTER 5. NUMERICAL RESULTS

For $p = 3$ we note that for $\theta_E = ch_E$, the error rate in the $L_2$-norm is. However, once a certain number of dofs is reached, the absolute error is again of the same magnitude as $\theta_E = 0$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$h_E = 0$</th>
<th>$h_E = ch_E$</th>
<th>$\theta = 1$</th>
<th>dofs</th>
<th>$h_E = 0$</th>
<th>$h_E = ch_E$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2197</td>
<td>1.70 $\times 10^{-2}$</td>
<td>1.71 $\times 10^{-2}$</td>
<td>0</td>
<td>2197</td>
<td>3.13 $\times 10^{-2}$</td>
<td>3.14 $\times 10^{-2}$</td>
<td>0</td>
</tr>
<tr>
<td>117649</td>
<td>7.79 $\times 10^{-4}$</td>
<td>7.82 $\times 10^{-4}$</td>
<td>3.82</td>
<td>117649</td>
<td>4.85 $\times 10^{-4}$</td>
<td>4.86 $\times 10^{-4}$</td>
<td>3.01</td>
</tr>
<tr>
<td>912673</td>
<td>5.50 $\times 10^{-7}$</td>
<td>5.63 $\times 10^{-7}$</td>
<td>3.80</td>
<td>912673</td>
<td>6.03 $\times 10^{-7}$</td>
<td>6.04 $\times 10^{-7}$</td>
<td>3.01</td>
</tr>
<tr>
<td>7189557</td>
<td>3.80 $\times 10^{-10}$</td>
<td>3.83</td>
<td>5.22 $\times 10^{-10}$</td>
<td>3.83</td>
<td>7189557</td>
<td>7.53 $\times 10^{-10}$</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm. Error rates in the $\| \cdot \|_h$-norm.

Table 5.20: Error rates for Example 1 with $d = 2$ and $p = 2$.

As we are now using an iterative solver, we can compare the time needed to solve the linear system (c.f. Table 5.22). We see that for $p \leq 2$, our method with performs better than $\theta_E = 0$, where we need less time. However, for $p = 3$, $\theta_E = 0$ performs slightly better than the other choices of $\theta_E$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$h_E = 0$</th>
<th>$h_E = ch_E$</th>
<th>$\theta = 1$</th>
<th>dofs</th>
<th>$h_E = 0$</th>
<th>$h_E = ch_E$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>125</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>125</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>729</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>729</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>4913</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>4913</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>35937</td>
<td>0.10</td>
<td>0.11</td>
<td>0.07</td>
<td>35937</td>
<td>0.10</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>274625</td>
<td>0.81</td>
<td>0.75</td>
<td>0.34</td>
<td>274625</td>
<td>0.81</td>
<td>0.75</td>
<td>0.34</td>
</tr>
<tr>
<td>2146689</td>
<td>18.51</td>
<td>15.77</td>
<td>3.49</td>
<td>2146689</td>
<td>41.78</td>
<td>41.81</td>
<td>188.04</td>
</tr>
<tr>
<td>16974593</td>
<td>871.13</td>
<td>744.82</td>
<td>59.57</td>
<td>16974593</td>
<td>11091.13</td>
<td>1032.82</td>
<td>1195.64</td>
</tr>
</tbody>
</table>

$p = 1$

Table 5.22: Time needed to solve the linear system for Example 1.
CHAPTER 5. NUMERICAL RESULTS

Figure 5.3: Decomposition of the 3D space time cylinder $Q = (0,1)^3$ into 256 subdomains for parallel computing on 256 cores.

Example 2

For this example, we considered again a piece-wise constant, but globally discontinuous $\nu$, with

$$\nu(x,t) = \begin{cases} 
\nu_1, & \text{for } (x,t) \in Q_1, \\
\nu_2, & \text{for } (x,t) \in Q_2,
\end{cases}$$

where $Q_1$ is the transparent part and $Q_2$ the non-transparent part of the figures in Figure 5.4. As for the choice of $\nu_1$ and $\nu_2$, we use one pair of values less than for $d = 1$, i.e., we consider

- $\nu_1 = 100$ and $\nu_2 = 1$,
- $\nu_1 = 1$ and $\nu_2 = 100$, and
- $\nu_1 = 795774$ and $\nu_2 = 165$. 
CHAPTER 5. NUMERICAL RESULTS

For the highest difference in $\nu$ dofs $\theta$, we interchange $\nu_1$ and $\nu_2$. The reference solution $u^{(fms)}$ was computed on a mesh with 21,835,009 dofs. Here we see, in contrast to $d = 1$, only reduced convergence rates. Moreover, the error rates in the $\|\cdot\|_h$-norm are again almost the same for both choices of $\theta_E$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = \theta_E / \nu_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_0$</td>
<td>Rate</td>
</tr>
<tr>
<td>133</td>
<td>$1.01 \times 10^{-2}$</td>
<td>0</td>
</tr>
<tr>
<td>841</td>
<td>$6.48 \times 10^{-3}$</td>
<td>0.64</td>
</tr>
<tr>
<td>5969</td>
<td>$2.11 \times 10^{-3}$</td>
<td>1.62</td>
</tr>
<tr>
<td>44,961</td>
<td>$6.21 \times 10^{-4}$</td>
<td>1.77</td>
</tr>
<tr>
<td>348,993</td>
<td>$1.75 \times 10^{-4}$</td>
<td>1.83</td>
</tr>
<tr>
<td>2,750,081</td>
<td>$4.45 \times 10^{-5}$</td>
<td>1.97</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm.

Table 5.23: Error rates for Example 2(a) with $d = 2$, $p = 1$, and $\nu_1 = 100$ and $\nu_2 = 1$.

If we interchange $\nu_1$ and $\nu_2$, the convergence rates in the $\|\cdot\|_h$-norm improve and are now linear, in contrast to $d = 1$, where we attained quadratic convergence.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = \theta_E / \nu_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_0$</td>
<td>Rate</td>
</tr>
<tr>
<td>133</td>
<td>$5.06 \times 10^{-2}$</td>
<td>0</td>
</tr>
<tr>
<td>841</td>
<td>$2.80 \times 10^{-2}$</td>
<td>0.86</td>
</tr>
<tr>
<td>5969</td>
<td>$1.54 \times 10^{-2}$</td>
<td>0.86</td>
</tr>
<tr>
<td>44,961</td>
<td>$8.42 \times 10^{-3}$</td>
<td>0.88</td>
</tr>
<tr>
<td>348,993</td>
<td>$3.48 \times 10^{-3}$</td>
<td>1.27</td>
</tr>
<tr>
<td>2,750,081</td>
<td>$9.42 \times 10^{-4}$</td>
<td>1.88</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm.

Table 5.24: Error rates for Example 2(a) with $d = 2$, $p = 1$, and $\nu_1 = 1$ and $\nu_2 = 100$.

For the highest difference in $\nu_1$ and $\nu_2$, the convergence behaviour is again different than for $d = 1$. Here, we reach almost optimal rates, i.e., quadratic convergence in the $L_2$-norm and linear convergence in the $\|\cdot\|_h$-norm (c.f. Table 5.23).
For the last case, i.e., \( \nu \) we observe that the convergence rates in both norms are worse than for \( \theta \text{ dofs} \).

Next we consider Example 2(b), where the coefficient \( \nu \) depends on space and time. Here, \( u(fms) \) was calculated on a mesh with 25,363,073 dofs. For \( \nu_1 = 100 \) and \( \nu_2 = 1 \), we observe that the convergence rates in both norms are worse than for \( d = 1 \).

For interchanged \( \nu_1 \) and \( \nu_2 \), the rates are again better than for the previous case. We obtain linear convergence rates in the \( \| \cdot \|_h \)-norm, but for the \( L_2 \)-norm, the rates are worse than linear.

For the last case, i.e., \( \nu_1 = 795774 \) and \( \nu_2 = 165 \), we notice a similar behaviour as for \( d = 1 \). Once a certain number of dofs is reached, the error in the \( \| \cdot \|_h \)-norm stops decreasing, while the \( L_2 \)-norm is still decreasing.
CHAPTER 5. NUMERICAL RESULTS

For a coefficient with non-smooth space-time interface, as in Example 2(c), we used a reference solution with 26428097 dfs. We have no convergence in the $\| \cdot \|_h$-norm for $\nu_1 = 100$ and $\nu_2 = 1$. The $L_2$-norm is fine, however we only get reduced rates. This is most likely due to the loss in regularity of the analytical solution. Another reason could be the that we are too close to the reference solution $u^{(fms)}$.

<table>
<thead>
<tr>
<th>dfs</th>
<th>$\theta_E = 0$ $| u - u_h |_0$ Rate</th>
<th>$\theta_E = h \nu_E$ $| u - u_h |_0$ Rate</th>
<th>$\theta_E = 0$ $| u - u_h |_h$ Rate</th>
<th>$\theta_E = h \nu_E$ $| u - u_h |_h$ Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>$2.46 \times 10^{-3}$ 0</td>
<td>$0.52 \times 10^{-3}$ 0</td>
<td>$5.57 \times 10^{-2}$ 0</td>
<td>$1.12 \times 10^{-1}$ 0</td>
</tr>
<tr>
<td>1031</td>
<td>$2.02 \times 10^{-3}$ 0.16</td>
<td>$2.89 \times 10^{-3}$ 1.18</td>
<td>$7.57 \times 10^{-2}$ -0.46</td>
<td>$8.13 \times 10^{-2}$ 0.46</td>
</tr>
<tr>
<td>7277</td>
<td>$7.18 \times 10^{-4}$ 1.62</td>
<td>$9.09 \times 10^{-4}$ 1.67</td>
<td>$4.50 \times 10^{-2}$ 0.75</td>
<td>$4.57 \times 10^{-2}$ 0.83</td>
</tr>
<tr>
<td>514617</td>
<td>$2.51 \times 10^{-4}$ 1.52</td>
<td>$2.90 \times 10^{-4}$ 1.65</td>
<td>$3.12 \times 10^{-2}$ 0.53</td>
<td>$3.12 \times 10^{-2}$ 0.55</td>
</tr>
<tr>
<td>423089</td>
<td>$1.31 \times 10^{-4}$ 0.93</td>
<td>$1.37 \times 10^{-4}$ 1.08</td>
<td>$3.23 \times 10^{-2}$ -0.05</td>
<td>$3.23 \times 10^{-2}$ -0.05</td>
</tr>
<tr>
<td>3330401</td>
<td>$6.64 \times 10^{-5}$ 0.98</td>
<td>$6.70 \times 10^{-5}$ 1.03</td>
<td>$4.40 \times 10^{-2}$ -0.45</td>
<td>$4.40 \times 10^{-2}$ -0.45</td>
</tr>
</tbody>
</table>

Table 5.29: Error rates for Example 2(c) with $d = 2$, $p = 1$, and $\nu_1 = 1$ and $\nu_2 = 1$.

However, for interchanged values of $\nu_1$ and $\nu_2$, the convergence rates in the $L_2$-norm as well as in the $\| \cdot \|_h$-norm are almost optimal, i.e., $\mathcal{O}(h^2)$ and $\mathcal{O}(h)$, respectively.

<table>
<thead>
<tr>
<th>dfs</th>
<th>$\theta_E = 0$ $| u - u_h |_0$ Rate</th>
<th>$\theta_E = h \nu_E$ $| u - u_h |_0$ Rate</th>
<th>$\theta_E = 0$ $| u - u_h |_h$ Rate</th>
<th>$\theta_E = h \nu_E$ $| u - u_h |_h$ Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>$3.91 \times 10^{-2}$ 0</td>
<td>$6.91 \times 10^{-2}$ 0</td>
<td>$2.92 \times 10^{-1}$ 0</td>
<td>$4.59 \times 10^{-1}$ 0</td>
</tr>
<tr>
<td>1031</td>
<td>$1.74 \times 10^{-2}$ 1.17</td>
<td>$2.82 \times 10^{-2}$ 1.29</td>
<td>$4.52 \times 10^{-1}$ -0.63</td>
<td>$4.71 \times 10^{-1}$ -0.04</td>
</tr>
<tr>
<td>7277</td>
<td>$5.62 \times 10^{-3}$ 1.63</td>
<td>$8.37 \times 10^{-3}$ 1.75</td>
<td>$2.62 \times 10^{-1}$ 0.79</td>
<td>$2.64 \times 10^{-1}$ 0.83</td>
</tr>
<tr>
<td>514617</td>
<td>$1.65 \times 10^{-3}$ 1.77</td>
<td>$2.26 \times 10^{-3}$ 1.89</td>
<td>$1.38 \times 10^{-1}$ 0.92</td>
<td>$1.39 \times 10^{-1}$ 0.93</td>
</tr>
<tr>
<td>423089</td>
<td>$4.84 \times 10^{-4}$ 1.76</td>
<td>$6.04 \times 10^{-4}$ 1.90</td>
<td>$7.80 \times 10^{-2}$ 0.95</td>
<td>$7.18 \times 10^{-2}$ 0.95</td>
</tr>
<tr>
<td>3330401</td>
<td>$1.29 \times 10^{-4}$ 1.91</td>
<td>$1.49 \times 10^{-4}$ 2.02</td>
<td>$3.90 \times 10^{-2}$ 0.88</td>
<td>$3.90 \times 10^{-2}$ 0.88</td>
</tr>
</tbody>
</table>

Table 5.30: Error rates for Example 2(c) with $d = 2$, $p = 1$, and $\nu_1 = 1$ and $\nu_2 = 100$.

And for the last coefficient choice, we again notice stagnating convergence rates in the $\| \cdot \|_h$-norm and reduced rates in the $L_2$-norm.
CHAPTER 5. NUMERICAL RESULTS

55

Table 5.33: Time needed to solve the linear system for Example 2, with

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = h_E/\nu_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_0$</td>
<td>Rate</td>
</tr>
<tr>
<td>164</td>
<td>$1.49 \times 10^{-5}$</td>
<td>0</td>
</tr>
<tr>
<td>1031</td>
<td>$1.37 \times 10^{-5}$</td>
<td>0.12</td>
</tr>
<tr>
<td>7277</td>
<td>$4.46 \times 10^{-6}$</td>
<td>1.62</td>
</tr>
<tr>
<td>54167</td>
<td>$1.47 \times 10^{-6}$</td>
<td>1.60</td>
</tr>
<tr>
<td>423089</td>
<td>$7.15 \times 10^{-7}$</td>
<td>1.04</td>
</tr>
<tr>
<td>3330401</td>
<td>$6.52 \times 10^{-7}$</td>
<td>0.13</td>
</tr>
</tbody>
</table>

As for solving times, we note that the case $\nu_1 = 1$ and $\nu = 100$ performed far worse than the other two choices of $\nu_1$ and $\nu_2$. If we compare $\theta_E = 0$ and $\theta_E = h_E$, then the solving times are very similar.

Table 5.31: Error rates for Example 2(c) with $d = 2$, $p = 1$, and $\nu_1 = 795774$ and $\nu_2 = 165$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = h_E/\nu_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_0$</td>
<td>Rate</td>
</tr>
<tr>
<td>133</td>
<td>0.01</td>
<td>0.01 128</td>
</tr>
<tr>
<td>841</td>
<td>0.01</td>
<td>0.01 128</td>
</tr>
<tr>
<td>5069</td>
<td>0.03</td>
<td>0.03 128</td>
</tr>
<tr>
<td>44961</td>
<td>0.11</td>
<td>0.11 128</td>
</tr>
<tr>
<td>348993</td>
<td>0.65</td>
<td>0.65 128</td>
</tr>
<tr>
<td>2750081</td>
<td>7.94</td>
<td>8.09 128</td>
</tr>
</tbody>
</table>

Table 5.32: Time needed to solve the linear system for Example 2, with $d = 2$, $p = 1$, and $\nu_1 = 100$ and $\nu_2 = 1$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = h_E/\nu_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_0$</td>
<td>Rate</td>
</tr>
<tr>
<td>133</td>
<td>0.02</td>
<td>0.01 128</td>
</tr>
<tr>
<td>841</td>
<td>0.02</td>
<td>0.02 128</td>
</tr>
<tr>
<td>5069</td>
<td>0.05</td>
<td>0.05 128</td>
</tr>
<tr>
<td>44961</td>
<td>0.22</td>
<td>0.21 128</td>
</tr>
<tr>
<td>348993</td>
<td>1.66</td>
<td>1.62 128</td>
</tr>
<tr>
<td>2750081</td>
<td>18.73</td>
<td>18.69 128</td>
</tr>
</tbody>
</table>

Table 5.33: Time needed to solve the linear system for Example 2, with $d = 2$, $p = 1$, and $\nu_1 = 1$ and $\nu_2 = 100$. 

Error rates in the $L_2$-norm.

Error rates in the $\|\cdot\|_h$-norm.
CHAPTER 5. NUMERICAL RESULTS

Table 5.34: Time needed to solve the linear system for Example 2, with \( d = 2 \), \( p = 1 \), and \( \nu_1 = 795774 \) and \( \nu_2 = 165 \).

Example 3

For this example, we choose \( \lambda \) exactly as in the case \( d = 1 \), i.e. \( \lambda = 1.45 \). Moreover, for \( p = 2 \) and \( p = 3 \), we chose \( c = 0.01 \). For piecewise linear basis functions, i.e. \( p = 1 \), the convergence rates are optimal in the \( \| \cdot \|_{L_2} \)-norm. However, in contrast to \( d = 1 \), the \( L_2 \)-norm only shows a reduced convergence rate, and the reduction is more severe for \( \theta_E = 0 \) (c.f. Table 5.16 and Table 5.35). This behaviour is due to the nature of our problem. When we assemble the right hand side of our linear system, we have to approximate the integral

\[
\int_E f (v_h + \theta_E h_E \partial_t v_h) \, d(x,t),
\]

where \( f \) contains the term \( |t - 0.5|^{0.45} \). This expression is difficult to numerically integrate and introduces the reduction in the convergence rates.

Table 5.35: Error rates for Example 3 with \( d = 2 \) and \( p = 1 \) and \( \lambda = 1.45 \).

In case of piecewise quadratic basis functions, the reduced rates now occur in both norms. The \( L_2 \)-norm behaves as for \( p = 1 \) and the \( \| \cdot \|_{L_2} \)-norm shows similar rates, but the difference between \( \theta_E = 0 \) and \( \theta_E = h_E \) is now less.
CHAPTER 5. NUMERICAL RESULTS

### 5.3 3D Examples

In the last section we consider now the case \( d = 3 \), which results in a 4D finite element method. Hence, there are no visualizations of the domain, or the decomposition for parallel computing. The sparsity pattern of our system matrix is more dense, as
the number of neighbouring elements is again higher in 4D than in 3D. Therefore, we again use the AMG-preconditioned GMRES method in order to solve the linear system. Moreover, the increase in the number of dofs after an uniform refinement step is much higher than in 3D, c.f. Table 5.39. Unfortunately, we have no 4D mesh generator available, which would enable us to create more complex four dimensional domains. We will only consider Example 1.

Example 1

For a problem with a highly smooth solution, our method gives optimal behaviour in the $\| \cdot \|_{h}$-norm. In the $L_2$-norm, we obtain, as before, optimal rates for $\theta_E = 0$ and $\theta_E = h_E$, whereas only reduced convergence for $\theta_E = 1$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = h_E$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_h$</td>
<td>$|u - u_h|_{L^2}$</td>
<td>$|u - u_h|_h$</td>
</tr>
<tr>
<td>625</td>
<td>$1.26 \times 10^{-4}$</td>
<td>$1.48 \times 10^{-1}$</td>
<td>$1.84 \times 10^{-1}$</td>
</tr>
<tr>
<td>6561</td>
<td>$5.69 \times 10^{-5}$</td>
<td>$1.30 \times 10^{-2}$</td>
<td>$1.18 \times 10^{-2}$</td>
</tr>
<tr>
<td>80620</td>
<td>$1.86 \times 10^{-3}$</td>
<td>$2.60 \times 10^{-4}$</td>
<td>$1.85 \times 10^{-4}$</td>
</tr>
<tr>
<td>1185921</td>
<td>$4.77 \times 10^{-4}$</td>
<td>$1.82 \times 10^{-3}$</td>
<td>$1.81 \times 10^{-3}$</td>
</tr>
<tr>
<td>17850625</td>
<td>$1.24 \times 10^{-4}$</td>
<td>$1.94 \times 10^{-3}$</td>
<td>$1.92 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm.

Table 5.39: Error rates for Example 1 with $d = 3$ and $p = 1$.

For piecewise quadratic basis functions we observe similar rates as for $d = 2$. The $L_2$-norm gives quadratic convergence for both choices of $\theta_E$. In the $\| \cdot \|_{h}$-norm, we observe optimal behaviour for both choices.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = h_E$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_h$</td>
<td>$|u - u_h|_{L^2}$</td>
<td>$|u - u_h|_h$</td>
</tr>
<tr>
<td>6561</td>
<td>$1.48 \times 10^{-4}$</td>
<td>$1.85 \times 10^{-1}$</td>
<td>$2.65 \times 10^{-4}$</td>
</tr>
<tr>
<td>83521</td>
<td>$2.31 \times 10^{-4}$</td>
<td>$2.60 \times 10^{-2}$</td>
<td>$2.83 \times 10^{-2}$</td>
</tr>
<tr>
<td>1185921</td>
<td>$4.78 \times 10^{-4}$</td>
<td>$1.93 \times 10^{-3}$</td>
<td>$3.75 \times 10^{-3}$</td>
</tr>
<tr>
<td>17850625</td>
<td>$1.17 \times 10^{-4}$</td>
<td>$2.01 \times 10^{-3}$</td>
<td>$2.20 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

Error rates in the $L_2$-norm.

Table 5.40: Error rates for Example 1 with $d = 3$ and $p = 2$.

Regarding the solving times for the linear system, the choice $\theta_E = h_E$ performs better than $\theta_E = 0$.

<table>
<thead>
<tr>
<th>dofs</th>
<th>$\theta_E = 0$</th>
<th>$\theta_E = h_E$</th>
<th>$\theta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|u - u_h|_h$</td>
<td>$|u - u_h|_{L^2}$</td>
<td>$|u - u_h|_h$</td>
</tr>
<tr>
<td>6561</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>6561</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>83521</td>
<td>0.30</td>
<td>0.29</td>
<td>0.25</td>
</tr>
<tr>
<td>1185921</td>
<td>1.14</td>
<td>1.90</td>
<td>1.56</td>
</tr>
<tr>
<td>17850625</td>
<td>59.48</td>
<td>55.57</td>
<td>41.01</td>
</tr>
</tbody>
</table>

$p = 1$

Table 5.41: Time needed to solve the linear system for Example 1.
Chapter 6

Conclusions & Outlook

In this thesis, we considered the solution of parabolic initial boundary value problems with variable coefficients, that can depend on space and time variables in a continuous or discontinuous way. However, instead of trying to solve this kind of problems in the usual way, where time $t$ has a special meaning, we choose an alternative way. We treat the time $t$ just as another variable, and consider now a problem in one dimension higher. Before we try to introduce a numerical method to solve such a space-time problem, we have to ensure that the continuous problem is well posed. We showed that the model problem (1.1) – (1.3) has indeed a unique solution in $\hat{H}^{1,0}(Q)$, where we followed the techniques presented by Ladyženskaya in [14]. An alternative way to obtain such a uniqueness and existence result was also presented, namely a formulation in Bochner spaces of abstract functions. Here we made use of the results by Steinbach [28] based on Babuška-Aziz’s theorem.

We continued by introducing a space-time finite element scheme based on local time-upwind stabilization. It satisfies a consistency identity for a sufficiently smooth solution of the model problem (1.1) – (1.3). To ensure that the finite element scheme has a unique solution, we proved ellipticity and boundedness of the stabilized space-time bilinear form. However, to ensure that the method is stable, it turned out that we have to choose the stabilization parameter $\theta_E = O(h_E)$ for all $E \in T_h$. Moreover, we were able to derive an a priori error estimate in the $\|\cdot\|_h$-norm. In addition to our space-time finite element scheme, we briefly described an alternative method introduced by Touloupoulos [32], which instead uses bubble functions for stabilization, and for which there is also an a priori error estimate available.

Next, we performed various numerical tests where we observed that our numerical results are in accordance with the a priori error estimate (4.35). Indeed, if the exact solution fulfils the requirements of Theorem 4.18, we obtain the optimal convergence rate of $O(h^p)$ wrt the $\|\cdot\|_h$, for polynomial degree $p$. Moreover, we see that the $L_2$-error has some interesting behaviour wrt the polynomial degree. For $p = 1$ and $p = 3$, we get convergence rates of order $O(h^{p+1})$, which is the usual result for elliptic problems. However, for $p = 2$, we only get a reduced rate of $O(h^{p})$. The reason behind this behaviour is unknown to the author and subject to future research.

In future work, we could try to derive some a priori error-estimate for the $L_2$-norm,
which was always studied numerically in our experiments in Chapter 5. Moreover, one could develop an a posteriori error estimator, which would enable us to use adaptive mesh refinement, leading to a space-time Adaptive Finite Element Method (AFEM). We mention that our scheme is prepared for AFEM, since we allow local mesh-sizes $h_E$ for $E \in \mathcal{T}_h$ under the condition of shape regularity of the element $E$. This will help in analysing problems with a discontinuous diffusion coefficient $\nu$. Then we can improve the solver for the huge algebraic linear system. For instance, we have to find special preconditioners that enable us to solve the problem also for a high number of dofs. We can then combine the space-time AFEM with Nested Iterations, which drastically reduces the solving time. The main future goal is the application to nonlinear parabolic problems and eddy current problems, which typically arise in electrical engineering.
Bibliography


Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Masterarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.
Die vorliegende Masterarbeit ist mit dem elektronisch übermittelten Textdokument identisch.

Linz, November 2017

______________________________
Andreas Schafelner
Curriculum Vitae

Name: Andreas Schafelner

Nationality: Austria

Date of Birth: 30 July, 1992

Place of Birth: Steyr, Austria

Education:
1998–2002 Volksschule I (elementary school)
Enns

2002–2006 Bundesrealgymnasium (secondary comprehensive school)
Enns

2006–2011 Höhere Technische Lehranstalt
Perg

2012–2016 Bachelor studies in Technical Mathematics,
Johannes Kepler University Linz

2016–2017 Master studies in Industry Mathematics,
Johannes Kepler University Linz

Special Activities:
July 2016 16th ECMI Modelling Week, Sofia, Bulgaria