VERLAG DER TECHNISCHEN UNIVERSITÄT GRAZ

Christian Köthe

Adaptive least-squares space-time finite element methods

CES 48

MONOGRAPHIC SERIES TU GRAZ
COMPUTATION IN ENGINEERING AND SCIENCE

Köthe Christian

Adaptive least-squares space-time finite element methods

Monographic Series TU Graz
Computation in Engineering and Science

Series Editors

Monographic Series TU Graz
Computation in Engineering and Science Volume 48

Köthe Christian

Adaptive least-squares space-time finite element methods

This work is based on the dissertation “Adaptive least-squares space-time finite element methods”, presented at Graz University of Technology, Institute of Applied Mathematics in 2025.

Supervision / Assessment:

Olaf Steinbach (Graz University of Technology)
Leszek Demkowicz (University of Texas at Austin)
Manfred Kaltenbacher (Graz University of Technology)

Cover	Verlag der Technischen Universität Graz
Cover photo	Vier-Spezies-Rechenmaschine
	by courtesy of the Gottfried Wilhelm Leibniz Bibliothek
	Niedersächsische Landesbibliothek Hannover
Printed by	Buchschmiede (DATAFORM Media GmbH)

2025 Verlag der Technischen Universität Graz

www.tugraz-verlag.at

Print
ISBN 978-3-99161-060-1

E-Book
ISBN 978-3-99161-061-8
DOI 10.3217/978-3-99161-060-1

Abstract

In this thesis we consider a minimal residual/least-squares approach for the solution of an abstract operator equation. In particular, the first objective is to apply a least-squares method in combination with a space-time discretization scheme for the solution of parabolic evolution equations. Moreover, a second objective is to consider the approach for the simulation of electric machines.

In order to take into account the different types of partial differential equations (PDEs) we present in a first step an abstract minimal residual framework together with a complete stability and a priori error analysis. Furthermore, it is shown that under a saturation assumption the method obtains an efficient and reliable error indicator, which can be used to drive an adaptive refinement scheme.

In a second step, we apply the least-squares framework together with a space-time discretization scheme to several parabolic PDEs, including the heat equation, the convection-diffusion equation and a heat equation with nonlinear reaction term. Several numerical examples are presented confirming our theoretical findings. In addition to that we use the inbuilt error estimator in an adaptive refinement scheme resulting in space-time meshes which are completely unstructured with respect to space and time.

Finally, we apply the method to the simulation of an electric machine. Here we consider a two-dimensional spatial domain and different models, including the magnetostatic approximation, a quasistatic approximation and the eddy current problem. In the latter two models we exploit a space-time discretization in order to consider the movement of the rotor within the mesh. Numerical examples are presented which demonstrate the correctness of the proposed method.

Zusammenfassung

In dieser Arbeit betrachten wir die Methode der kleinsten Fehlerquadrate zur Lösung einer abstrakten Operatorgleichung. Einerseits liegt der Schwerpunkt auf der Anwendung der Methode in Kombination mit einem Raum-Zeit Diskretisierungsverfahren zur Lösung parabolischer partieller Differentialgleichungen. Zum anderen soll dieser Ansatz für die Berechnung elektromagnetischer Felder am Elektromotor eingesetzt werden.

Um die dabei auftretenden partiellen Differentialgleichungen mit dem selben Zugang behandeln zu können, wird zunächst ein abstraktes Konzept vorgestellt. Im Rahmen dieses Konzepts wird eine vollständige Stabilitäts-und Fehleranalyse durchgeführt. Mithilfe einer Saturationsannahme wird gezeigt, dass diese Methode einen Fehlerschätzer besitzt. Dieser kann in einem adaptiven Verfahren zur Netzverfeinerung verwendet werden.

Als nächstes wird die Anwendung dieser Methode auf parabolische Evolutionsprobleme demonstriert. Dabei wird ein Raum-Zeit Diskretisierungsverfahren verwendet. Insbesondere werden die Wärmeleitungsgleichung, die Konvektions- und Diffusionsgleichung und eine Wärmeleitungsgleichung mit nichtlinearem Reaktionsterm behandelt. Die theoretischen Ergebnisse werden durch numerische Beispiele bestätigt und der eingebaute Fehlerschätzer wird in einem adaptiven Verfahren zur Netzverfeinerung verwendet. Damit ergeben sich völlig unstrukturierte Zerlegungen des Gebietes in Raum und Zeit.

Abschließend wird dieser Ansatz für die Berechnung elektromagnetischer Felder am Elektromotor diskutiert. Als physikalisches Modell der Maschine werden die Gleichungen der Magnetostatik, eine quasistatische Erweiterung dieses Modells und das Wirbelstromproblem betrachtet. In den beiden letztgenannten Modellen erlaubt uns eine Raum-Zeit Diskretisierung die Bewegung des Rotors in der Triangulierung des Raum-Zeit Zylinders abzubilden. Numerische Beispiele zeigen eine korrekte Berechnung des magnetischen Feldes.

Acknowledgements

First, I would like to express my gratitude to my supervisor Prof. Olaf Steinbach for giving me the opportunity to do my PhD within the SFB CREATOR in the research project *C04: Space-Time Finite Element Methods*. Moreover, I would like to thank him for guiding this thesis and for all his support during the years.

I would like to thank Prof. Leszek Demkowicz and Prof. Manfred Kaltenbacher for agreeing to review this thesis.

Further, I would like to acknowledge the support by the Austrian Science Fund (FWF) under the Grant Collaborative Research Centre TRR361/90: CREATOR Computational Electric Machine Laboratory.

Finally, I would like to thank all my colleagues at the Institute of Applied Mathematics for the nice working atmosphere. In particular, many thanks to Assoc.Prof. Günther Of and to Richard, Mario and Michael for all the scientific discussions, for the exchange of ideas and also for the social support.

Contents

1 Introduction

Many phenomena in engineering science and natural science are described in terms of partial differential equations (PDEs). Prominent examples are Maxwell's equations to model electromagnetic phenomena, Stokes and Navier-Stokes equations to model fluid problems, the heat equation to model heat transfer problems or the equations of elasticity to model structural mechanical problems, just to mention a few. For the numerical solution of PDEs different numerical schemes like e.g., finite volume [86], finite difference [86], boundary elements [156] or finite elements [45] are used. We focus in this thesis on the latter approach. The finite element method (FEM) has been established as the standard numerical method for the solution of PDEs describing physical fields, see [102]. For an efficient and accurate numerical simulation adaptive finite elements are of great interest. One reason for this is as mentioned in [170] that the overall accuracy of the numerical solution deteriorates in the presence of local singularities. A remedy for this problem is to refine the mesh around the singularity. However, the question is how to detect the regions which have to be refined. Another reason is that one is interested in reliable estimates for the accuracy of the computed numerical solution since a priori error estimates only provide information about the asymptotic behaviour of the error [170]. The tool to address both issues are local error indicators which can be determined a posteriori from the numerical solution and the given data. In general the local error indicators are used in an adaptive algorithm which has the form, cf. [37],

$$\text{SOLVE} \rightarrow \text{ESTIMATE} \rightarrow \text{MARK} \rightarrow \text{REFINE}. \tag{1.1}$$

In more detail this means that first the PDE is solved, and then a local error indicator is used to estimate the error. Well-known error estimators are e.g. residual based error estimators [11], hierachical based error estimators [14], estimators based on the gradient recovery technique [181] or goal oriented error estimators [13], just to name a few. Afterwards the elements are selected according to some marking strategy. Possible strategies are e.g. the Dörfler marking strategy [58] or a bulk criterion like the maximum marking strategy, see [73, 129]. The marked elements are then refined via e.g. a bisection algorithm [9, 121, 163, 167] or red-green-blue refinement [34, 73]. In literature this is also referred to as h-adaptivity, since it locally decreases the mesh size h, cf. [146]. For more information about adaptive finite element methods we refer the interested reader to the textbooks [4, 13, 139, 171]. A natural question which arises is whether the numerical solutions computed from the adaptive scheme in (1.1) converge to the true solution. We refer to the work [37] which describes

an axiomatic framework to prove optimal rates for adaptive finite element methods which builds upon the earlier works [20, 58, 128, 162] as mentioned in [146].

Least-squares methods

In this thesis we focus on least-squares or minimal residual finite element methods [21]. These kinds of methods obtain an inbuilt error estimator. The idea behind these methods is to exploit the connection of finite element methods and energy minimization principles, cf. [21]. For elliptic problems this connection is given e.g. in terms of the Rayleigh-Ritz principle, where one seeks for a minimizer to an unconstrained quadratic convex functional. The functional describes the energy and is given via the variational formulation. In order to generalize this idea to arbitrary PDEs one considers an artificial energy, namely the energy induced by the residual of the PDE, cf. [21]. To put this in a more mathematical context, we consider Hilbert spaces X, Y, a given right-hand side $f \in Y^*$ in the dual of Y and a linear operator $B : X \to Y^*$, which is assumed to be an isomorphism and given in terms of the partial differential equation. The standard finite element method is based on an operator equation, where one seeks for $u \in X$ such that

$$Bu = f \quad \text{in } Y^*. \tag{1.2}$$

In a least-squares method one considers the equivalent minimization problem

$$u = \arg\min_{w \in X} \frac{1}{2} \|Bw - f\|_{Y^*}^2. \tag{1.3}$$

The least-squares method seems at first glance challenging as the residual is measured in a dual norm and hence cannot be computed locally to define an error estimator. Furthermore, it is mentionend in [21] that a least-squares finite element method obtained from (1.3) should also be practical. This means that the finite element spaces used in a discretization scheme for (1.3) should not be more difficult to work with than those normally encountered in typical Galerkin or mixed Galerkin methods for (1.2). Also, the finite element matrices and the right-hand side in the least-squares method should be easily computable. Therefore, one common approach is to rewrite the differential operator $B : X \to Y^*$ by a first order system operator $\mathcal{B} : U \to V$, where V is a space with L^2-regularity. This is usually referred to as first order system least-squares (FOSLS) methods, see e.g. [32]. This method has been extensively studied, see e.g., the textbook [21] for an overview of basic results on least-squares finite element methods or the works [2, 17] which deal with adaptivity in the context of FOSLS and the references given therein. FOSLS methods for parabolic equations are considered in e.g. [72, 81, 119, 120]. A subclass of the FOSLS familiy are so-called constrained first order system least-squares methods (CFOSLS). We mention

the works [3, 132, 147, 172]. However, the reformulation of the PDE as a first order system comes with the fact that one has to assume that the right-hand side f has L^2-regularity. This may be in practical applications not the space of interest and is to some extent unnatural as right-hand sides of partial differential equations are usually considered in Sobolev spaces with negative order, cf. [106]. Although there are some workarounds, see [70, 71], where a FOSLS method with a load in a Sobolev space with negative order is analyzed by replacing the source term with its finite element approximation, we will for the mentioned reasons consider in this thesis a different approach. We prefer to stay in the norm induced by the operator B. In order to get a practical method we use the Riesz operator $A : Y \to Y^*$ to lift the abstract norm in Y^* to a norm in Y, which can be evaluated locally on each finite element. This setting allows us to derive a unified framework, which can be applied to various PDEs. This approach has been studied in e.g. [6, 164] in the context of parabolic equations, [46] in connection with convection-diffusion equations, [28, 36, 54, 51, 52, 53] in case of Discontinuous Petrov Galerkin (DPG) methods or [47, 127], just to name a few.

Space-time methods

Of particular interest in this thesis are space-time discretization schemes which are used for the numerical solution of time-dependent partial differential equations. As mentioned in e.g. [118, 146, 175], the standard procedures for the numerical solution of time-dependent problems are the vertical method of lines and the horizontal method of lines or Rothe's method, see [142, 166]. In the vertical method of lines one first discretizes with respect to the spatial variables with e.g. a finite element method. The semidiscretization in space leads to a system of ordinary differential equations which can be solved via some time stepping scheme, see [87, 88] for an overview about time-stepping schemes. In Rothe's method one performs the two semidiscretizations steps the other way round. This means first one performs a discretization with respect to time with some time stepping scheme. The resulting boundary value problems are then discretized e.g. by means of a finite element method. The fast solution of such semidiscretization schemes lead to the so called parallel-in-time methods. A good overview over the history of such methods is given by Gander, see [74].

We will focus on space-time methods where time is treated just like an additional spatial variable. This approach has gained a lot of interest recently and an overview about recent advances is given in [161, 113]. A classification of space-time methods can also be found in [146, 175]. A full space-time method has the advantage that it allows for adaptivity with respect to space and time simultaneously, see e.g., [159, 160] for a space-time method in the sense of [157], [72] in case of a FOSLS method or [110] for a space-time finite element method based on upwind stabilization. Moreover, the handling of moving domains, where the movement is a priori known, is easier as

a moving spatial d-dimensional domain viewed in the $d+1$-dimensional space-time cylinder is just a picture, i.e., it can be meshed using standard tools. This fact is exploited in [77, 78, 83] in case of a rotating electric motor. In addition to that it is more flexible with respect to parallelization in space and time as it overcomes the sequential behaviour of the standard semidiscretizations schemes. Also in the context of optimal control problems constrained by time-dependent PDEs full space-time discretizations are of interest as they allow for a simultaneous solution of the forward and backward problem at once, see e.g., [16, 111, 112] for a related approach. This is in contrast to time stepping schemes which require to first step forward in time and then step backward in time, see also [118]. The treatment of a time-dependent PDE on a $d+1$-dimensional domain comes with the fact that a global linear system must be solved at once, which also increases the memory demand, see also [146, 175]. Therefore, fast solvers and preconditioning are essential which are not within the scope of thesis, see e.g., [75, 114] for related approaches in case of space-time tensor-product meshes.

Objective and outline of the thesis

The objective of this thesis is to combine the least-squares/minimal residual method with a fully unstructured space-time discretization scheme in the spirit of [157]. Moreover, a second objective is to apply the minimal residual method for the simulation of electric machines based on the Maxwell system. Here we will consider the magnetostatic approximation, i.e., a time-independent problem, as well as a quasistatic approximation and the eddy current problem, i.e., time-dependent problems. In the latter two models, a space-time discretization enables us to consider the movement of the rotor within the mesh. In order to account for the different PDEs, we will present an abstract unified least-squares approach based on [106]. This framework can be applied to various linear and nonlinear PDEs.

The remainder of this thesis is organized as follows: In Chapter 2 we introduce the required function spaces, repeat basic functional analytic results on the solvability of abstract operator equations and recall basic notions on finite element discretizations. In Chapter 3 we present an abstract least-squares framework for linear problems including a full stability and error analysis. Furthermore, we show that the inbuilt error estimator is efficient and reliable. We also discuss an extension to nonlinear problems. In Chapter 4 we apply the abstract theory to the heat equation, to a convection-diffusion equation and to a semilinear parabolic equation, which has the form of the Schlögl model, see e.g., [39, 151]. For the heat equation we also present a comparison to the FOSLS method introduced by Führer and Karkulik in [72]. Finally, in Chapter 5 we demonstrate the application of the least-squares method for the simulation of an electric machine.

2 Preliminaries

In this chapter we introduce the required function spaces and recall some well-known results on the existence and uniqueness of solutions of variational formulations, which we will use throughout this thesis.

Throughout the whole thesis we consider $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ to be a domain (open, connected, non-empty), which is bounded and its boundary $\Gamma := \partial\Omega$ is Lipschitz, see [156, Def. 2.1]. Further, let $T > 0$ be a given final time horizon. The corresponding space-time cylinder is then denoted by $Q := \Omega \times (0, T)$ with the boundaries $\Sigma := \partial\Omega \times (0, T)$, $\Sigma_0 := \overline{\Omega} \times \{0\}$ and $\Sigma_T := \overline{\Omega} \times (0, T)$ such that $\partial Q = \Sigma \cup \Sigma_0 \cup \Sigma_T$.

2.1 Function spaces

We give a brief introduction to the function spaces used in this thesis. For a more detailed presentation we refer to classical textbooks like e.g., [1, 24, 66, 123, 180]. A good source for spaces used in variational formulations for space-time methods is the thesis [175] and the references given therein.

2.1.1 Sobolev spaces

Let $d \in \mathbb{N} := \{1, 2, 3 \dots\}$. We call a vector $\alpha \in \mathbb{N}_0^d$ with components $\alpha_i \in \mathbb{N}_0 := \{0, 1, 2, 3, \dots\}$ multi index. The length $|\alpha|$ and the factorial $\alpha!$ are given as

$$|\alpha| := \sum_{i=1}^{d} \alpha_i, \quad \alpha! := \alpha_1! \alpha_2! ... \alpha_d!.$$

For $x \in \Omega$ we write

$$x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} ... x_d^{\alpha_d}$$

and for a smooth function $u : \Omega \to \mathbb{R}$ the notation $D^\alpha u$ means

$$D^\alpha u(x) := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_d^{\alpha_d}} u(x).$$

For $k \in \mathbb{N}_0$ the space of k times continuously differentiable functions reads

$$C^k(\Omega) = \{u : \Omega \to \mathbb{R} : \ D^\alpha u \in C(\Omega), \text{ for all } |\alpha| \leq k\},$$

and the corresponding norm is

$$\|u\|_{C^k(\Omega)} := \sum_{|\alpha|\leq k} \sup_{x\in\Omega} |D^\alpha u(x)|.$$

Note that for $k = 0$, $C(\Omega) := C^0(\Omega)$ corresponds to the space of continuous functions. The space of inifinitely often differentiable functions is given by

$$C^\infty(\Omega) := \bigcap_{k\geq 0} C^k(\Omega).$$

For a function $u : \Omega \to \mathbb{R}$ the support is defined as

$$\operatorname{supp} u := \overline{\{x \in \Omega : \ u(x) \neq 0\}}$$

and the space of infinitely often differentiable functions with compact support reads

$$C_0^\infty(\Omega) := \{u \in C^\infty(\Omega) : \ \operatorname{supp} u \Subset \Omega\}.$$

For $1 \leq p < \infty$ we define the space of p-integrable functions as

$$\mathcal{L}^p(\Omega) := \left\{ u : \Omega \to \mathbb{R} \text{ measurable} : \ \int_\Omega |u(x)|^p \mathrm{d}x < \infty \right\}$$

and for $p = \infty$ we have

$$\mathcal{L}^\infty(\Omega) := \left\{ u : \Omega \to \mathbb{R} \text{ measurable} : \ \operatorname*{ess\,sup}_{x\in\Omega} |u(x)| < \infty \right\}.$$

We call the factor space

$$L^p(\Omega) := \mathcal{L}^p(\Omega)/\sim, \quad u \sim v :\Leftrightarrow u = v \text{ almost everywhere (a.e.)}$$

Lebesgue space of p-integrable functions. Equipped with the norms

$$\|u\|_{L^p(\Omega)} := \left(\int_\Omega |u(x)|^p \mathrm{d}x \right)^{\frac{1}{p}}, \ 1 \leq p < \infty,$$
$$\|u\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{x\in\Omega} |u(x)|$$

it can be shown, see e.g. [1, Thm. 2.16], [24, Thm. 1.1.8], that these spaces are Banach spaces. For $p = 2$ one can even show that the space $L^2(\Omega)$ is a Hilbert space with respect to the inner product

$$\langle u, v\rangle_{L^2(\Omega)} := \int_\Omega u(x)v(x) \ \mathrm{d}x$$

for $u, v \in L^2(\Omega)$, see [1, Cor. 2.18]. Further we denote with $L^p_{loc}(\Omega)$ the Lebesgue space of functions, which are p-integrable on each compact set $K \subset \Omega$, i.e.,

$$L^p_{loc}(\Omega) = \{u : \Omega \to \mathbb{R} \text{ measurable} : u \in L^p(K) \ \forall K \subset \Omega \text{ compact}\}.$$

For the introduction of Sobolev spaces we need the concept of a weak derivative.

DEFINITION 2.1 (cf. [1, 1.62, p.21], [24, Def. 1.2.4], [156, Def. 2.3]). *Let $u \in L^1_{loc}(\Omega)$ and $\alpha \in \mathbb{N}_0^d$ be a multi index. We say that u admits a weak derivative of order α if there exists a function $v \in L^1_{loc}(\Omega)$ satisfying*

$$\int_\Omega v(x)\varphi(x)\,\mathrm{d}x = (-1)^{|\alpha|}\int_\Omega u(x)D^\alpha\varphi(x)\,\mathrm{d}x \quad \textit{for all } \varphi \in C_0^\infty(\Omega).$$

If such a function v exists, we write $D^\alpha u := v$.

Now, we can for $p \geq 1$ and $k \in \mathbb{N}_0$ define the Sobolev spaces

$$\begin{aligned} W^{k,p}(\Omega) &:= \{u \in L^p(\Omega) :\ D^\alpha u \in L^p(\Omega),\ |\alpha| \leq k\}, \\ W_0^{k,p}(\Omega) &:= \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{W^{k,p}(\Omega)}} \end{aligned} \tag{2.1}$$

with the corresponding norms

$$\begin{aligned} \|u\|_{W^{k,p}(\Omega)} &:= \left(\sum_{|\alpha|\leq k} \|D^\alpha u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}} \quad \text{for } 1 \leq p < \infty, \\ \|u\|_{W^{k,\infty}(\Omega)} &:= \max_{0\leq|\alpha|\leq k} \|D^\alpha u\|_{L^\infty(\Omega)}, \end{aligned}$$

and the seminorms

$$\begin{aligned} |u|_{W^{k,p}(\Omega)} &:= \left(\sum_{|\alpha|=k} \|D^\alpha u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}} \quad \text{for } 1 \leq p < \infty, \\ |u|_{W^{k,\infty}(\Omega)} &:= \max_{|\alpha|=k} \|D^\alpha u\|_{L^\infty(\Omega)}. \end{aligned}$$

It holds that $(W^{k,p}(\Omega), \|\cdot\|_{W^{k,p}(\Omega)})$ is a Banach space, see [24, Thm. 1.3.2], [1, Thm. 3.3] and that $C^\infty(\Omega) \cap W^{k,p}(\Omega)$ is dense in $W^{k,p}(\Omega)$, see [24, Thm. 1.3.4] which goes back to Meyers and Serin [126]. Note that we have the inclusion [1, p. 60]

$$W_0^{k,p}(\Omega) \subset W^{k,p}(\Omega) \subset L^p(\Omega).$$

Moreover, we have by [1, Thm. 3.6] that $W^{k,p}(\Omega)$ is separable if $1 \leq p < \infty$, and uniformly convex as well as reflexive if $1 < p < \infty$. For $p = 2$ we use the notation

$$H^k(\Omega) := W^{k,2}(\Omega), \quad H_0^k(\Omega) := W_0^{k,2}(\Omega)$$

which are separable Hilbert spaces with inner product

$$\langle u, v\rangle_{H^k(\Omega)} := \sum_{0\leq|\alpha|\leq k} \langle D^\alpha u, D^\alpha v\rangle_{L^2(\Omega)} \tag{2.2}$$

for $u, v \in H^k(\Omega)$. The definition of the Sobolev spaces (2.1) can be extended to $0 < s \in \mathbb{R}$. For this we write $s = k + \sigma$, where $k \in \mathbb{N}_0$ and $\sigma \in (0,1)$. Then we can define the spaces

$$\begin{aligned} W^{s,p}(\Omega) &:= \left\{ u \in W^{k,p}(\Omega) : \ |u|_{W^{s,p}} < \infty \right\}, \\ W_0^{s,p}(\Omega) &:= \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{W^{s,p}(\Omega)}} \end{aligned} \tag{2.3}$$

with the corresponding norm

$$\|u\|_{W^{s,p}} = \left(\|u\|_{W^{k,p}}^p + |u|_{W^{s,p}(\Omega)}^p \right)^{\frac{1}{p}},$$

where

$$|u|_{W^{s,p}(\Omega)}^p := \sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{|x-y|^{d+p\sigma}} \mathrm{d}x\mathrm{d}y$$

denotes the Sobolev-Slobodeckii seminorm, see [123, p.74]. Again we use for $p = 2$ the notation

$$H^s(\Omega) := W^{s,2}(\Omega), \quad H_0^s(\Omega) := W_0^{s,2}(\Omega),$$

which are Hilbert spaces with the inner product (2.2) for $s = k \in \mathbb{N}_0$ and

$$\langle u, v \rangle_{H^s(\Omega)} := \langle u, v \rangle_{H^k(\Omega)} + \sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{(D^\alpha u(x) - D^\alpha u(y))(D^\alpha v(x) - D^\alpha v(y))}{|x-y|^{d+2\sigma}} \mathrm{d}x\mathrm{d}y$$

for $s = k + \sigma$ with $k \in \mathbb{N}_0$ and $\sigma \in (0,1)$. For $s > 0$ we define, see e.g. [64, Def. B47], the space $H^{-s}(\Omega)$ as dual space of $H_0^s(\Omega)$ equipped with the norm

$$\|f\|_{H^{-s}(\Omega)} := \sup_{0 \neq v \in H_0^s(\Omega)} \frac{\langle f, v \rangle_\Omega}{\|v\|_{H^s(\Omega)}},$$

where $\langle \cdot, \cdot \rangle_\Omega$ denotes the duality paring as the extension of the L^2-inner product. Of special interest in our analysis will be the space $H_0^1(\Omega)$. In this space there holds due to the assumptions on the domain Ω the Poincaré inequality, see [64, Lem. B.61], i.e., there exists a constant $c_p > 0$ such that

$$\|u\|_{L^2(\Omega)} \leq c_p \|\nabla u\|_{[L^2(\Omega)]^d} \quad \text{for all } u \in H_0^1(\Omega). \tag{2.4}$$

As a consequence we have that

$$\|u\|_{H_0^1(\Omega)} := \|\nabla u\|_{[L^2(\Omega)]^d} = \sqrt{\sum_{|\alpha|=1} \|D^\alpha u\|_{L^2(\Omega)}^2}$$

defines an equivalent norm on $H_0^1(\Omega)$, which we will frequently use. Note that instead of the vector valued space $[L^2(\Omega)]^d$ we will for better readability simply write $L^2(\Omega)$ in the following as it is clear from the context that it has to be understood as a vector valued space. Further we have the inclusion, see e.g. [66, Sec. 5.9.1, p. 299]

$$H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega).$$

2.1.2 Bochner spaces

For the analysis of the differential operators given by parabolic evolution equations we use the concept of Bochner Sobolev spaces. Here a function $u : Q \to \mathbb{R}$, $(x,t) \mapsto u(x,t)$ is considered as a function of t with values in a Banach space X, i.e., $t \mapsto u(t,\cdot) \in X$. For the definition of these spaces we proceed similar as in Section 2.1.1.

Let $(X, \|\cdot\|_X)$ be a Banach space. For $p \in [1,\infty)$ we define the space of p-integrable vector valued functions as

$$\mathcal{L}^p(0,T;X) := \left\{ u : (0,T) \to X \text{ measurable} : \int_0^T \|u(t)\|_X^p \,\mathrm{d}t < \infty \right\}$$

and for $p = \infty$

$$\mathcal{L}^\infty(0,T;X) := \left\{ u : (0,T) \to X \text{ measurable} : \operatorname*{ess\,sup}_{t\in(0,T)} \|u(t)\|_X < \infty \right\}.$$

We call the factor space

$$L^p(0,T;X) := \mathcal{L}^p(0,T;X)/\sim, \quad u \sim v :\Leftrightarrow u = v \text{ a.e.}$$

Bochner space of p-integrable functions. If we equip these spaces with the norms

$$\begin{aligned} \|u\|_{L^p(0,T;X)} &:= \left(\int_0^T \|u(t)\|_X^p \,\mathrm{d}t \right)^{\frac{1}{p}}, \quad 1 \le p < \infty, \\ \|u\|_{L^\infty(0,T;X)} &:= \operatorname*{ess\,sup}_{t\in(0,T)} \|u(t)\|_X \end{aligned}$$

one can show that they are Banach spaces, see [176, Prop. 23.2, Problem 23.12], [143, Satz 1.24]. For a reflexive Banach space X, $1 \le p < \infty$ and $q > 0$ such that $\frac{1}{p} + \frac{1}{q} = 1$, the mapping

$$\begin{aligned} R : L^q(0,T;X^*) &\to (L^p(0,T;X))^* \\ \langle Ru, v\rangle &:= \int_0^T \langle u(t), v(t)\rangle_{X^*\times X} \,\mathrm{d}t \quad \text{for all } v \in L^p(0,T;X) \end{aligned}$$

is an isometric isomorphism, see e.g. [143, Satz 1.30]. Hence, we can identify the dual $(L^p(0,T;X))^*$ with $L^q(0,T;X^*)$. Moreover, there holds, see [143, Bsp. 1.26]

$$L^p(Q) = L^p(0,T;L^p(\Omega)).$$

We denote with $L^p_{loc}(0,T;X)$ the Bochner space of functions, which are p-integrable on each compact set $K \subset (0,T)$. Similar as for the classical Sobolev spaces we need the concept of a weak derivative to define the Bochner Sobolev spaces.

DEFINITION 2.2 (cf. [66, Sec. 5.9.2, p. 301], [48, Def. 4.17]). *Let $(X, \|\cdot\|)$ be a Banach space and $u \in L^1_{loc}(0,T;X)$. We say the function u obtains a weak derivative of order j if there exists a function $v \in L^1_{loc}(0,T;X)$ such that*

$$\int_0^T v(t)\varphi(t)\,\mathrm{d}t = (-1)^j \int_0^T u(t)\partial_t^j \varphi(t) \quad \textit{for all } \varphi \in C_0^\infty(0,T).$$

If such a function v exists we write $\partial_t^j u := v$.

Now, we can define the Bochner Sobolev spaces

$$W^{k,p}(0,T;X) := \left\{ u \in L^p(0,T;X) : \ \partial_t^j u \in L^p(0,T;X) \text{ for } 0 \le j \le k \right\}.$$

For our analysis it is sufficient to consider the space $W^{1,p}(0,T;X)$. One can show that this space is a Banach space for $p \ge 1$ with respect to the norm

$$\|u\|_{W^{1,p}(0,T;X)} := \left(\|u\|^p_{L^p(0,T;X)} + \|\partial_t u\|^p_{L^p(0,T;X)} \right)^{\frac{1}{p}},$$

cf. [8, p. 3], [48, Thm. 5.2], [66, Sec. 5.9.2, p. 203]. For $p = 2$ we use the notation

$$H^k(0,T;X) := W^{k,2}(0,T;X).$$

For $1 \le p \le \infty$ we have that $W^{1,p}(0,T;X)$ is continuously embedded into the space $C([0,T]\,;X)$, i.e., $W^{1,p}(0,T;X) \subset C([0,T]\,;X)$ and there exists a constant $C > 0$ such that

$$\|u\|_{C([0,T];X)} \le C\|u\|_{W^{1,p}(0,T;X)},$$

see [66, Sec. 5.9.2, Thm.2]. Here $C([0,T]\,;X)$ denotes the space of continuous functions on the intervall $[0,T]$ with values in the Banach space X which is equipped with the norm

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

This result ensures that initial condtions $u(0)$ and terminal conditions $u(T)$ are well-defined. We use the notations

$$W^{1,p}_{0,}(0,T;X) := \left\{ u \in W^{1,p}(0,T;X) : \ u(0) = 0 \text{ in } X \right\},$$
$$W^{1,p}_{,0}(0,T;X) := \left\{ u \in W^{1,p}(0,T;X) : \ u(T) = 0 \text{ in } X \right\}$$

to indicate spaces with zero initial or terminal conditions, respectively. Of particular interest in our analysis are the spaces

$$L^2(0,T;H^1_0(\Omega)) \quad \text{and} \quad W(Q) := L^2(0,T;H^1_0(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)).$$

The Bochner space $L^2(0,T;H_0^1(\Omega))$ allows the characterization, see [154, p. 196, p. 204], [175, p. 23]

$$L^2(0,T;H_0^1(\Omega)) = \left\{u \in L^2(Q) : \nabla_x u \in L^2(Q),\ u|_\Sigma = 0\right\}. \tag{2.5}$$

Moreover, it is a Hilbert space with respect to the inner product

$$\begin{aligned}\langle p, q\rangle_{L^2(0,T;H_0^1(\Omega))} &:= \int_0^T \langle p(t), q(t)\rangle_{H_0^1(\Omega)} \mathrm{d}t \\ &= \int_0^T \int_\Omega \nabla_x p(x,t) \cdot \nabla_x q(x,t)\,\mathrm{d}x\mathrm{d}t = \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)}\end{aligned}$$

for all $p, q \in L^2(0,T;H_0^1(\Omega))$, see [175]. Note that the second to last equality is due to (2.5). The dual space $(L^2(0,T;H_0^1(\Omega)))^*$ is identified with $L^2(0,T;H^{-1}(\Omega))$ and there holds due to (2.4) the Poincaré inequality

$$\|u\|_{L^2(Q)} \le c_p \|\nabla_x u\|_{L^2(Q)} \quad \text{for all } u \in L^2(0,T;H_0^1(\Omega)). \tag{2.6}$$

The space $W(Q)$ in more detail reads

$$W(Q) = \left\{u \in L^2(0,T;H_0^1(\Omega)) :\ \partial_t u \in L^2(0,T;H^{-1}(\Omega))\right\}.$$

It is a Hilbert space with inner product, see [175], cf. [49, Chap. XVIII, §1, Prop.6]

$$\langle u, v\rangle_{W(Q)} := \langle u, v\rangle_{L^2(0,T;H_0^1(\Omega))} + \int_0^T \langle \partial_t u(t), \partial_t v(t)\rangle_{H^{-1}(\Omega)}\,\mathrm{d}t$$

and the corresponding norm reads

$$\|u\|_{W(Q)} = \left(\|\nabla_x u\|^2_{L^2(Q)} + \|\partial_t u\|^2_{L^2(0,T;H^{-1}(\Omega))}\right)^{\frac{1}{2}}.$$

Further we have that $W(Q)$ is continuously embedded into the space $C([0,T]\,;L^2(\Omega))$, see [66, Thm. 3, p. 303], [143, Lem. 2.44], which means that initial and terminal conditions are well-defined and there holds the integration by parts formula [143, Lem. 2.44]

$$\begin{aligned}&\int_0^T \langle \partial_t u(t), v(t)\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}\,\mathrm{d}t \\ &\qquad = \langle u(T), v(T)\rangle_{L^2(\Omega)} - \langle u(0), v(0)\rangle_{L^2(\Omega)} - \int_0^T \langle u(t), \partial_t v(t)\rangle_{H_0^1(\Omega)\times H^{-1}(\Omega)}\,\mathrm{d}t\end{aligned}$$

Finally, by [66, Thm. 3, p.303] the mapping $t \mapsto \|u(t)\|^2_{L^2(\Omega)}$ is absolutely continuous and for the derivative we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2_{L^2(\Omega)} = 2\langle \partial_t u(t), u(t)\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \quad \text{for a.e. } t \in (0,T).$$

2.2 Variational methods

In this section we recall some well-known results on the solvability of abstract operator equations. We start with the Lemma of Lax-Milgam.

LEMMA 2.3 (Lax-Milgram). *Let V be a Hilbert space, let $A : V \to V^*$ be linear and let $f \in V^*$. Further, assume $A : V \to V^*$ to be bounded and elliptic, i.e., there exist constants $c_1^A, c_2^A > 0$ such that*

$$\langle Au, v\rangle_{V^*\times V} \leq c_2^A \|u\|_V \|v\|_V, \quad \langle Au, u\rangle_{V^*\times V} \geq c_1^A \|u\|_V^2$$

for all $u, v \in V$. Then, the variational problem:

$$\text{Find } u \in V : \ \langle Au, v\rangle_{V^*\times V} = \langle f, v\rangle_{V^*\times V} \quad \forall v \in V,$$

has a unique solution. For the unique solution $u \in V$ there holds the a priori estimate

$$\|u\|_V \leq \frac{1}{c_1^A} \|f\|_{V^*}.$$

Proof. A proof can be found in [64, Lemma 2.2]. □

An important property of elliptic operators is that they can be used to define equivalent norms.

LEMMA 2.4. *Let $(V, \langle ., .\rangle_V)$ be a Hilbert space and $A : V \to V^*$ be a linear, bounded, self-adjoint and elliptic operator. Then, for all $u, v \in V$*

$$\langle u, v\rangle_A := \langle Au, v\rangle_{V^*\times V}$$

defines an inner product on V and

$$\|u\|_A := \sqrt{\langle u, u\rangle_A} = \sqrt{\langle Au, u\rangle_{V^*\times V}}$$

defines an equivalent norm to $\|.\|_V$ which satisfies

$$\sqrt{c_1^A}\|u\|_V \leq \|u\|_A \leq \sqrt{c_2^A}\|u\|_V \quad \text{for all } u \in V.$$

Proof. A proof can be found in [118, Lem. 2.10]. □

The next theorem generalizes the Lemma of Lax-Milgram.

THEOREM 2.5 (Banach-Nečas-Babuška). *Let X be a Banach space and Y be a reflexive Banach space. Further, let $f \in Y^*$ and $B : X \to Y^*$ be linear and bounded, i.e., there exists a constant $c_2^B > 0$ such that*

$$\|Bu\|_{Y^*} \leq c_2^B \|u\|_X \quad \forall u \in X.$$

The variational problem:

$$\text{Find } u \in X : \ \langle Bu, v\rangle_{Y^* \times Y} = \langle f, v\rangle_{Y^* \times Y} \quad \forall v \in Y,$$

has a unique solution if and only if the conditions

BNB 1

$$\exists\, c_1^B > 0 : \ \inf_{0 \neq u \in X} \sup_{0 \neq v \in Y} \frac{\langle Bu, v\rangle_{Y^* \times Y}}{\|u\|_X \|v\|_Y} \geq c_1^B,$$

BNB 2

$$\forall v \in Y, \ (\ \forall u \in X, \langle Bu, v\rangle_{Y^* \times Y} = 0\ ) \Rightarrow v = 0,$$

are fulfilled. Moreover, for the unique solution $u \in X$ there holds the a priori estimate

$$\|u\|_X \leq \frac{1}{c_1^B} \|f\|_{Y^*}.$$

Proof. A proof can be found in [64, Theorem 2.6]. □

REMARK 2.6. *The condition (BNB 1) ensures that the operator B is bounded from below, i.e., it is injective, and its range is closed, cf. [64, Lem. A.39, Cor. A.45] and [134, p. 440]. The condition (BNB 2) gives surjectivity of B, i.e., $ran(B) = Y^*$, see [64, Cor. A.45, Cor. A.46]. A slightly generalized version of Thm. 2.5 can be found e.g. in [134, Thm. 6.6.1]. There if and only if is replaced by if and the condition (BNB 2) is exchanged by the assumption that $f \in \mathrm{ran}(B)$. This means that $f \in \ker(B^*)^\perp$ due to the Closed Range Theorem, see e.g. [134, Thm. 5.17.3]. Hence, this leads to a compatibility condition on f which reads*

$$\langle f, q\rangle_H = 0 \quad \textit{for all } q \in \ker(B^*).$$

The next theorem gives an existence and uniqueness result for the solution of abstract saddle point systems.

THEOREM 2.7 ([64, cf. Thm. 2.34]). *Let X, Y be two reflexive Banach spaces and $f \in Y^*$, $g \in X^*$ be two given right-hand sides. Further, let $A : Y \to Y^*$, $B : X \to Y^*$ be linear and bounded operators. The saddle point system which seeks for $(u, p) \in X \times Y$ such that*

$$\begin{aligned} \langle Ap, q\rangle_{Y^* \times Y} + \langle Bu, q\rangle_{Y^* \times Y} &= \langle f, q\rangle_{Y^* \times Y} \quad \textit{for all } q \in Y, \\ \langle B^* p, v\rangle_{X^* \times X} &= \langle g, v\rangle_{X^* \times X} \quad \textit{for all } v \in X \end{aligned} \tag{2.7}$$

has a unique solution if and only if the conditions

(i)

$$\exists\, c_1^A > 0 : \inf_{0\neq p\in\ker(B^*)} \sup_{0\neq q\in\ker(B^*)} \frac{\langle Ap, q\rangle_{Y^*\times Y}}{\|p\|_Y \|q\|_Y} \geq c_1^A,$$

(ii)

$$\forall q \in \ker(B^*),\ (\forall p \in \ker(B^*), \langle Ap, q\rangle_{Y^*\times Y} = 0) \Rightarrow q = 0,$$

(iii)

$$\exists c_1^B > 0 : \inf_{0\neq u\in X} \sup_{0\neq q\in Y} \frac{\langle Bu, q\rangle_{Y^*\times Y}}{\|u\|_X \|q\|_Y} \geq c_1^B$$

hold true. Moreover, for the unique solution $(u, p) \in X \times Y$ there hold the a priori estimates

$$\begin{aligned}
\|p\|_Y &\leq \frac{1}{c_1^A}\|f\|_{Y^*} + \frac{1}{c_1^B}\left(1 + \frac{c_2^A}{c_1^A}\right)\|g\|_{X^*},\\
\|u\|_X &\leq \frac{1}{c_1^B}\left(1 + \frac{c_2^A}{c_1^A}\right)\|f\|_{Y^*} + \frac{c_2^B}{[c_1^B]^2}\left(1 + \frac{c_2^A}{c_1^A}\right)\|g\|_{X^*}.
\end{aligned}$$

REMARK 2.8. *The conditions (i) and (ii) in Theorem 2.7 ensure that the operator A is an isomorphism on* $\ker(B^*)$. *The condition (iii) on B ensures that the adjoint operator B^* is surjective, see [64, Lemma A.39, Theorem A.56]. If the operator $A : Y \to Y^*$ is Y-elliptic conditions (i) and (ii) are clearly fulfilled, see also [64, Rem. 2.35]*

The next theorem is on the solvability of abstract evolution equations. For this, we first need the definition of an evolution triple which is also known as Gelfand triple.

DEFINITION 2.9 ([176, Def. 23.11]). *The triple (V, H, V^*) is called an evolution triple if*

(i) $(V, \|\cdot\|_V)$ *is a real separable and reflexive Banach space,*

(ii) $(H, (\cdot,\cdot)_H)$ *is a real separable Hilbert space,*

(iii) the embedding $V \hookrightarrow H$ is dense and continuous, i.e., $V \subset H$, $\overline{V}^{\|\cdot\|_H} = H$ and there exists a $c > 0$ such that $\|x\|_H \leq c\|x\|_V$ for all $x \in V$.

An example for such a triple is given by means of the spaces $(H_0^1(\Omega), L^2(\Omega), H^{-1}(\Omega))$, see e.g. [176, Def. 23.11] or [143, p. 89]. Now, we are in a position to state the following Theorem.

THEOREM 2.10. *Let the triple (V, H, V^*) be an evolution triple and the operator $A : L^2(0,T;V) \to L^2(0,T;V^*)$ be linear. Further, assume that A is bounded and elliptic, i.e., there exist constants $c_1^A, c_2^A > 0$ such that*

$$\begin{aligned} |\langle Au, v\rangle| &\leq c_2^A \|u\|_{L^2(0,T;V)} \|v\|_{L^2(0,T;V)}, \\ \langle Au, u\rangle &\geq c_1^A \|u\|^2_{L^2(0,T;V)} \end{aligned}$$

for all $u, v \in L^2(0,T;V)$, where $\langle ., . \rangle$ denotes the duality pairing. Suppose furthermore that $F \in L^2(0,T;V^*)$ and $u_0 \in H$. Then, the initial value problem

$$\partial_t u + Au = F \quad \text{in } L^2(0,T;V^*), \quad u(0) = u_0,$$

has a unique solution $u \in L^2(0,T;V)$ with weak derivative $\partial_t u \in L^2(0,T;V^*)$.

Proof. A proof can be found for instance in [64, Theorem 6.6] or [176, Theorem 23A, Corollary 23.24]. □

We will also deal with nonlinear operator equations. In order to investigate such equations we state the following result which goes back to Zarantonello [177, pp. 503], [178, pp. 174].

THEOREM 2.11 ([177, Thm. 25B]). *Let $(V, (\cdot,\cdot)_V)$ be a real Hilbert space, $f \in V^*$ and $B : V \to V^*$ a nonlinear operator which is strongly monotone and Lipschitz continuous, i.e., there are numbers $c_1^B > 0$ and $c_2^B > 0$ such that for all $u, v \in V$*

$$\langle B(u) - B(v), u - v\rangle_{V^* \times V} \geq c_1^B \|u - v\|_V^2$$

and

$$\|B(u) - B(v)\|_{V^*} \leq c_2^B \|u - v\|_V.$$

Then the nonlinear operator equation

$$B(u) = f \quad \text{in } V^*$$

has a unique solution, which depends continuously on f. More precisely, it follows from $B(u_j) = f_j$, $j = 1, 2$ that

$$\|u_1 - u_2\|_V \leq \frac{1}{c_1^B} \|f_1 - f_2\|_{V^*}$$

and the inverse operator $B^{-1} : V^* \to V$ is Lipschitz continuous with constant $\frac{1}{c_1^B}$.

2.3 Discretization

In order to solve the operator equations and the related variational formulations numerically we use the finite element method (FEM). The idea of the finite element method is to consider conforming finite dimensional subspaces of the trial and test spaces which can be spanned in terms of basis functions with local support. The aim of this section is to recall some basic notions of finite element discretizations, introduce the finite element spaces and state some approximation properties of these spaces which we will use throughout the thesis. For a more detailed presentation we refer to [15, 24, 45, 60, 156]. In order to cover discretizations for a spatial bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with $d = 1, 2, 3$ as well as for a space-time cylinder $Q = \Omega \times (0, T) \subset \mathbb{R}^{d+1}$ we consider a domain $D \subset \mathbb{R}^n$ for $n = 1, 2, 3, 4$. In addition to that we consider domains D whose boundary is polygonal for $n = 2$, or polyhedral for $n = 3$ or polychoral for $n = 4$.

We start with the definition of an n-dimensional simplex $\tau \subset \mathbb{R}^n$.

Definition 2.12 ([60, cf. Def. 3.11]). *Let the points $z_1, ..., z_{n+1} \in \mathbb{R}^n$ be given such that the vectors $(z_j - z_1)_{j=2}^{n+1}$ are linearly independent. The interior of the convex hull $\tau = (\mathrm{conv}\{z_1, ..., z_{n+1}\})^\circ$ is called a (non-degenerate) n-dimensional simplex in $\mathbb{R}^n$. The corner points $z_1, ..., z_n$ are called nodes or vertices of τ. For $r \in \{0, ..., n-1\}$ and $\tilde{z}_1, ..., \tilde{z}_{r+1} \in \{z_1, ..., z_{n+1}\}$ such that the vectors $(\tilde{z}_j - \tilde{z}_1)_{j=2}^{r+1}$ are linearly independent we call $\tilde{\tau} = (\mathrm{conv}\{\tilde{z}_1, ..., \tilde{z}_{r+1}\})^\circ$ r-dimensional face of τ.*

The n-dimensional simplex $\tau \subset \mathbb{R}^n$ is an interval for $n = 1$, a triangle for $n = 2$, a tetrahedron for $n = 3$ or a pentatope [130] for $n = 4$. We call a decomposition $\mathcal{T}_h := \{\tau_1, ..., \tau_N\}$ of the domain $D \subset \mathbb{R}^n$ into nonoverlapping simplicial elements $\tau_\ell \subset \mathbb{R}^n$ a triangulation or mesh, i.e.,

$$\overline{D} = \bigcup_{\ell=1}^{N} \overline{\tau}_\ell \quad \text{and} \quad \tau_\ell \cap \tau_k = \emptyset \text{ for } k \neq \ell.$$

Further we call a triangulation admissible if for two simplicial elements $\tau_k, \tau_\ell \in \mathcal{T}_h$ with $k \neq \ell$ we have either

$$\overline{\tau}_k \cap \overline{\tau}_\ell = \emptyset \quad \text{or} \quad \overline{\tau}_\ell \cap \overline{\tau}_k = \overline{S},$$

where S is an $n - j$-dimensional face of τ_k and τ_ℓ with $j = 1, ..., n$. This means in an admissible triangulation two neighboring simplicial elements join either a node ($n = 1, 2, 3, 4$), an edge ($n = 2, 3, 4$), a triangle ($n = 3, 4$) or a tetrahedron ($n = 4$),

i.e., we avoid hanging nodes, see e.g. [156]. We use $\widetilde{M}$ to denote the number of vertices $\{x_i\}_{i=1}^{\widetilde{M}}$ of a triangulation. For an element $\tau \in \mathcal{T}_h$ we denote with

$$h_\tau := |\tau|^{\frac{1}{n}} = \left(\int_\tau \mathrm{d}x \right)^{\frac{1}{n}}$$

the local mesh size. The global maximal and minimal mesh sizes are given as

$$h := h_{max,\mathcal{T}_h} := \max_{\tau \in \mathcal{T}_h} h_\tau \quad \text{and} \quad h_{min,\mathcal{T}_h} := \min_{\tau \in \mathcal{T}_h} h_\tau.$$

In addition to that we define the diameter of an element $\tau \in \mathcal{T}_h$ as

$$d_\tau := \sup_{x,y \in \tau} |x - y|,$$

which coincides with the longest edge of the element and

$$r_\tau := \sup_{x \in \tau} \{ r > 0 : \ \overline{B_r(x)} \subset \overline{\tau} \}$$

which is the radius of the largest ball that can be inscribed in τ. We say that a family of admissible triangulations $(\mathcal{T}_h)_{h \in I}$ is shape regular, see [156], if there exists a constant $c_F > 0$ independent of $\mathcal{T}_h$ such that

$$d_\tau \leq c_F r_\tau \quad \text{for all } \tau \in \mathcal{T}_h \text{ and } h \in I.$$

We say that a family of admissible triangulations $(\mathcal{T}_h)_{h \in I}$ is globally quasi-uniform if there exists a constant $c_G > 1$ independent of h such that

$$\frac{h_{max,\mathcal{T}_h}}{h_{min,\mathcal{T}_h}} \leq c_G \quad \text{for all } h \in I.$$

Moreover, $(\mathcal{T}_h)_{h \in I}$ is locally quasi-uniform if there exists a constant $c_L > 0$ independent of h such that

$$\frac{h_{\tau_\ell}}{h_{\tau_j}} \leq c_L,$$

for all neighboring elements $\overline{\tau}_\ell \cap \overline{\tau}_j \neq \emptyset$ in $\mathcal{T}_h$ and for all $h \in I$. Now, we are in a position to define the finite element spaces which we use throughout the thesis. For this let $\mathcal{T}_h$ be an admissible triangulation of D. The space of globally continuous and piecewise polynomial functions of order k is defined as

$$S_h^k(\mathcal{T}_h) := \left\{ v_h \in C(\overline{D}) : \ v_h|_\tau \in \mathbb{P}_k(\tau) \text{ for all } \tau \in \mathcal{T}_h \right\}, \tag{2.8}$$

where $\mathbb{P}_k(\tau)$ denotes the set of all polynomials up to order k, i.e.,

$$\mathbb{P}_k(\tau) := \left\{ p : \tau \to \mathbb{R} : p(x) = \sum_{\alpha \in \mathbb{N}_0^n, \, |\alpha| \leq k} c_\alpha x^\alpha, \ c_\alpha \in \mathbb{R} \right\}.$$

Sometimes we will also use the notation $S_h^k(D)$ for the space defined in (2.8). Of particular interest in our considerations is the space $S_h^1(\mathcal{T}_h)$ of piecewise linear finite elements. This space allows the representation

$$S_h^1(\mathcal{T}_h) = \operatorname{span}\{\varphi_i\}_{i=1}^{\widetilde{M}} \subset H^1(D),$$

where φ_i are the usual nodal basis functions satisfying $\varphi_i(x_k) = \delta_{ik}$. Note that the dimension of this space is equal to the number of vertices of the triangulation $\mathcal{T}_h$. For $n = 1, 2, 3$ one can show the following approximation property.

THEOREM 2.13 ([156, Thm. 9.10]). *Let $u \in H^s(\mathcal{T}_h)$ with $s \in [\sigma, 2]$ and $\sigma = 0, 1$. Then there holds the approximation property*

$$\inf_{v_h \in S_h^1(\mathcal{T}_h)} \|u - v_h\|_{H^\sigma(\mathcal{T}_h)} \leq c h^{s-\sigma} |u|_{H^s(\mathcal{T}_h)}. \tag{2.9}$$

REMARK 2.14. *The approximation property (2.9) is also valid for $n = 4$, see e.g. [118, Thm. 2.35]. However, for the proof one needs to construct a quasi-interpolation operator since the nodal interpolation operator is not a well-defined operator for functions in $H^2(D)$ with $D \subset \mathbb{R}^4$.*

3 Least-squares/minimal residual method

In this chapter we aim to derive a least-squares method for the solution of an abstract operator equation $Bu = f$. In particular, we consider the Hilbert spaces X, Y, a given right-hand side $f \in Y^*$ and an abstract operator $B : X \to Y^*$ which in practical applications is given in terms of a partial differential operator. Then, we seek for $u \in X$ which solves

$$Bu = f \quad \text{in } Y^*. \tag{3.1}$$

The usual way to treat (3.1) is to consider some conforming, finite dimensional subspaces $X_H \subset X$ and $Y_h \subset Y$ and to solve the Galerkin-Petrov variational formulation

$$\langle Bu_H, q_h \rangle_H = \langle f, q_h \rangle_H \quad \text{for all } q_h \in Y_h, \tag{3.2}$$

where $\langle ., . \rangle_H$ denotes the dual paring. Existence and uniqueness of (3.2) is based on a discrete inf-sup condition as well as $\dim(X_H) = \dim(Y_h)$, see Theorem 2.5. However, the choice $\dim(X_H) = \dim(Y_h)$ is in certain applications too restrictive. Therefore, we are interested in an alternative solution method which comes along without the latter condition on the discrete spaces since this will enlarge the set of possible stable pairs (X_H, Y_h), cf. [118]. For this reason we consider a minimal residual method for the solution of (3.1). This method looks for a minimizer $u \in X$ to the quadratic functional

$$\mathcal{J}(w) = \frac{1}{2}\|Bw - f\|^2_{Y^*}. \tag{3.3}$$

From (3.3) it can be seen that the residual is measured with respect to the natural norm given by the mapping properties of B. However, in many applications the norm on Y^* is a norm in a Sobolev space with a negative exponent, and therefore one may ask how to deal with this norm. We will address this question in Section 3.1, where we will also derive the minimal residual method and show that its solution is on the continuous level equivalent to the solution of the operator equation (3.1). Further, we will discuss its discretization and show a priori error estimates. The fact that the method is well-posed for $\dim(X_H) \neq \dim(Y_h)$ enables us to show that this method gives rise to an efficient and reliable a posteriori error indicator, which can be used to drive an adaptive refinement scheme. Moreover, at several points we will highlight the connection to a Petrov-Galerkin method with optimal test space [52], as this will give us some fruitful insights on the choice of the test space Y_h. In Section 3.2 we will extend the approach to the solution of nonlinear operator equations. We will discuss a formal derivation of the method as well as possible solution methods.

For the presentation of the results in the linear case we will follow the lines in [106]. However, there are several papers which already deal with a similar approach, see e.g., [46, 47, 127] in case of minimal residual methods, [6] in case of linear parabolic evolution equations, [36, 52, 84, 85] for DPG methods. For a least-squares method in the context of boundary element methods we refer the interested reader to [158] for elliptic problems as well as to [93] for the wave equation in a space-time setting.

3.1 An abstract framework for linear problems

Throughout this section we let $X \subset H \subset X^*$ and $Y \subset H \subset Y^*$ be Gelfand triples of Hilbert spaces, where X^*, Y^* denote the dual spaces of X and Y with respect to H with the duality pairing $\langle f, q\rangle_H$ for $f \in Y^*$ and $q \in Y$. Further, we consider an operator $A : Y \to Y^*$ which we assume to be linear, bounded, self-adjoint and Y-elliptic. With this it immediately follows by Lemma 2.4 that the operator A implies a norm in Y. In what follows we will use the induced norm by the operator A, i.e.,

$$\|q\|_Y := \sqrt{\langle Aq, q\rangle_H} \quad \text{for all } q \in Y.$$

Moreover, in order to derive a least-squares method for the abstract operator equation (3.1) we introduce some assumptions on the operator $B : X \to Y^*$.

Assumption 3.1. *We assume that the operator B fulfills the following conditions:*

(i) B is linear and bounded, i.e., there exists a constant $c_2^B > 0$ such that

$$\|Bv\|_{Y^*} \leq c_2^B \|v\|_X$$

for all $v \in X$.

(ii) B satisfies an inf-sup condition, i.e., there exists a constant $c_1^B > 0$ such that

$$\sup_{0 \neq q \in Y} \frac{\langle Bv, q\rangle_H}{\|q\|_Y} \geq c_1^B \|v\|_X \tag{3.4}$$

for all $v \in X$.

(iii) B is surjective.

3.1.1 Solvability analysis

In view of the Assumptions 3.1 an application of Theorem 2.5 gives that $B : X \to Y^*$ is an isomorphism and that the operator equation (3.1) obtains a unique solution $u \in X$. For the derivation of a minimal residual method we use the following representation of the dual norm

$$\|f\|_{Y^*} := \sup_{0 \neq q \in Y} \frac{\langle f, q \rangle_H}{\|q\|_Y}. \tag{3.5}$$

LEMMA 3.2 ([106, Lem. 2.1]). *For the dual norm* (3.5) *there holds*

$$\|f\|_{Y^*} = \sqrt{\langle A^{-1} f, f \rangle_H} \quad \text{for all } f \in Y. \tag{3.6}$$

Proof. Let $f \in Y^*$ and consider $p_f \in Y$ as the unique solution of the variational problem

$$\langle A p_f, q \rangle_H = \langle f, q \rangle_H \quad \forall q \in Y.$$

This means $p_f = A^{-1} f$ and using Lemma 2.3 it holds that $\|p_f\|_Y \leq \|f\|_{Y^*}$. Now, we can estimate

$$\big\langle A^{-1} f, f \big\rangle_H = \langle p_f, f \rangle_H \leq \|p_f\|_Y \|f\|_{Y^*} \leq \|f\|_{Y^*}^2.$$

On the other hand we have

$$\begin{aligned} \|f\|_{Y^*} &= \sup_{0 \neq q \in Y} \frac{\langle f, q \rangle_H}{\|q\|_Y} = \sup_{0 \neq q \in Y} \frac{\langle A p_f, q \rangle_H}{\|q\|_Y} \\ &\leq \|p_f\|_Y = \|A^{-1} f\|_Y = \sqrt{\langle f, A^{-1} f \rangle_H}. \end{aligned}$$

Combining the estimates we get

$$\|f\|_{Y^*} = \sqrt{\langle A^{-1} f, f \rangle_H} = \|p_f\|_Y. \qquad \square$$

REMARK 3.3. *In literature the operator $A : Y \to Y^*$ is called Riesz operator and $p_f \in Y$ is called Riesz representative of $f \in Y^*$. Further, the equality $\|f\|_{Y^*} = \|A p_f\|_{Y^*} = \|p_f\|_Y = \|A^{-1} f\|_Y$ tells us that the Riesz operator is an isometry which is a well-known property of this operator, see e.g., [25, Thm. 5.5], [143, Satz 10.3].*

Now, we are in a position to consider the quadratic functional (3.3). We first note that due to Lemma 3.2 we can write

$$\begin{aligned} \mathcal{J}(w) &= \frac{1}{2} \|Bw - f\|_{Y^*} = \frac{1}{2} \big\langle A^{-1}(Bw - f), Bw - f \big\rangle_H \\ &= \frac{1}{2} \big\langle B^* A^{-1} B w, w \big\rangle_H - \big\langle B^* A^{-1} f, w \big\rangle_H + \frac{1}{2} \big\langle A^{-1} f, f \big\rangle_H. \end{aligned}$$

A necessary condition for a minimizer to the quadratic functional is that it needs to satisfy the first-order optimality system, which in this case reads

$$0 = \mathrm{D}\mathcal{J}(w)(v) = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{J}(w+tv)\bigg|_{t=0} = \left\langle B^*A^{-1}(Bw-f), v\right\rangle_H$$

for all $v \in X$. This means a candidate $u \in X$ for the minimizer of the quadratic functional has to solve

$$Su := B^*A^{-1}Bu = B^*A^{-1}f \quad \text{in } X^*. \tag{3.7}$$

In the following we will collect some properties of the operator S. This will help us to state existence and uniqueness results to the solution of (3.7).

LEMMA 3.4 ([106, Lem. 2.2]). *The operator $S := B^*A^{-1}B : X \to X^*$ is bounded and elliptic, i.e.,*

$$\|Su\|_{X^*} \le c_2^S \|u\|_X, \quad \langle Su, u\rangle_H \ge c_1^S \|u\|_X^2 \quad \textit{for all } u \in X,$$

where $c_2^S = [c_2^B]^2$, $c_1^S = [c_1^B]^2$.

Proof. Let $u \in X$. Then, using that A^{-1} is an isometry, see Remark 3.3, and that B is bounded we obtain

$$\begin{aligned}\|Su\|_{X^*} &= \sup_{0\neq v\in X} \frac{\langle Su, v\rangle_H}{\|v\|_X} = \sup_{0\neq v\in X} \frac{\langle A^{-1}Bu, Bv\rangle_H}{\|v\|_X} \\ &\le \sup_{0\neq v\in X} \frac{\|A^{-1}Bu\|_Y \|Bv\|_{Y^*}}{\|v\|_X} = \sup_{0\neq v\in X} \frac{\|Bu\|_{Y^*}\|Bv\|_{Y^*}}{\|v\|_X} \le [c_2^B]^2 \|u\|_X.\end{aligned}$$

In order to prove that $S : X \to X^*$ is X-elliptic we define $p_u = A^{-1}Bu \in Y$. With this we get

$$\langle Su, u\rangle_H = \left\langle A^{-1}Bu, Bu\right\rangle_H = \langle p_u, Ap_u\rangle_H = \|p_u\|_Y^2.$$

Moreover, using the inf-sup condition (3.4) of the operator B, we obtain

$$c_1^B \|u\|_X \le \sup_{0\neq q\in Y} \frac{\langle Bu, q\rangle_H}{\|q\|_Y} = \sup_{0\neq q\in Y} \frac{\langle Ap_u, q\rangle_H}{\|q\|_Y} \le \|p_u\|_Y,$$

i.e.,

$$\langle Su, u\rangle_H = \|p_u\|_Y^2 \ge [c_1^B]^2 \|u\|_X^2. \qquad \square$$

REMARK 3.5. *Since the operator $S : X \to X^*$ is elliptic it immediately follows from Lemma 2.4 that*

$$\|v\|_S := \sqrt{\langle Sv, v\rangle_H} = \sqrt{\langle A^{-1}Bv, Bv\rangle_H} = \|Bv\|_{Y^*}$$

defines an equivalent norm on X satisfying

$$c_1^B \|v\|_X \le \|v\|_S \le c_2^B \|v\|_X \quad \textit{for all } v \in X.$$

The properties of the operator S ensure that the first-order optimality system (3.7) is uniquely solvable due to the Lemma of Lax-Milgram, see Lemma 2.3. Using [179, Sec. 2.2, Thm. 2.A] the solution $u \in X$ to (3.7) is indeed a minimizer to the quadratic functional (3.3) since for the second directional derivative we have

$$\mathrm{D}^2\mathcal{J}(u)(v) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{J}(u+tv)\bigg|_{t=0} = \langle Sv, v\rangle_H > 0 \quad \forall v \in X \setminus \{0\}, \tag{3.8}$$

due to the ellipticity of S. The following theorem summarizes our findings so far and draws the connection between solving the operator equation (3.1) and looking for a minimizer to the quadratic functional (3.3).

THEOREM 3.6. *Let $A : Y \to Y^*$ be linear, bounded, self-adjoint and Y-elliptic. Let $B : X \to Y^*$ be linear, bounded, satisfying an inf-sup condition and surjective. Further, let $f \in Y^*$ be a given right-hand side. Then, the following statements hold true:*

(i) The problem: Find $u \in X$ such that $Bu = f$ in Y^* has a unique solution.

(ii) The problem: $\mathcal{J}(w) = \frac{1}{2}\|Bw - f\|^2_{Y^*} \to \min!$ has a unique solution.

(iii) Statements (i) and (ii) are equivalent and obtain the same solution.

Proof. It remains to prove (iii). Let $u \in X$ be a solution to (i), i.e., $Bu = f$ in Y^*. Then, $u \in X$ is also a solution to the first-order optimality system since $B^*A^{-1}(Bu - f) = 0$ in X^*. Moreover, the second order directional derivative (3.8) is positive and hence u is minimizer, i.e., solves (ii). For the other direction we assume $u \in X$ to be a solution to (ii). Then, $u \in X$ solves the first-order optimality system $B^*A^{-1}(Bu - f) = 0$ in X^*. Using that the operators $A : Y \to Y^*$ and $B : X \to Y^*$ are isomorphisms due to the assumptions it has to hold $Bu = f$ in Y^*, i.e., $u \in X$ solves (i). □

REMARK 3.7. *Theorem 3.6 states that one can consider the equivalent minimization problem for the solution of an operator equation. While the corresponding variational formulation of the operator equation can be of Petrov-Galerkin type the resulting operator S of the first-order optimality system will always be an elliptic operator, i.e., the corresponding variational formulation will be of Galerkin-Bubnov type.*

REMARK 3.8. *The first-order optimality system* (3.7) *is sometimes called the normal equation, cf. [21, p. 52]. In our case this equation also involves the operator A^{-1}.*

In the following we will deal with the first-order optimality system (3.7) in more detail. The Galerkin-Bubnov variational formulation is to find $u \in X$ such that

$$\langle Su, v\rangle_H = \left\langle B^*A^{-1}f, v\right\rangle_H \tag{3.9}$$

is satisfied for all $v \in X$.

REMARK 3.9. *Note that* (3.9) *can also be viewed as a Petrov-Galerkin variational formulation with optimal test space. Indeed, choosing* $Y^{opt} := A^{-1}B(X)$ *the variational formulation* (3.9) *reads to find* $u \in X$ *sucht that*

$$\langle Bu, q\rangle_H = \langle f, q\rangle_H \quad \text{for all } q \in Y^{opt}. \tag{3.10}$$

This point of view is taken e.g. in [36, 52, 53, 84].

Since the operator $S = B^*A^{-1}B$ involves the operator A^{-1} which in practical applications is hard to compute we introduce an additional variable $p := A^{-1}(f - Bu)$. This is the Riesz representative of the residual. Now, (3.7) can be rewritten as a mixed system, which reads to find $(u, p) \in X \times Y$ such

$$Ap + Bu = f \quad \text{in } Y^*, \quad B^*p = 0 \quad \text{in } X^*. \tag{3.11}$$

The related variational formulation is to find $(u, p) \in X \times Y$ such that

$$\langle Ap, q\rangle_H + \langle Bu, q\rangle_H = \langle f, q\rangle_H, \quad \langle p, Bv\rangle_H = 0 \tag{3.12}$$

for all $(v, q) \in X \times Y$. Note that by construction it holds that $p \equiv 0$, since $p \in \ker(B^*) = \mathrm{ran}(B)^\perp = \{0\}$.

3.1.2 Discretization

For the discretization of the mixed system (3.12) we consider conforming finite dimensional subspaces $X_H = \mathrm{span}\{\varphi_i\}_{i=1}^{M_X} \subset X$ and $Y_h = \mathrm{span}\{\psi_j\}_{j=1}^{M_Y} \subset Y$. The discrete variational formulation of (3.12) is then to find $(u_H, p_h) \in X_H \times Y_h$ such that

$$\langle Ap_h, q_h\rangle_H + \langle Bu_H, q_h\rangle_H = \langle f, q_h\rangle_H, \quad \langle p_h, Bv_h\rangle_H = 0, \tag{3.13}$$

for all $(v_H, q_h) \in X_H \times Y_h$. Using the Galerkin isomorphism $u_H \leftrightarrow \underline{u} \in \mathbb{R}^{M_X}$ and $p_h \leftrightarrow \underline{p} \in \mathbb{R}^{M_Y}$ the variational formulation (3.13) has an equivalent representation as a linear system of algebraic equations which reads

$$\begin{pmatrix} A_h & B_h \\ B_h^\top & \end{pmatrix} \begin{pmatrix} \underline{p} \\ \underline{u} \end{pmatrix} = \begin{pmatrix} \underline{f} \\ \underline{0} \end{pmatrix}, \tag{3.14}$$

where, for $i, j = 1, \ldots, M_Y$ and $k = 1, \ldots, M_X$,

$$A_h[j, i] = \langle A\psi_i, \psi_j\rangle_H, \quad B_h[j, k] = \langle B\varphi_k, \psi_j\rangle_H, \quad f_j = \langle f, \psi_j\rangle_H.$$

For the unique solvability of (3.13) and hence (3.14) we assume in the following the discrete inf-sup condition

$$c_S\|v_H\|_X \leq \sup_{0\neq q_h\in Y_h} \frac{\langle Bv_H, q_h\rangle_H}{\|q_h\|_Y} \quad \text{for all } v_H \in X_H, \tag{3.15}$$

with a constant $c_S > 0$, see Theorem 2.7. In order to derive an error estimate we consider the equivalent Schur complement system of (3.14). Using $\underline{p} = A_h^{-1}(\underline{f} - B_h\underline{u})$ we obtain the discrete Schur complement system

$$S_h\underline{u} := B_h^\top A_h^{-1} B_h \underline{u} = B_h^\top A_h^{-1}\underline{f}. \tag{3.16}$$

In the following we collect some properties of the Schur complement matrix S_h.

LEMMA 3.10. *Assume the discrete inf-sup stability condition* (3.15) *to be satisfied. Then the Schur complement matrix $S_h = B_h^\top A_h^{-1} B_h$ is positive definite and there hold the norm equivalence inequalities*

$$[c_S]^2 \, \|v_H\|_X^2 \;\le\; (S_h\underline{v}, \underline{v}) \;\le\; [c_2^B]^2 \|v_H\|_X^2 \tag{3.17}$$

for all $\underline{v} \in \mathbb{R}^{M_X} \leftrightarrow v_H \in X_H$.

Proof. For $v_H \in X_H \leftrightarrow \underline{v} \in \mathbb{R}^{M_X}$ we introduce $\underline{p}_v := A_h^{-1} B_h \underline{v} \in \mathbb{R}^{M_Y} \leftrightarrow p_{v_H h} \in Y_h$, which solves

$$\langle A p_{v_H h}, q_h \rangle_H = \langle B v_H, q_h \rangle_H \quad \text{for all } q_h \in Y_h.$$

Thus, we obtain

$$(S_h\underline{v}, \underline{v}) = (A_h^{-1} B_h \underline{v}, B_h \underline{v}) = (\underline{p}_v, B_h\underline{v}) = (A_h \underline{p}_v, \underline{p}_v) = \|p_{v_H h}\|_Y^2$$

The discrete inf-sup stability condition (3.15) implies

$$c_S \|v_H\|_X \le \sup_{0 \neq q_h \in Y_h} \frac{\langle B v_H, q_h \rangle_H}{\|q_h\|_Y} = \sup_{0 \neq q_h \in Y_h} \frac{\langle A p_{v_H h}, q_h \rangle_H}{\|q_h\|_Y} \le \|p_{v_H h}\|_Y,$$

and hence,

$$c_S^2 \|v_H\|_X^2 \le \|p_{v_H h}\|_Y^2 = (S_h\underline{v}, \underline{v}).$$

For the upper estimate we note that

$$\|p_{v_H h}\|_Y \le c_2^B \|v_H\|_X,$$

and therefore the inequality

$$(S_h\underline{v}, \underline{v}) = \|p_{v_H h}\|_Y^2 \le [c_2^B]^2 \|v_H\|_X^2$$

follows. □

Now, we are in a position to prove the following best approximation result, see also [6, Thm. 2.1].

LEMMA 3.11. *Assume the discrete inf-sup stability condition* (3.15). *Then, the Schur complement system* (3.16) *admits a unique solution* $\underline{u} \in \mathbb{R}^{M_X} \leftrightarrow u_H \in X_H$ *with the a priori estimate*

$$\|u_H\|_X \leq \frac{1}{c_S}\|f\|_{Y^*}. \tag{3.18}$$

Moreover, the error estimate

$$\|u - u_H\|_X \leq \frac{c_2^B}{c_S} \inf_{v_H \in X_H} \|u - v_H\|_X, \tag{3.19}$$

holds true, where $u \in X$ *solves* (3.9).

Proof. The unique solvability of (3.16) follows since the matrix S_h is positive definite, see Lemma 3.10. For the a priori estimate we consider

$$(S_h\underline{u}, \underline{u}) = (B_h^\top A_h^{-1}\underline{f}, \underline{u}) = (A_h^{-1}\underline{f}, B_h\underline{u}) \leq \|\underline{f}\|_{A_h^{-1}}\|B_h\underline{u}\|_{A_h^{-1}}.$$

If we use

$$\|B_h\underline{u}\|_{A_h^{-1}}^2 = (A_h^{-1}B_h\underline{u}, B_h\underline{u}) = (S_h\underline{u}, \underline{u}),$$

we obtain the estimate

$$(S_h\underline{u}, \underline{u}) \leq \|\underline{f}\|_{A_h^{-1}}^2 = (A_h^{-1}\underline{f}, \underline{f}) = (A_h\underline{p}_f, \underline{p}_f) = \langle Ap_{fh}, p_{fh}\rangle_H,$$

where $\underline{p}_f \in \mathbb{R}^{M_Y} \leftrightarrow p_{fh} \in Y_h$ solves

$$\langle Ap_{fh}, q_h\rangle_H = \langle f, q_h\rangle_H \quad \text{for all } q_h \in Y_h.$$

Since $\|p_{fh}\|_Y \leq \|f\|_{Y^*}$ and using (3.17) we finally obtain the estimate

$$c_S^2\|u_H\|_X^2 \leq (S_h\underline{u}, \underline{u}) \leq \|f\|_{Y^*}^2,$$

which gives the a priori estimate. In order to prove the best approximation result we define the Galerkin projection $G_H : X \to X_H$, $u \mapsto u_H = G_H u$ which maps an arbitrary element $u \in X$ via $f = Bu \in Y^*$ to a solution $u_H \in X_H \leftrightarrow \underline{u} \in \mathbb{R}^{M_X}$ of (3.16). The mapping G_H is linear due to the linearity of the involved operators. Furthermore, an application of the a priori estimate (3.18) and the fact that $Bu = f$ gives

$$\|G_H u\|_X = \|u_H\|_X \leq \frac{1}{c_S}\|f\|_{Y^*} = \frac{1}{c_S}\|Bu\|_{Y^*} \leq \frac{c_2^B}{c_S}\|u\|_X,$$

i.e., G_H is bounded. In addition to that G_H is indeed a projection. To see this we consider $w_H := G_H u_H$. By defintion of the mapping G_H it holds that $w_H \in X_H \leftrightarrow \underline{w} \in \mathbb{R}^{M_X}$ solves (3.16) for the right-hand side $f = Bu_H \leftrightarrow \underline{f} = B_h\underline{u} \in \mathbb{R}^{M_Y}$, i.e.,

$$S_h\underline{w} = B_h^\top A_h^{-1}\underline{f} = B_h^\top A_h^{-1}B_h\underline{u} = S_h\underline{u}.$$

Now, using that the Schur complement matrix is invertible we obtain $\underline{w} = \underline{u}$, which gives $G_H u_H = u_H$ and hence

$$G_H^2 u = G_H(G_H u) = G_H u_H = u_H,$$

i.e., G_H is a projection. Finally, using $G_H v_H = v_H$, $\|I - G_H\|_{X\to X} = \|G_H\|_{X\to X}$, see [173, Lem. 5] and the boundedness of G_H we obtain the estimate

$$\begin{aligned}\|u - u_H\|_X &= \|(I - G_H)u\|_X = \|(I - G_H)(u - v_H)\|_X \\ &\leq \|I - G_H\|_{X\to X}\|u - v_H\|_X = \|G_H\|_{X\to X}\|u - v_H\|_X \\ &\leq \frac{c_2^B}{c_S}\|u - v_H\|_X.\end{aligned}$$

Since $v_H \in X_H$ was arbitrary the assertion follows. □

3.1.3 A posteriori error estimator

So far, we have seen that a minimizer $u \in X$ to (3.3) can be computed from the first-order optimality system (3.7) and that this system obtains the representation as the saddle point system (3.11), when introducing the Riesz representative of the residual $p = A^{-1}(f - Bu)$. While in the continuous setting we have $p \equiv 0$, the discrete mixed variational formulation (3.13) gives $p_h \in Y_h \leftrightarrow \underline{p} = A_h^{-1}(\underline{f} - B_h\underline{u}) \neq \underline{0}$ as the Galerkin matrix B_h is in general not invertible, cf. [106]. This is a consequence of the fact that we only need injectivity on the discrete level for the operator B, see the discrete inf-sup condition (3.15), in order to get well-posedness of (3.13), see Theorem 2.7. Therefore, the Galerkin matrix B_h is in general a rectangular matrix as $\dim(X_H) \neq \dim(Y_h)$. In this sense, the minimal residual approach for the solution of $Bu = f$ gives us on the discrete level more flexibility in the choice of stable pairs (X_H, Y_h) than a standard Galerkin-Petrov scheme where one has to enforce $\dim(X_H) = \dim(Y_h)$ in order to get solvability. However, we introduce some additional unknowns which in terms of computational costs can be seen as a disadvantage, cf. [118]. But, in the following we will show that the additional unknown p_h can be used to define an error estimator for the true error $\|u - u_H\|_X$. This has also been demonstrated, e.g., in [36, Thm. 2.1] or [127, Prop. 3.8]

Lemma 3.12 ([106, Lem. 2.5]). *Let $(u_H, p_h) \in X_H \times Y_h$ be the unique solution of (3.13). Then, there holds the efficiency estimate*

$$\|p_h\|_Y \leq c_2^B\|u - u_H\|_X. \tag{3.20}$$

Proof. We subtract (3.13) from (3.12) for $q = q_h \in Y_h \subset Y$ and $v = v_H \in X_H \subset X$ to obtain the Galerkin orthogonalities

$$\langle A(p - p_h), q_h\rangle_H + \langle B(u - u_H), q_h\rangle_H = 0, \quad \langle p - p_h, Bv_H\rangle_H = 0 \tag{3.21}$$

for $(v_H, q_h) \in X_H \times Y_h$. Since $p \equiv 0$ they boil down to

$$\langle Ap_h, q_h\rangle_H = \langle B(u - u_H), q_h\rangle_H, \quad \langle Bv_H, p_h\rangle_H = 0$$

for all $(v_H, q_h) \in X_H \times Y_h$. Testing with $q_h = p_h \in Y_h$ gives

$$\|p_h\|_Y^2 = \langle Ap_h, p_h\rangle_H = \langle B(u - u_H), p_h\rangle_H \leq c_2^B \|u - u_H\|_X \|p_h\|_Y,$$

from which we conclude the result. □

In order to prove a reliability estimate we make use of the fact, see e.g. [28, Rem. 2.1], [127, Thm. 3.7], [164, Prop. 5.1], that the discrete inf-sup stability condition (3.15) implies the existence of a Fortin projector [23, 68].

LEMMA 3.13. *Assume the discrete inf-sup stability condition* (3.15). *Then, there exists a mapping* $\Pi_h : Y \to Y_h$, *satisfying*

$$\langle r - \Pi_h r, Bv_H\rangle_H = 0 \text{ for all } v_H \in X_H, \quad \|\Pi_h r\|_Y \leq \frac{c_2^B}{c_S}\|r\|_Y \text{ for all } r \in Y. \quad (3.22)$$

Proof. The mapping can be constructed when defining $\Pi_h : Y \to Y_h$, $r \mapsto \Pi_h r$, as the second component of the solution $(w_H, r_h) \in X_H \times Y_h$ of the variational problem

$$\langle Ar_h, q_h\rangle_H - \langle Bw_H, q_h\rangle_H = 0, \quad \langle r_h, Bv_H\rangle_H = \langle r, Bv_H\rangle_H$$

for all $(v_H, q_h) \in X_H \times Y_h$. This mixed system reads in terms of algebraic equations as

$$A_h \underline{r} - B_h \underline{w} = 0, \quad B_h^\top \underline{r} = \underline{g},$$

where $\underline{g}[i] = \langle r, B\varphi_i\rangle$, $i = 1, ..., M_X$. The unique solvability of this system follows since the associated Schur complement system $S_h \underline{w} = \underline{g}$ admits a unique solution due to Lemma 3.10. Therefore, the mapping is well-defined and fulfills the condition $\langle r - \Pi_h r, Bv_H\rangle_H = 0$, for all $v_H \in X_H$. Moreover, we have

$$\|r_h\|_Y^2 = \langle Ar_h, r_h\rangle_H = \langle Bw_H, r_h\rangle_H = \langle Bw_H, r\rangle_H \leq c_2^B \|w_H\|_X \|r\|_Y.$$

When using the discrete inf-sup condition (3.15) we obtain

$$c_S\|w_H\|_X \leq \sup_{0\neq q_h \in Y} \frac{\langle Bw_H, q_h\rangle_H}{\|q_h\|_Y} = \sup_{0\neq q_h \in Y_h} \frac{\langle Ar_h, q_h\rangle_H}{\|q_h\|_Y} \leq \|r_h\|_Y.$$

Finally, combining the latter two estimates we get

$$\|r_h\|_Y \leq \frac{c_2^B}{c_S}\|r\|_Y,$$

which finishes the proof. □

Now, we can prove a reliability estimate for $p_h \in Y_h$, see [36, cf. Thm. 2.1] in the context of DPG methods.

LEMMA 3.14. *Let $(u_H, p_h) \in X_H \times Y_h$ be the unique solution to (3.13) and let $u \in X$ be the unique solution to (3.1). Then, $p_h \in Y_h$ is reliable modulo the data oscillation term $\operatorname{osc}(f) = \|f \circ (I - \Pi_h)\|_{Y^*}$, i.e., it satisfies*

$$\|u - u_H\|_X \leq \frac{1}{c_S} \frac{c_2^B}{c_1^B} \|p_h\|_Y + \frac{1}{c_1^B} \operatorname{osc}(f) . \tag{3.23}$$

Proof. Using the inf-sup condition (3.4) and the properties of the Fortin projector (3.22) we obtain

$$\begin{aligned} c_1^B \|u - u_H\|_X &\leq \sup_{0 \neq q \in Y} \frac{\langle B(u - u_H), q \rangle_H}{\|q\|_Y} \\ &= \sup_{0 \neq q \in Y} \frac{\langle B(u - u_H), q - \Pi_h q \rangle_H + \langle B(u - u_H), \Pi_h q \rangle_H}{\|q\|_Y} \\ &= \sup_{0 \neq q \in Y} \frac{\langle Bu, q - \Pi_h q \rangle_H + \langle Ap_h, \Pi_h q \rangle_H}{\|q\|_Y} \\ &= \sup_{0 \neq q \in Y} \frac{\langle f, q - \Pi_h q \rangle_H + \langle Ap_h, \Pi_h q \rangle_H}{\|q\|_Y} \leq \operatorname{osc}(f) + \frac{c_2^B}{c_S} \|p_h\|_Y, \end{aligned}$$

which gives the assertion. □

As pointed out in [36, Thm. 2.1, Rem. 2.6], [127, Rem. 3.9] for the data oscillation term it holds

$$\begin{aligned} \operatorname{osc}(f) &= \sup_{0 \neq q \in Y} \frac{\langle f, (I - \Pi_h) q \rangle_H}{\|q\|_Y} = \sup_{0 \neq q \in Y} \frac{\langle Bu, (I - \Pi_h) q \rangle_H}{\|q\|_Y} \\ &= \sup_{0 \neq q \in Y} \frac{\langle B(u - v_H), (I - \Pi_h) q \rangle_H}{\|q\|_Y} \leq c_2^B \|I - \Pi_h\|_{Y \to Y} \|u - v_H\|_X, \end{aligned}$$

i.e., with $\|I - \Pi_h\|_{Y \to Y} = \|\Pi_h\|_{Y \to Y}$, see again [173, Lem. 5] and (3.22) we get

$$\operatorname{osc}(f) \leq \frac{[c_2^B]^2}{c_S} \inf_{v_H \in X_H} \|u - v_H\|_X.$$

This means the data oscillation term is at least of the same order as the true error $\|u - u_H\|_X$. However, for sufficiently smooth data we can expect that the data oscillation term is of higher order than $\|u - u_H\|_X$, in particular when using trial and test spaces which are defined with respect to different polynomial degrees of the involved basis functions, see [36]. In this case the error indicator p_h is asymptotically reliable.

In what follows we will give an alternative proof for the reliability estimate, which follows the lines [106]. This is motivated by the paper [59] which states that small data oscillations imply a so-called saturation assumption. Furthermore, it draws a connection to the hierarchical error estimators, in particular the well-known $h - \frac{h}{2}$ estimator, see e.g., [37, 67]. To prove a reliability estimate in this context it is necessary to introduce an additional ansatz space $\bar{X}_H \subset X$ which satisfies $X_H \subset \bar{X}_H$. For this trial space we assume the discrete inf-sup stability condition

$$\bar{c}_S \|\bar{v}_H\|_X \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle B\bar{v}_H, q_h \rangle_H}{\|q_h\|_Y} \quad \text{for all } \bar{v}_H \in \bar{X}_H, \tag{3.24}$$

with a constant $\bar{c}_S > 0$. Due to $X_H \subset \bar{X}_H$ it holds that (3.24) implies the discrete inf-sup stability condition (3.15). In addition to that we consider $(\bar{u}_H, \bar{p}_h) \in \bar{X}_H \times Y_h$ as the unique solution to the variational formulation

$$\langle A\bar{p}_h, q_h \rangle_H + \langle B\bar{u}_H, q_h \rangle_H = \langle f, q_h \rangle_H, \quad \langle \bar{p}_h, B\bar{v}_H \rangle_H = 0 \tag{3.25}$$

for all $(\bar{v}_H, q_h) \in \bar{X}_H \times Y_h$. Now, we are able to prove the following estimate.

Lemma 3.15 (cf. [106, Lem. 2.6]). *Let $(u_H, p_h) \in X_H \times Y_h$ and $(\bar{u}_H, \bar{p}_h) \in \bar{X}_H \times Y_h$ be the unique solutions of the Galerkin variational formulations* (3.13) *and* (3.25)*, respectively. Assume the saturation assumption*

$$\|u - \bar{u}_H\|_X \leq \eta \, \|u - u_H\|_X \quad \textit{for some } \eta \in (0, 1). \tag{3.26}$$

Then the error estimator $\|p_h\|_Y$ is reliable, satisfying

$$\|u - u_H\|_X \leq \frac{1}{1 - \eta} \frac{1}{\bar{c}_S} \|p_h\|_Y \,. \tag{3.27}$$

Proof. When subtracting the Galerkin variational formulation (3.25) from (3.13) we obtain the Galerkin orthogonality

$$\langle B(\bar{u}_H - u_H), q_h \rangle_H = \langle A(p_h - \bar{p}_h), q_h \rangle_H \quad \text{for all } q_h \in Y_h. \tag{3.28}$$

Since $\bar{u}_H - u_H \in \bar{X}_H$ we conclude from the discrete inf-sup stability condition (3.24)

$$\begin{aligned} \bar{c}_S \|\bar{u}_H - u_H\|_X &\leq \sup_{0 \neq q_h \in Y_h} \frac{\langle B(\bar{u}_H - u_H), q_h \rangle_H}{\|q_h\|_Y} \\ &= \sup_{0 \neq q_h \in Y_h} \frac{\langle A(p_h - \bar{p}_h), q_h \rangle_H}{\|q_h\|_Y} \leq \|p_h - \bar{p}_h\|_Y \,. \end{aligned}$$

Further, by the second equation in (3.25) we have that $\langle \bar{p}_h, B(\bar{u}_H - u_H) \rangle_H = 0$. With this and the Galerkin orthogonality (3.28) we can estimate

$$\begin{aligned} \|p_h - \bar{p}_h\|_Y^2 &= \langle A(p_h - \bar{p}_h), p_h - \bar{p}_h \rangle_H = \langle B(\bar{u}_H - u_H), p_h - \bar{p}_h \rangle_H \\ &= \langle B(\bar{u}_H - u_H), p_h \rangle_H = \langle A(p_h - \bar{p}_h), p_h \rangle_H \leq \|p_h - \bar{p}_h\|_Y \|p_h\|_Y, \end{aligned}$$

i.e.,

$$\|p_h - \bar{p}_h\|_Y \leq \|p_h\|_Y.$$

Hence, we obtain

$$\|\bar{u}_H - u_H\|_X \leq \frac{1}{\bar{c}_S} \|p_h - \bar{p}_h\|_Y \leq \frac{1}{\bar{c}_S} \|p_h\|_Y,$$

Finally, using the triangle inequality and the saturation assumption (3.26) we get

$$\|u - u_H\|_X \leq \|u - \bar{u}_H\|_X + \|\bar{u}_H - u_H\|_X \leq \eta \|u - u_H\|_X + \frac{1}{\bar{c}_S} \|p_h\|_Y$$

from which the assertion follows. □

Since $X_H \subset \bar{X}_H$ we expect that the solution $\bar{u}_H$ will be a better approximation to the true solution $u \in X$ than $u_H \in X_H$. The saturation assumption (3.26) is a condition which quantifies how much better this approximation is, cf. [4, p. 88]. However, as pointed out in [118, Rem. 5.9] it is hard to verify that such a condition is fulfilled for explicit examples, see also [38].

3.1.4 On the choice of the test space Y_h

In this subsection we want to address the question on how large do we have to choose Y_h such that the discrete inf-sup condition (3.15) holds.

Let us consider the variational formulation (3.9). The discrete variational formulation is to find $\bar{u}_H \in X_H$ such that

$$\langle S\bar{u}_H, v_H \rangle_H = \left\langle B^* A^{-1} f, v_H \right\rangle_H \quad \text{for all } v_H \in X_H. \tag{3.29}$$

Obviously, (3.29) admits a unique solution due to the ellipticity of S, see Lemma 3.4. As pointed out in e.g. [52, 84] it can be viewed as an ideal Petrov-Galerkin method which seeks for $\bar{u}_H \in X_H$ satisfying

$$\langle B\bar{u}_H, q_H \rangle_H = \langle f, q_H \rangle_H \quad \text{for all } q_H \in Y_H^{opt} := T(X_H) := A^{-1}B(X_H),$$

where Y_H^{opt} is called optimal test space and $T : X \to Y$ is called trial-to-test operator. The choice $Y_h = Y_H^{opt}$ yields immediately the discrete inf-sup condition since

$$\sup_{0 \neq q_h \in Y_H^{opt}} \frac{\langle Bv_H, q_h \rangle_H}{\|q_h\|_Y} \geq \frac{\langle Bv_H, A^{-1}Bv_H \rangle_H}{\|A^{-1}Bv_H\|_Y} = \frac{\langle Sv_H, v_H \rangle_H}{\|v_H\|_S} = \|v_H\|_S \geq c_1^B \|v_H\|_X.$$

However, the optimal test space is in general not realizable as the trial-to-test operator T does not allow for a direct evaluation. Hence, $\bar{u}_H$ is in general not computable.

From the Schur complement system (3.16) we infer that the computable solution $u_H \in X_H$ satisfies

$$\langle Bu_H, q_h\rangle_H = \langle f, q_h\rangle_H \quad \text{for all } q_h \in Y_H^{prac} := P_h Y_H^{opt} \subset Y_h, \tag{3.30}$$

where $P_h : Y \to Y_h$ denotes the orthogonal projection. This means the solution u_H is computed with respect to a practical realization $Y_H^{prac} \subset Y_h$ of the optimal test space, which is the orthogonal projection of the optimal test space Y_H^{opt} onto Y_h. Note that (3.30) is also called practical Petrov-Galerkin method, see e.g. [85].

REMARK 3.16. *In [52] it is shown how to construct optimal test functions for the practical Petrov-Galerkin method* (3.30) *in a Discontinuous Galerkin (DG) setting. An analysis of the DPG method in case of the Poisson equation can be found in [51, 53, 85]. In these references an ultraweak formulation based on a first-order reformulation of the Poisson equation as well as a primal formulation without a first-order reformulation are discussed. We use the saddle point formulation* (3.13) *which is equivalent to* (3.30)*, see e.g. [28, Prop. 2.2]. The saddle point formulation involves the trial space X_H and the test space Y_h and not explicitly the practical optimal test space Y_H^{prac}. Moreover, we will use a Continuous Galerkin (CG) method for the discretization. Note that this approach was also followed by [46].*

A first insight into the connection between the optimal testspace Y_H^{opt} and the test space Y_h gives the following result.

THEOREM 3.17. *Assume the discrete inf-sup condition* (3.15)*. Then there holds*

$$\inf_{q_h \in Y_h} \|\bar{p} - q_h\|_Y \leq \sqrt{1 - \left(\frac{c_S}{c_2^B}\right)^2}\, \|\bar{p}\|_Y \quad \textit{for all } \bar{p} \in Y_H^{opt} \tag{3.31}$$

Proof. First note that an element $\bar{p} \in Y_H^{opt}$ can be associated with an element $v_H \in X_H$ via $\bar{p} = A^{-1}Bv_H$. Now, we can define $\bar{p}_h \in Y_h$ as the unique solution to the variational problem

$$\langle A\bar{p}_h, q_h\rangle_H = \langle Bv_H, q_h\rangle_H = \langle A\bar{p}, q_h\rangle_H \quad \text{for all } q_h \in Y_h$$

satisfying the orthogonality relation

$$\|\bar{p} - \bar{p}_h\|_Y^2 = \|\bar{p}\|_Y^2 - \|\bar{p}_h\|_Y^2.$$

Note that $\bar{p}_h \in Y_H^{prac}$. Now, we can estimate

$$\begin{aligned}
\|\bar{p}_h\|_Y^2 = \langle A\bar{p}_h, \bar{p}_h\rangle_H &= \left(\frac{\langle A\bar{p}_h, \bar{p}_h\rangle_H}{\|\bar{p}_h\|_Y}\right)^2 \\
&= \left(\sup_{0\neq q_h \in Y_h} \frac{\langle A\bar{p}_h, q_h\rangle_H}{\|q_h\|_Y}\right)^2 = \left(\sup_{0\neq q_h \in Y_h} \frac{\langle Bv_H, q_h\rangle_H}{\|q_h\|_Y}\right)^2 \\
&\geq c_S^2 \|v_H\|_X^2 \geq \left(\frac{c_S}{c_2^B}\right)^2 \|v_H\|_S^2 = \left(\frac{c_S}{c_2^B}\right)^2 \|\bar{p}\|_Y^2.
\end{aligned}$$

With this we obtain

$$\|\bar{p} - \bar{p}_h\|_Y^2 \leq \left(1 - \left(\frac{c_S}{c_2^B}\right)^2\right) \|\bar{p}\|_Y^2$$

Finally, using

$$\inf_{q_h \in Y_h} \|\bar{p} - q_h\|_Y \leq \|\bar{p} - \bar{p}_h\|_Y$$

gives the desired result. □

REMARK 3.18. *A slightly stronger statement is proven in [28, Prop. 2.5]. There it is shown that in* (3.31) *even equality holds.*

The assertion of Theorem 3.17 can be also reversed in the sense that for given space X_H it gives us an abstract condition on how rich the test space Y_h in comparison to the optimal test space Y_H^{opt} has to be in order to fulfill a discrete inf-sup stability condition.

THEOREM 3.19 ([106, Thm. 2.7]). *For a given finite element space $X_H \subset X$ let $Y_H^{opt} = A^{-1}B(X_H)$ be the associated optimal test space. Further, let $Y_h \subset Y$ such that*

$$\inf_{q_h \in Y_h} \|\bar{p} - q_h\|_Y \leq \delta \|\bar{p}\|_Y \quad \text{for all } \bar{p} \in Y_H^{opt} \tag{3.32}$$

is satisfied for some $\delta \in (0,1)$. Then, there holds the discrete inf-sup stability condition

$$c_1^B \sqrt{(1-\delta^2)} \|v_H\|_X \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle Bv_H, q_h \rangle_H}{\|q_h\|_Y} \quad \text{for all } v_H \in X_H. \tag{3.33}$$

Proof. First note that an arbitrary element $v_H \in X_H$ can be uniquely identified with an element $\bar{p} \in Y_H^{opt}$ via $\bar{p} = A^{-1}Bv_H$ since the involved operators are isomorphisms. Then we can compute

$$\|v_H\|_S^2 = \|\bar{p}\|_Y^2.$$

In addition to that we define $\bar{p}_h \in Y_h$ as unique solution of the Galerkin variational formulation

$$\langle A\bar{p}_h, q_h \rangle_H = \langle Bv_H, q_h \rangle_H = \langle A\bar{p}, q_h \rangle_H \quad \text{for all } q_h \in Y_h,$$

satisfying the bound

$$\|\bar{p}_h\|_Y \leq \|\bar{p}\|_Y,$$

Cea's lemma,

$$\|\bar{p} - \bar{p}_h\|_Y \leq \inf_{q_h \in Y_h} \|\bar{p} - q_h\|_Y,$$

and the orthogonality relation

$$\|\bar{p}\|_Y^2 = \|\bar{p}_h\|_Y^2 + \|\bar{p} - \bar{p}_h\|_Y^2.$$

Then, from (3.32) we obtain

$$\|\bar{p} - \bar{p}_h\|_Y \leq \delta \|\bar{p}\|_Y.$$

Hence, we can write

$$\begin{aligned}\left(\sup_{0\neq q_h\in Y_h} \frac{\langle Bv_H, q_h\rangle_H}{\|q_h\|_Y}\right)^2 &= \left(\sup_{0\neq q_h\in Y_h} \frac{\langle A\bar{p}_h, q_h\rangle_H}{\|q_h\|_Y}\right)^2 = \left(\frac{\langle A\bar{p}_h, \bar{p}_h\rangle_H}{\|\bar{p}_h\|_Y}\right)^2 \\ &= \|\bar{p}_h\|_Y^2 = \|\bar{p}\|_Y^2 - \|\bar{p} - \bar{p}_h\|_Y^2 \\ &\geq (1-\delta^2)\|\bar{p}\|_Y^2 \\ &= (1-\delta^2)\|v_H\|_S^2\end{aligned}$$

implying the inf-sup stability condition (3.33). □

3.1.5 Discretization dependent norm for the ansatz space

As pointed out in [106], in some applications, e.g., when considering space-time finite elements for parabolic evolution equations as in [157], it might be useful to define a discretization dependent norm $\|.\|_{X,h}$ on X in order to establish a discrete inf-sup condition. So in this subsection we will show that the main results of the least-squares framework are still valid. Again, we will follow the presentation in [106].

Let $\|.\|_{X,h} : X \to \mathbb{R}$ define a norm on X which satisfies $\|v\|_{X,h} \leq \|v\|_X$ for all $v \in X$ and which is asymptotically equivalent to the original norm on X, i.e., $\|\cdot\|_{X,h} \to \|\cdot\|_X$ as $h \to 0$. Assume that

$$\tilde{c}_S \|v_H\|_{X,h} \leq \sup_{0\neq q_h\in Y_h} \frac{\langle Bv_H, q_h\rangle_H}{\|q_h\|_Y} \quad \text{for all } v_H \in X_H. \tag{3.34}$$

Then the following stability and error estimate holds.

Lemma 3.20 ([106, cf. Lem. 2.8]). *Assume the discrete inf-sup stability condition* (3.34). *Then the Schur complement matrix* $S_h = B_h^\top A_h^{-1} B_h$ *is positive definite and there hold the inequalities*

$$[\tilde{c}_S]^2 \|v_H\|_{X,h}^2 \leq (S_h \underline{v}, \underline{v}) \leq [c_2^B]^2 \|v_H\|_X^2 \tag{3.35}$$

for all $\underline{v} \in \mathbb{R}^{M_X} \leftrightarrow v_H \in X_H$. *Furthermore, the Schur complement system* (3.16) *as well as the mixed system* (3.13) *admit a unique solution and there holds the error estimate*

$$\|u - u_H\|_{X,h} \leq \frac{c_2^B}{\tilde{c}_S} \inf_{v_H\in X_H} \|u - v_H\|_X. \tag{3.36}$$

Proof. The proof of the inequalities (3.35) follows the lines of the proof of Lemma 3.10 replacing the discrete inf-sup stability condition (3.15) with (3.34). This ensures the unique solvability of (3.16), (3.13), respectively, since $\|.\|_{X,h}$ defines a norm on $X_H \subset X$. The proof of the error estimate works in the same way as demonstrated in Lemma 3.11, using (3.35) instead of (3.17). □

In order to show that the error indicator is efficient and reliable, we can proceed similar as in Section 3.1.3. We consider an additional ansatz space $\bar{X}_H \subset X$, which satisfies $X_H \subset \bar{X}_H$. For this trial space we assume the discrete inf-sup stability condition

$$\bar{c}_S \|\bar{v}_H\|_{X,h} \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle B\bar{v}_H, q_h \rangle_H}{\|q_h\|_Y} \quad \text{for all } \bar{v}_H \in \bar{X}_H, \tag{3.37}$$

with a constant $\bar{c}_S > 0$. Note that (3.37) implies the discrete inf-sup stability condition (3.34). Now, we can provide the following result.

LEMMA 3.21. *Let $u \in X$ be the unique solution to (3.1), $(u_H, p_h) \in X_H \times Y_h$ be the unique solution to (3.13) and $(\bar{u}_H, \bar{p}_h)$ be the unique solution to the variational formulation (3.25). Further, assume that the operator B is also bounded with respect to the discrete norm $\|\cdot\|_{X,h}$, i.e.,*

$$\langle Bv, q_h \rangle_H \leq c_2^B \|v\|_{X,h} \|q_h\|_Y \quad \forall v \in X,\ \forall q_h \in Y_h, \tag{3.38}$$

and the saturation assumption

$$\|u - \bar{u}_H\|_{X,h} \leq \eta \|u - u_H\|_{X,h} \quad \text{for some } \eta \in (0,1). \tag{3.39}$$

Then there hold the inequalities

$$\frac{1}{c_2^B} \|p_h\|_Y \leq \|u - u_H\|_{X,h} \leq \frac{1}{1-\eta} \frac{1}{\bar{c}_S} \|p_h\|_Y,$$

i.e., the error indicator p_h is efficient and reliable.

Proof. The efficiency estimate can be proven in the same way as in Lemma 3.12 using (3.38), i.e.,

$$\begin{aligned} \|p_h\|_Y^2 &= \langle Ap_h, p_h \rangle_H = \langle B(u - u_H), p_h \rangle_H \\ &\leq c_2^B \|u - u_H\|_{X,h} \|p_h\|_Y, \end{aligned}$$

from which the first inequality follows. For the reliability estimate we can follow the lines of the proof of Lemma 3.15. We recall the main steps. Using the Galerkin

orthogonality (3.28) and that $\bar{u}_H - u_H \in \bar{X}_H$ we conclude from the discrete inf-sup stability condition (3.37)

$$\begin{aligned}\bar{c}_S \|\bar{u}_H - u_H\|_{X,h} &\leq \sup_{0 \neq q_h \in Y_h} \frac{\langle B(\bar{u}_H - u_H), q_h \rangle_H}{\|q_h\|_Y} \\ &= \sup_{0 \neq q_h \in Y_h} \frac{\langle A(p_h - \bar{p}_h), q_h \rangle_H}{\|q_h\|_Y} \leq \|p_h - \bar{p}_h\|_Y \,.\end{aligned}$$

Further, using that $\langle \bar{p}_h, B(\bar{u}_H - u_H) \rangle_H = 0$, which follows from (3.25), and the Galerkin orthogonality (3.28) we obtain

$$\begin{aligned}\|p_h - \bar{p}_h\|_Y^2 &= \langle A(p_h - \bar{p}_h), p_h - \bar{p}_h \rangle_H = \langle B(\bar{u}_H - u_H), p_h - \bar{p}_h \rangle_H \\ &= \langle B(\bar{u}_H - u_H), p_h \rangle_H = \langle A(p_h - \bar{p}_h), p_h \rangle_H \\ &\leq \|p_h - \bar{p}_h\|_Y \|p_h\|_Y.\end{aligned}$$

Hence, we conclude

$$\|\bar{u}_H - u_H\|_{X,h} \leq \frac{1}{\bar{c}_S} \|p_h\|_Y.$$

Finally, using the triangle inequality and the saturation assumption (3.39) we get

$$\|u - u_H\|_{X,h} \leq \|u - \bar{u}_H\|_{X,h} + \|\bar{u}_H - u_H\|_{X,h} \leq \eta \|u - u_H\|_{X,h} + \frac{1}{\bar{c}_S} \|p_h\|_Y$$

from which the assertion follows. □

3.2 Extension to the nonlinear case

In this section we will extend the approach seen in Section 3.1 to the nonlinear case. In the context of DPG methods there are some works dealing with nonlinear problems. We want to mention the work [41], where a Gauß-Newton scheme is used to solve a nonlinear minimal residual problem. A steepest descent approach is considered in the PhD thesis [116], which is based on the works [26, 27], where they solved a transonic flow problem. A PDE-constrained residual minimization approach is considered [31]. Finally, in [35] the DPG framework is applied to a nonlinear diffusion problem and quasi-optimal a priori and reliable and efficient a posteriori error estimates are shown.

For the motivation of the least-squares approach in the nonlinear case we consider two Hilbert spaces X, Y, and a nonlinear operator $B : X \to Y^*$. Then we are interested to find a solution $u \in X$ to the nonlinear operator equation

$$B(u) = f \quad \text{in } Y^*. \tag{3.40}$$

The discretization of (3.40) reads to find $u_H \in X_H \subset X$ such that

$$\langle B(u_H), q_h \rangle_H = \langle f, q_h \rangle_H \quad \text{for all } q_h \in Y_h \subset Y.$$

This leads to a system of nonlinear equations which can be solved with e.g. Newton's method [57]. In Newton's method one computes a sequence of iterates $\{u_H^k\}_k \subset X_H$ which satisfy

$$u_H^{k+1} = u_H^k + w_H^k,$$

where $w_H^k \in X_H$ is the solution of the Newton system

$$\big\langle B'(u_H^k) w_H^k, q_h \big\rangle_H = -\big\langle B(u_H^k) - f, q_h \big\rangle_H \quad \text{for all } q_h \in Y_h. \tag{3.41}$$

For well-posedness of (3.41) one has to assume some discrete inf-sup stability conditon as well as $\dim(X_H) = \dim(Y_h)$. However, as pointed out in [118] the latter condition is in some applications too restrictive as it narrows possible stable pairs (X_H, Y_h). Therefore, in the following we will consider a minimal residual approach for the solution of (3.40), which seeks for a minimizer $u \in X$ to the least-squares functional

$$J(w) = \frac{1}{2} \|B(w) - f\|_{Y^*}^2 \tag{3.42}$$

We start with an abstract derivation of the method and then discuss solution strategies involving Newton's as well as Gauß-Newton's method.

3.2.1 Derivation of the method

As in the linear setting we consider $X \subset H \subset X^*$ and $Y \subset H \subset Y^*$ to be Gelfand triples of Hilbert spaces, where X^*, Y^* denote the dual spaces with respect to H with the duality pairing $\langle f, q \rangle_H$ for $f \in Y^*$ and $q \in Y$. Further, let $A : Y \to Y^*$ be a linear, bounded, self-adjoint and Y-elliptic operator. In addition to that we consider a nonlinear operator $B : X \to Y^*$ which we assume to be an isomorphism and twice Fréchet differentiable.

The minimal residual method aims to find a minimizer $u \in X$ to (3.42) which due to Lemma 3.2 reads

$$\begin{aligned} \mathcal{J}(w) &= \frac{1}{2} \|B(w) - f\|_{Y^*}^2 = \frac{1}{2} \big\langle A^{-1}(B(w) - f), B(w) - f \big\rangle_H \\ &= \frac{1}{2} \big\langle A^{-1} B(w), B(w) \big\rangle_H - \big\langle A^{-1} f, B(w) \big\rangle_H + \frac{1}{2} \big\langle A^{-1} f, f \big\rangle_H \end{aligned}$$

A minimizer $u \in X$ to (3.42) has to fulfill the first-order optimality system which in this case reads

$$\mathrm{D}\mathcal{J}(u)(v) = \big\langle B'(u)^* A^{-1}(B(u) - f), v \big\rangle_H = 0 \quad \text{for all } v \in X. \tag{3.43}$$

In order to solve (3.43) we can apply Newton's method, which gives the well-known Newton minimization algorithm [133].

ALGORITHM 3.22. *Choose an initial guess* $u^0 \in X$.

For $k = 0, 1, 2...$, *until convergence do*

(i) Compute $w_u^k \in X$ *as solution to*

$$\mathrm{D}^2\mathcal{J}(u^k)(w_u^k, v) = -\mathrm{D}\mathcal{J}(u^k)(v) \quad \text{for all } v \in X. \tag{3.44}$$

(ii) Set $u^{k+1} = u^k + w_u^k$.

In particular, (3.44) reads in our case

$$\begin{aligned}\left\langle B'(u^k)^* A^{-1} B'(u^k) w_u^k, v\right\rangle_H + \left\langle B''(u^k)(w_u^k, v), A^{-1}(B(u^k) - f)\right\rangle_H \\ = -\left\langle B'(u^k)^* A^{-1}(B(u^k) - f), v\right\rangle_H\end{aligned} \tag{3.45}$$

for all $v \in X$. In our numerical examples we do not apply Newton's method directly to (3.43). Instead, we first introduce similar as in the linear case the Riesz representative of the residual $p := A^{-1}(f - B(u)) \in Y$. Then (3.43) can be rewritten as a mixed system to find $(u, p) \in X \times Y$ such that

$$Ap + B(u) = f \text{ in } Y^*, \quad B'(u)^* p = 0 \text{ in } X^*. \tag{3.46}$$

The related variational formulation of the mixed system is then to find $(u, p) \in X \times Y$ such that

$$\langle Ap, q\rangle_H + \langle B(u), q\rangle_H = \langle f, q\rangle_H, \quad \langle p, B'(u)v\rangle_H = 0 \tag{3.47}$$

holds for all $(v, q) \in X \times Y$. Now, we apply Newton's method to the operator $G : Y \times X \to Y^* \times X^*$ defined by

$$G(p, u) := \begin{bmatrix} G_1(p, u) \\ G_2(p, u) \end{bmatrix} := \begin{bmatrix} Ap + B(u) - f \\ B'(u)^* p \end{bmatrix},$$

which gives the following algorithm for the solution of (3.43).

ALGORITHM 3.23. *Choose an initial guess* $z^0 := (p^0, u^0) \in Y \times X$.

For $k = 0, 1, 2, ...$, *until convergence do*

(i) Compute $w^k = (w_p^k, w_u^k) \in Y \times X$ *as solution to*

$$\left\langle G'(p^k, u^k)(w_p^k, w_u^k), (q, v)\right\rangle_H = -\left\langle G(p^k, u^k), (q, v)\right\rangle_H \quad \forall (q, v) \in Y \times X. \tag{3.48}$$

(ii) Set $z^{k+1} = z^k + w^k$, *where* $z^k = (p^k, u^k)$.

In more detail (3.48) reads to find $(w_p^k, w_u^k) \in Y \times X$ such that

$$\begin{aligned} \left\langle Aw_p^k, q\right\rangle_H + \left\langle B'(u^k)w_u^k, q\right\rangle_H &= -\left\langle G_1(p^k, u^k), q\right\rangle_H, \\ \left\langle w_p^k, B'(u^k)v\right\rangle_H + \left\langle p^k, B''(u^k)(w_u^k, v)\right\rangle_H &= -\left\langle G_2(p^k, u^k), v\right\rangle_H \end{aligned} \tag{3.49}$$

is satisfied for all $(q, v) \in Y \times X$. The reason for the consideration of the optimality system (3.43) in its mixed form (3.46) is that it allows the computation of the solution u as well as the Riesz representative of the residual simultaneously. This is in particular interesting if one wants to use the Riesz lift of the residual to drive an adaptive refinement scheme. Furthermore, we want to mention that the implementation of Algorithm 3.23 can be done via the automatic differentiability capabilities of `Netgen/NGSolve`, which felt more convenient. Both Algorithms, i.e., 3.22, 3.23, respectively, have in common that they need the second derivative of B. However, in some cases one might be interested in a solution method which does not need the second derivative of the operator B. One possibility is to use a Gauß-Newton method for computing a minimizer to the nonlinear least-squares functional. In Gauß-Newton's method the second directional derivative in (3.44) is approximated by

$$\mathrm{D}^2\mathcal{J}(u^k)(w_u^k, v) \approx \left\langle B'(u^k)^* A^{-1} B'(u^k) w_u^k, v\right\rangle_H.$$

This results in the following algorithm.

ALGORITHM 3.24. *Choose an initial guess $u^0 \in X$.*

For $k = 0, 1, 2...$, until convergence do

(i) Compute $w_u^k \in X$ as solution to

$$\left\langle B'(u^k)^* A^{-1} B'(u^k) w_u^k, v\right\rangle_H = -\left\langle B'(u^k)^* A^{-1}(B(u^k) - f), v\right\rangle_H \tag{3.50}$$

for all $v \in X$.

(ii) Set $u^{k+1} = u^k + w_u^k$.

REMARK 3.25. *In operator form* (3.50) *reads*

$$B'(u^k)^* A^{-1}\left[B'(u^k)w_u^k + B(u^k) - f\right] = 0,$$

i.e., the Gauß-Newton search direction $w_u^k \in X$ is the minimizer of the linearized least-squares functional

$$w_u^k = \operatorname*{arg\,min}_{v \in X} \frac{1}{2}\left\|B'(u^k)v + B(u^k) - f\right\|_{Y^*}^2.$$

REMARK 3.26. *In order to overcome the realization of the inverse of A and to have the Riesz representative of the residual explicitly available one can implement step (i) in Algorithm 3.24 in the following way: First determine $p^k \in Y$ as solution to*

$$\left\langle Ap^k, q\right\rangle_H = \left\langle B(u^k) - f, q\right\rangle_H \quad \text{for all } q \in Y.$$

Then solve the mixed system to find $(w_p^k, w_u^k) \in Y \times X$ such that

$$\left\langle Aw_p^k, q\right\rangle_H + \left\langle B'(u^k)w_u^k, q\right\rangle_H = 0, \quad \left\langle w_p^k, B'(u^k)v\right\rangle_H = \left\langle p^k, B'(u^k)v\right\rangle_H$$

is satisfied for all $(q, v) \in Y \times X$.

A comparison between (3.45) and (3.50) reveals that for the Gauß-Newton method to be well-posed one only needs stability of the first order derivative B', which is in practice often beneficial. For the well-posedness of Newton's method one needs additional conditions on the second order derivative B'' of the nonlinear operator B. However, in general the convergence of Gauß-Newton's method is slower than that of Newton's method, cf. [57, p. 199], [133, p. 257]. In what follows we assume that the operator $B'(u) : X \to Y^*$ satisfies an inf-sup condition and a boundedness estimate uniformly for all $u \in X$, i.e.,

$$c_1^{B'}\|v\|_X \leq \sup_{0\neq q\in Y} \frac{\langle B'(u)v, q\rangle_H}{\|q\|_Y}, \quad \|B'(u)v\|_{Y^*} \leq c_2^{B'}\|v\|_X \tag{3.51}$$

for all $v \in X$ with constants $c_1^{B'}$, $c_2^{B'} > 0$ independent of u. Then we can show that the operator $S_u := B'(u)^*A^{-1}B'(u) : X \to X^*$ obtains the following properties.

LEMMA 3.27. *Under the assumptions of* (3.51) *the operator $S_u := B'(u)^*A^{-1}B'(u) : X \to X^*$ is bounded and elliptic uniformly for any $u \in X$, i.e.,*

$$\|S_u w\|_{X^*} \leq c_2^{S_u}\|w\|_X, \quad \langle S_u w, w\rangle_H \geq c_1^{S_u}\|w\|_X^2 \quad \text{for all } w \in X,$$

where $c_2^{S_u} = [c_2^{B'}]^2$, $c_1^{S_u} = [c_1^{B'}]^2$.

Proof. The proof follows the lines of the proof of Lemma 3.4 □

COROLLARY 3.28. *Under the assumptions* (3.51) *the variational formulation* (3.50) *admits a unique solution and the Gauß-Newton search direction $w_u^k \in X$ is a descent direction.*

Proof. Unique solvability of (3.50) follows immediately from Lemma 3.27. For the unique solution $w_u^k \in X$ to (3.50) it holds

$$\begin{aligned} \mathrm{D}\mathcal{J}(u^k)(w_u^k) &= \left\langle B'(u^k)^*A^{-1}(B(u^k) - f), w_u^k\right\rangle_H \\ &= -\left\langle B'(u^k)^*A^{-1}B'(u^k)w_u^k, w_u^k\right\rangle_H \leq -[c_1^{B'}]^2\left\|w_u^k\right\|_X^2 < 0 \end{aligned}$$

for $w_u^k \neq 0$. Hence, the assertion follows. □

REMARK 3.29. *If the operator $B'(u) : X \to Y^*$ does not satisfy the conditions in* (3.51) *one can use a steepest descent method to solve* (3.42). *This involves the computation of a Riesz representative of*

$$\mathrm{D}\mathcal{J}(u) = B'(u)^* A^{-1}(B(u) - f) \in X^*.$$

This can be done via an auxiliary boundary value problem, where B' is solely used to compute the right-hand side. Hence, no conditions on B' are needed. However, this comes at the cost of a lower convergence rate of the iterative scheme compared to that of a Gauß-Newton method. The steepest descent approach is considered e.g. in [116].

3.2.2 Discretization

For the discretization we consider conforming finite dimensional subspaces $X_H = \mathrm{span}\{\varphi_i\}_{i=1}^{M_X} \subset X$ and $Y_h = \mathrm{span}\{\psi_i\}_{i=1}^{M_Y} \subset Y$, which are defined with respect to some admissible decomposition of the computational domain into shape regular simplicial finite elements of mesh size H,h, respectively. The discrete version of the Newton algorithm 3.23 starts with an initial guess $(p_h^0, u_H^0) \in Y_h \times X_H$ and then computes iteratively $(p_h^{k+1}, u_H^{k+1}) = (p_h^k, u_H^k) + (w_{ph}^k, w_{uH}^k)$, where $(w_{ph}^k, w_{uH}^k) \in Y_h \times X_H$ solves

$$\begin{aligned} \big\langle A w_{ph}^k, q_h \big\rangle_H + \big\langle B'(u_H^k) w_{uH}^k, q_h \big\rangle_H &= -\big\langle G_1(p_h^k, u_H^k), q_h \big\rangle_H, \\ \big\langle w_{ph}^k, B'(u_H^k) v_H \big\rangle_H + \big\langle p_h^k, B''(u_H^k)(w_{uH}^k, v_H) \big\rangle_H &= -\big\langle G_2(p_h^k, u_H^k), v_H \big\rangle_H \end{aligned} \tag{3.52}$$

for all $(q_h, v_H) \in Y_h \times X_H$. Introducing

$$\begin{aligned} A_h[i,j] &= \langle A\psi_j, \psi_i \rangle_H \quad i,j = 1, ..., M_Y, \\ \underline{B}(\underline{u})[i] &= \langle B(u_H), \psi_i \rangle, \quad i = 1, ..., M_Y, \\ B_h'(\underline{u})[i,j] &= \langle B'(u_H)\varphi_j, \psi_i \rangle_H, \quad i = 1, ..., M_Y,\ j = 1, ..., M_X, \\ B_h''(\underline{u}, \underline{p})[i,j] &= \langle p_h, B''(u_H)(\varphi_j, \varphi_i) \rangle_H, \quad i,j = 1, ..., M_X \end{aligned}$$

with the identification $u_H \in X_H \leftrightarrow \underline{u} \in \mathbb{R}^{M_X}$, $p_h \in Y_h \leftrightarrow \underline{p} \in \mathbb{R}^{M_Y}$ we can write the discrete version of Algorithm 3.23 as following.

ALGORITHM 3.30. *Choose an initial guess $(p_h^0, u_H^0) \in Y_h \times X_H \leftrightarrow (\underline{p}^0, \underline{u}^0) \in \mathbb{R}^{M_Y} \times \mathbb{R}^{M_X}$.*

For $k = 0, 1, 2, ...,$ until convergence do

(i) Solve the algebraic system of equations

$$\begin{bmatrix} A_h & B_h'(\underline{u}^k) \\ B_h'(\underline{u}^k)^\top & B_h''(\underline{u}^k, \underline{p}^k) \end{bmatrix} \begin{bmatrix} \underline{w}_p^k \\ \underline{w}_u^k \end{bmatrix} = - \begin{bmatrix} A_h \underline{p}^k + \underline{B}(\underline{u}^k) - \underline{f} \\ (B_h'(\underline{u}^k))^\top \underline{p}^k \end{bmatrix}.$$

(ii) Set $(p_h^{k+1}, u_H^{k+1}) \leftrightarrow (\underline{p}^{k+1}, \underline{u}^{k+1}) = (\underline{p}^k, \underline{u}^k) + (\underline{w}_p^k, \underline{w}_u^k)$.

In view of Remark 3.26 the discrete version of the Gauß-Newton algorithm starts with an initial guess $u_H^0 \in X_H$ and then computes iteratively $u_H^{k+1} = u_H^k + w_{uH}^k$, where we first determine $p_h^k \in Y_h$ as solution to

$$\left\langle Ap_h^k, q_h \right\rangle_H = \left\langle B(u_H^k) - f, q_h \right\rangle_H \quad \text{for all } q_h \in Y_h.$$

In the second step we compute $(w_{ph}, w_{uH}) \in Y_h \times X_H$ such that

$$\begin{aligned} \left\langle Aw_{ph}^k, q_h \right\rangle_H + \left\langle B'(u_H^k) w_{uH}^k, q_h \right\rangle_H &= 0 \\ \left\langle w_{ph}^k, B'(u_H^k) v_H \right\rangle_H &= \left\langle p_h^k, B'(u_H^k) v_H \right\rangle_H \end{aligned}$$

holds for all $(q_h, v_H) \in Y_h \times X_H$. Thus, we can state the following algorithm.

ALGORITHM 3.31. *Choose an initial guess* $u_H^0 \in X_H \leftrightarrow \underline{u}^0 \in \mathbb{R}^{M_X}$.

For $k = 0, 1, 2, ...,$ *until convergence do*

(i) Set $\underline{p}^k = A_h^{-1}(\underline{B}(\underline{u}^k) - \underline{f})$.

(ii) Solve the algebraic system of equations

$$\begin{bmatrix} A_h & B_h'(\underline{u}^k) \\ B_h'(\underline{u}^k)^\top & \end{bmatrix} \begin{bmatrix} \underline{w}_p^k \\ \underline{w}_u^k \end{bmatrix} = \begin{bmatrix} 0 \\ B_h'(\underline{u}^k)^\top \underline{p}^k \end{bmatrix}.$$

(iii) Set $u_H^{k+1} \leftrightarrow \underline{u}^{k+1} = \underline{u}^k + \underline{w}_u^k$.

REMARK 3.32. *For the practical implementation of the Newton algorithm 3.30 and the Gauß-Newton algorithm 3.31 we will use their damped version [57, p.109-172], [133]. This is to overcome the local convergence behaviour of Newton's or Gauß-Newton's method, respectively, as also mentionend in [76]. In the damped version of these algorithms we perform an update* $(\underline{p}^{k+1}, \underline{u}^{k+1}) = (\underline{p}^k, \underline{u}^k) + \tau(\underline{w}_p^k, \underline{w}_u^k)$, $\underline{u}^{k+1} = \underline{u}^k + \tau \underline{w}_u^k$, *respectively, where the parameter* $\tau \in (0, 1]$ *is chosen according to some line search strategy.*

4 Parabolic evolution equations

In this chapter we apply the minimal residual/least-squares framework described in Chapter 3 in combination with the conforming space-time finite element discretization scheme described in [157] to parabolic evolution equations. In particular, we consider the heat equation, the convection-diffusion equation and a semilinear heat equation as model problems. For the heat equation and the convection-diffusion equation a discrete inf-sup stability condition will be shown with respect to a mesh-dependent norm which ensures stability of the mixed system (3.14). The finite element matrices used to set up the mixed system (3.14) will be implemented using standard finite element libraries. Several numerical examples will be presented which confirm our theoretical findings and a comparison to the FOSLS method of Führer and Karkulik in [72] in case of the heat equation will be given. For the nonlinear problem a solution via Newton's and Gauß-Newton's method will be presented.

4.1 Heat equation

We consider the Dirichlet boundary value problem for the heat equation

$$\partial_t u(x,t) - \Delta_x u(x,t) = f(x,t) \quad \text{for } (x,t) \in Q := \Omega \times (0,T), \tag{4.1a}$$

$$u(x,t) = 0 \quad \text{for } (x,t) \in \Sigma := \Gamma \times (0,T), \tag{4.1b}$$

$$u(x,0) = 0 \quad \text{for } x \in \Omega, \tag{4.1c}$$

where $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ is a bounded Lipschitz domain with boundary $\Gamma = \partial\Omega$, $T > 0$ is a given time horizon and f is a given right-hand side.

The common space-time variational formulation of Bochner type [64] is to find $u \in X = L^2(0,T; H_0^1(\Omega)) \cap H_{0,}^1(0,T; H^{-1}(\Omega)$ such that

$$\langle Bu, q \rangle_Q := \langle \partial_t u, q \rangle_Q + \langle \nabla_x u, \nabla_x q \rangle_{L^2(Q)} = \langle f, q \rangle_Q \tag{4.2}$$

is satisfied for all $q \in Y := L^2(0,T; H_0^1(\Omega))$. Stable space-time discretization schemes for this variational formulation are considered in, e.g., [108, 153, 157]. We will employ a minimal residual discretization which we described in Chapter 3. The idea to use a least-squares approach for the solution of the heat equation was already considered by Andreev [6]. However, as pointed out in [82] a verification of the LBB condition for discrete pairs (X_h, Y_h) has been restricted to pairs of finite element spaces with

respect to partitions of Q that permit a decomposition into time-slabs [164, 165]. Therefore, stability of the scheme for fully unstructured decompositions of space and time are not covered. Using the abstract framework in Section 3.1 in the spirit of the space-time discretization scheme [157] we will show that the necessary discrete inf-sup condition on the operator B is satisfied with respect to a mesh-dependent norm, see also [106]. This gives us stability on the discrete level with respect to arbitrary decompositions of the space-time domain Q into simplicial elements. This result also enables us to prove that the inbuilt error indicator is efficient and under some saturation assumption reliable. In Section 4.1.2 we will describe an alternative least-squares approach, namely the space-time first order system least-squares (FOSLS) method introduced by Führer and Karkulik in [72]. The reformulation as a first order system comes with the fact that the residual is measured in a norm which is localizable, i.e., in a norm which allows for L^2-regularity. However, the handling of loads which belong to Sobolev spaces of negative order is more involved and not straightforward, see e.g. [70, 71] for a related approach. We also want to refer to [172], where a space-time discretization using a constrained first order system least-squares (CFOSLS) method is proposed, which is based on the works [3, 132]. In [172] discretizations for the heat equation, a scalar conservation law and the wave equation are considered and numerical examples in $(d+1)$ dimensions with $d = 2, 3$ are presented. The approach uses a slightly different reformulation of the original PDE as a first order system than in [72]. Finally, in Section 4.1.3 we will present some numerical examples, which will confirm our theoretical findings.

4.1.1 Minimal residual method

For the solution of (4.2) we consider the space-time finite element method proposed in [157] and apply it with respect to the least-squares setting developed in Section 3.1. In this context we have the Bochner spaces

$$X := L^2(0,T;H^1_0(\Omega)) \cap H^1_{0,}(0,T;H^{-1}(\Omega)),\ Y := L^2(0,T;H^1_0(\Omega)),\ H := L^2(Q)$$

with the corresponding norms

$$\|p\|_Y := \|\nabla_x p\|_{L^2(Q)}, \quad \|u\|_X := \sqrt{\|\partial_t u\|^2_{Y^*} + \|\nabla_x u\|^2_{L^2(Q)}}.$$

The operators $A : Y \to Y^*$ and $B : X \to Y^*$ are defined in the variational sense satisfying

$$\langle Ap, q\rangle_Q := \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)}, \quad \langle Bu, q\rangle_Q := \langle \partial_t u, q\rangle_Q + \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)}, \tag{4.3}$$

for all $p,\ q \in Y$ and $u \in X$, where $\langle\cdot,\cdot\rangle_Q$ denotes the duality pairing. In order to apply the abstract framework from Section 3.1 we have to make sure that the operators A, B fulfill the assumptions. This will be done in the following.

LEMMA 4.1. *The operator $A : Y \to Y^*$ defined in* (4.3) *is bounded, self-adjoint and elliptic with constants $c_1^A = c_2^A = 1$.*

Proof. The boundedness of A follows from the Cauchy-Schwarz inequality, i.e.,

$$\langle Ap, q\rangle_Q = \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)} \leq \|\nabla_x p\|_{L^2(Q)} \|\nabla_x q\|_{L^2(Q)} = \|p\|_Y \|q\|_Y.$$

From the definiton of A it follows

$$\langle Ap, q\rangle_Q = \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)} = \langle \nabla_x q, \nabla_x p\rangle_{L^2(Q)} = \langle Aq, p\rangle_Q,$$

which gives the self-adjointness. The ellipticity immediately follows from

$$\langle Ap, p\rangle_Q = \langle \nabla_x p, \nabla_x p\rangle_{L^2(Q)} = \|p\|_Y^2. \qquad \square$$

LEMMA 4.2. *The operator $B : X \to Y^*$ defined in* (4.3) *is bounded with $c_2^B = \sqrt{2}$, fulfills the inf-sup condition* (3.4) *with $c_1^B = 1$ and is surjective.*

Proof. Using the estimate $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b > 0$ we obtain

$$\begin{aligned}\langle Bu, q\rangle_Q &= \langle \partial_t u, q\rangle_Q + \langle \nabla_x u, \nabla_x q\rangle_{L^2(Q)} \leq \|\partial_t u\|_{Y^*}\|q\|_Y + \|u\|_Y\|q\|_Y \\ &\leq \sqrt{2}\sqrt{\|\partial_t u\|_{Y^*}^2 + \|u\|_Y^2}\|q\|_Y = \sqrt{2}\|u\|_X\|q\|_Y,\end{aligned}$$

which gives boundedness of the operator B. In order to prove an inf-sup condition we proceed as in [106]. We define $w_u := A^{-1}\partial_t u \in Y$, i.e., w_u solves the variational problem

$$\langle \nabla_x w_u, \nabla_x q\rangle_{L^2(Q)} = \langle \partial_t u, q\rangle_Q \quad \text{for all } q \in Y.$$

Note that $w_u \in Y$ is nothing than the Riesz representative of $\partial_t u \in Y^*$. Thus, by Lemma 3.2 and Remark 3.3 we conclude $\|w_u\|_Y = \|\partial_t u\|_{Y^*}$. For $\bar{q} := u + w_u \in Y$ we then have

$$\begin{aligned}\langle Bu, \bar{q}\rangle_Q &= \langle Bu, u + w_u\rangle_Q \\ &= \langle \partial_t u, u + w_u\rangle_Q + \langle \nabla_x u, \nabla_x (w + w_u)\rangle_{L^2(Q)} \\ &= \langle \nabla_x w_u, \nabla_x (u + w_u)\rangle_{L^2(Q)} + \langle \nabla_x u, \nabla_x (w + w_u)\rangle_{L^2(Q)} \\ &= \langle \nabla_x (u + w_u), \nabla_x (w + w_u)\rangle_{L^2(Q)} \\ &= \|u + w_u\|_Y^2 = \|\bar{q}\|_Y^2.\end{aligned}$$

Further we have

$$\begin{aligned}\|\bar{q}\|_Y^2 &= \|\nabla_x (u + w_u)\|_{L^2(Q)}^2 \\ &= \|\nabla_x u\|_{L^2(Q)}^2 + 2\langle \partial_t u, u\rangle_Q + \|\nabla_x w_u\|_{L^2(Q)}^2 \\ &\geq \|\nabla_x u\|_{L^2(Q)}^2 + \|\nabla_x w_u\|_{L^2(Q)}^2 = \|u\|_X^2,\end{aligned}$$

where we used

$$\begin{aligned}\langle \partial_t u, u\rangle_Q &= \int_0^T \langle \partial_t u(t), u(t)\rangle_\Omega \,\mathrm{d}t = \frac{1}{2}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), u(t)\rangle_{L^2(\Omega)}\,\mathrm{d}t \\ &= \frac{1}{2}\|u(T)\|^2_{L^2(\Omega)} - \frac{1}{2}\|u(0)\|^2_{L^2(\Omega)} = \frac{1}{2}\|u(T)\|^2_{L^2(\Omega)} \geq 0.\end{aligned}$$

Now, we obtain

$$\langle Bu, \bar{q}\rangle_Q \geq \|u\|_X \|\bar{q}\|_Y,$$

and therefore we conclude the inf-sup stability condition

$$\sup_{0\neq q\in Y} \frac{\langle Bu, q\rangle_Q}{\|q\|_Y} \geq \frac{\langle Bu, \bar{q}\rangle_Q}{\|\bar{q}\|_Y} \geq \|u\|_X \quad \text{for all } u \in X.$$

It remains to prove that B is surjective. For this let $q \in Y \setminus \{0\}$. Then we define

$$u_q(x,t) := \int_0^t q(x,s)\,\mathrm{d}s.$$

Since $u_q(x,0) = 0$ and $\partial_t u_q = q \in Y$ it holds that $u_q \in X$. Further we compute

$$\begin{aligned}\langle Bu_q, q\rangle_Q &= \langle \partial_t u_q, q\rangle_{L^2(Q)} + \langle \nabla_x u_q, \nabla_x q\rangle_{L^2(Q)} \\ &= \|q\|^2_{L^2(Q)} + \int_0^T \langle \nabla_x u_q(t), \nabla_x \partial_t u_q(t)\rangle_{L^2(\Omega)}\,\mathrm{d}t \\ &= \|q\|^2_{L^2(Q)} + \frac{1}{2}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\langle \nabla_x u_q(t), \nabla_x u_q(t)\rangle_{L^2(\Omega)}\,\mathrm{d}t \\ &= \|q\|^2_{L^2(Q)} + \frac{1}{2}\|\nabla_x u_q(T)\|^2_{L^2(\Omega)} > 0,\end{aligned}$$

which gives surjectivity of B and concludes the proof. □

Remark 4.3. *A proof of the inf-sup condition on B can also be found in e.g. [157, Thm. 2.1] with a constant $c_1^B = \frac{1}{2\sqrt{2}}$. It also follows from the inf-sup identity [65, Thm. 2.1]. However, the idea for the choice of the test function $\bar{q} = u + A^{-1}\partial_t u$ can already be found in [64, Thm. 6.6]. Note that it holds $\bar{q} \in Y^{opt} = A^{-1}B(X)$, i.e., it is an optimal test function, cf. Rem. 3.9 and Sec. 3.1.4. Indeed, using $Bu = \partial_t u + Au$, a straightforward computation gives*

$$\bar{q} = u + w_u = u + A^{-1}\partial_t u = A^{-1}(\partial_t u + Au) = A^{-1}Bu.$$

The properties of the operator B ensure that the variational problem (4.2) admits a unique solution $u \in X$. Together with the properties of the operator A and by Lemma 3.6 it holds

$$u = \arg\min_{w\in X} \frac{1}{2}\|Bw - f\|^2_{Y^*}.$$

A solution to the minimization problem can be obtained from the mixed system (3.12), which in this particular case reads to find $(u,p) \in X \times Y$ such that

$$\begin{aligned} \langle \nabla_x p, \nabla_x q \rangle_{L^2(Q)} + \langle \partial_t u, q \rangle_Q + \langle \nabla_x u, \nabla_x q \rangle_{L^2(Q)} &= \langle f, q \rangle_Q, \\ \langle \partial_t v, p \rangle_Q + \langle \nabla_x v, \nabla_x p \rangle_{L^2(Q)} &= 0 \end{aligned} \tag{4.4}$$

is satisfied for all $(v,q) \in X \times Y$.

For the discretization of (4.4) we consider the finite dimensional subspaces $X_H \subset X$ and $Y_h \subset Y$, where we assume the inclusion $X_H \subset Y_h$. As already used in the proof of Lemma 4.2 the norm $\|u\|_X$ allows the representation

$$\|u\|_X = \sqrt{\|w_u\|_Y^2 + \|u\|_Y^2}, \quad \|w_u\|_Y = \|\partial_t u\|_{Y^*},$$

where $w_u := A^{-1}\partial_t u$ is the Riesz representative of $\partial_t u \in Y^*$, i.e., it solves the variational formulation

$$\langle \nabla_x w_u, \nabla_x q \rangle_{L^2(Q)} = \langle \partial_t u, q \rangle_Q \quad \text{for all } q \in Y.$$

In view of this observation we define $w_{uh} \in Y_h$ as the unique solution of the Galerkin variational formulation

$$\langle \nabla_x w_{uh}, \nabla_x q_h \rangle_{L^2(Q)} = \langle \partial_t u, q_h \rangle_Q \quad \text{for all } q_h \in Y_h.$$

With this we can define the discrete norm

$$\|u\|_{X,h} := \sqrt{\|u\|_Y^2 + \|w_{uh}\|_Y^2} \le \|u\|_X. \tag{4.5}$$

This allows us to prove the following properties of the operator B.

LEMMA 4.4. *Let $X_H \subset X$ and $Y_h \subset Y$ be finite dimensional subspaces of X, Y, respectively with $X_H \subset Y_h$. Then the operator B satisfies the discrete boundedness estimate*

$$\langle Bu, q_h \rangle_Q \le \sqrt{2}\|u\|_{X,h}\|q_h\|_Y \quad \textit{for all } u \in X,\ q_h \in Y_h,$$

as well as the discrete inf-sup condition

$$\|u_H\|_{X,h} \le \sup_{0 \neq q_h \in Y_h} \frac{\langle Bu_H, q_h \rangle_Q}{\|q_h\|_Y} \quad \textit{for all } u_H \in X_H. \tag{4.6}$$

Proof. From the definition of the discrete norm (4.5) and using the Cauchy-Schwarz inequality as well as $(a+b)^2 \le 2(a^2+b^2)$ we obtain

$$\begin{aligned} \langle Bu, q_h \rangle_Q &= \langle \partial_t u, q_h \rangle_Q + \langle \nabla_x u, \nabla_x q_h \rangle_{L^2(Q)} \\ &= \langle \nabla_x w_{uh}, \nabla_x q_h \rangle_{L^2(Q)} + \langle \nabla_x u, \nabla_x q_h \rangle_{L^2(Q)} \\ &\le \left(\|w_{uh}\|_Y + \|u\|_Y\right) \|q_h\|_Y \\ &\le \sqrt{2}\|u\|_{X,h}\|q_h\|_Y, \end{aligned}$$

which gives the discrete boundedness estimate. The proof of the discrete inf-sup condition follows the lines as in the continuous case, see Lemma 4.2. We recall the main steps. First, we define $\bar{q}_h := u_H + w_{u_H h}$, where $w_{u_H h} \in Y_h$ solves

$$\langle \nabla_x w_{u_H h}, \nabla_x q_h \rangle_{L^2(Q)} = \langle \partial_t u, q_h \rangle_Q \quad \text{for all } q_h \in Y_h.$$

As $X_H \subset Y_h$ we have $\bar{q}_h \in Y_h$. Then we can compute

$$\langle Bu_H, \bar{q}_h \rangle_Q = \|\bar{q}_h\|_Y^2.$$

Further we have due to $\langle \partial_t u_H, u_H \rangle_Q \geq 0$ the estimate

$$\|\bar{q}_h\|_Y^2 = \|u_H\|_Y^2 + \|w_{u_H h}\|_Y^2 + 2\langle \partial_t u_H, u_H \rangle_Q \geq \|u_H\|_{X,h}^2,$$

which implies

$$\langle Bu_H, \bar{q}_h \rangle_Q \geq \|u_H\|_{X,h} \|\bar{q}_h\|_Y.$$

Finally, we obtain

$$\sup_{0 \neq q_h \in Y_h} \frac{\langle Bu_H, q_h \rangle_Q}{\|q_h\|_Y} \geq \frac{\langle Bu_H, \bar{q}_h \rangle_Q}{\|\bar{q}_h\|_Y} \geq \|u_H\|_{X,h},$$

which gives the desired result. □

REMARK 4.5. *A similar proof of the discrete inf-sup condition can be also found in [157, Thm. 3.1].*

The discrete inf-sup stability condition shown in Lemma 4.4 ensures the unique solvability of the mixed variational formulation (3.13), which in this particular case is to find $(u_H, p_h) \in X_H \times Y_h$ such that

$$\begin{aligned} \langle \nabla_x p_h, \nabla_x q_h \rangle_{L^2(Q)} + \langle \partial_t u_H, q_h \rangle_Q + \langle \nabla_x u_H, \nabla_x q_h \rangle_{L^2(Q)} &= \langle f, q_h \rangle_Q, \\ \langle \partial_t v_H, p_h \rangle_Q + \langle \nabla_x v_H, \nabla_x p_h \rangle_{L^2(Q)} &= 0 \end{aligned} \tag{4.7}$$

is satisfied for all $(v_H, q_h) \in X_H \times Y_h$. The related error estimate follows from (3.36) and reads with $c_2^B = \sqrt{2}$ and $\tilde{c}_S = 1$

$$\|u - u_H\|_{X,h} \leq \sqrt{2} \inf_{v_H \in X_H} \|u - v_H\|_X.$$

In case of a piecewise linear finite element space for the trial space X_H and a sufficient regular solution $u \in H^s(Q)$ for some $s \in [1, 2]$ we conclude the error estimate, see e.g., [157]

$$\|\nabla_x (u - u_H)\|_{L^2(Q)} \leq \|u - u_H\|_{X,h} \leq c H^{s-1} |u|_{H^s(Q)}.$$

In view of Lemma 3.21 we can define $\bar{X}_H = Y_h \cap X$ and show the discrete inf-sup condition (3.37) similar as demonstrated in Lemma 4.4 with $\bar{c}_S = 1$. Using that $c_2^B = \sqrt{2}$ and assuming the saturation assumption (3.39) for some $\eta \in (0, 1)$ we conclude for the error indicator

$$\frac{1}{\sqrt{2}} \|\nabla_x p_h\|_{L^2(Q)} \leq \|u - u_H\|_{X,h} \leq \frac{1}{1 - \eta} \|\nabla_x p_h\|_{L^2(Q)}.$$

4.1.2 FOSLS method

In this section we describe the space-time first order least-squares finite element method [72]. For the derivation of the method we assume that $f \in L^2(Q)$.

The basic idea, which was already mentioned in [21, Ch. 9.1.4] is to rewrite the heat equation (4.1) as a first oder system. For this we introduce the auxiliary variable $\sigma = -\nabla_x u$. Then we obtain the first order system

$$\begin{aligned} \partial_t u(x,t) + \operatorname{div}_x \sigma(x,t) &= f(x,t) && \text{for } (x,t) \in Q := \Omega \times (0,T), && \text{(4.8a)} \\ \sigma(x,t) + \nabla_x u(x,t) &= 0 && \text{for } (x,t) \in Q && \text{(4.8b)} \\ u(x,t) &= 0 && \text{for } (x,t) \in \Sigma := \Gamma \times (0,T), && \text{(4.8c)} \\ u(x,0) &= 0 && \text{for } x \in \Omega. && \text{(4.8d)} \end{aligned}$$

Now, we can define the spaces

$$\begin{aligned} U &:= \{(u,\sigma) \in X \times L^2(Q)^d : \ \partial_t u + \operatorname{div}_x \sigma \in L^2(Q)\}, \\ V &:= L^2(Q) \times L^2(Q)^d, \end{aligned}$$

where $X = L^2(0,T;H_0^1(\Omega)) \cap H_{0,}^1(0,T;H^{-1}(\Omega))$ with the corresponding norms for the spaces U, V

$$\begin{aligned} \|(u,\sigma)\|_U^2 &:= \|u\|_X^2 + \|\sigma\|_{L^2(Q)}^2 + \|\partial_t u + \operatorname{div}_x \sigma\|_{L^2(Q)}^2, \\ \|(w,\varrho)\|_V^2 &:= \|w\|_{L^2(Q)}^2 + \|\varrho\|_{L^2(Q)}^2, \end{aligned}$$

for all $(u,\sigma) \in U$, $(w,\varrho) \in V$. Further we can define the operator

$$\mathcal{B} : U \to V, \ (u,\sigma) \mapsto \mathcal{B}(u,\sigma) = \begin{bmatrix} \partial_t u + \operatorname{div}_x \sigma \\ \sigma + \nabla_x u \end{bmatrix},$$

and the right-hand side $\mathcal{F} := [f,0]^T$. Then the operator equation of the first order system (4.8) reads to find $(u,\sigma) \in U$ such that

$$\mathcal{B}(u,\sigma) = \mathcal{F} \quad \text{in } V. \tag{4.9}$$

For the solution of (4.9) one considers a least-squares approach which seeks for a minimizer $(u,\sigma) \in U$ to the functional

$$J(v,\tau) = \frac{1}{2}\|\mathcal{B}(v,\tau) - \mathcal{F}\|_V^2 = \frac{1}{2}\left[\|\partial_t v + \operatorname{div}_x \tau - f\|_{L^2(Q)}^2 + \|\tau + \nabla_x v\|_{L^2(Q)}^2\right].$$

The minimizer of this functional solves the first-order optimality system

$$\mathcal{B}^*\mathcal{B}(u,\sigma) = \mathcal{B}^*\mathcal{F} \quad \text{in } U^*, \tag{4.10}$$

i.e., we have to solve the variational formulation

$$\langle \mathcal{B}(u,\sigma), \mathcal{B}(v,\tau)\rangle_V = \langle \mathcal{F}, \mathcal{B}(v,\tau)\rangle_V \quad \text{for all } (v,\tau) \in U, \tag{4.11}$$

which in more detail reads

$$\begin{aligned} b(u,\sigma;v,\tau) := \langle \partial_t u + \operatorname{div}_x \sigma, \partial_t v + \operatorname{div}_x \tau\rangle_{L^2(Q)} + \langle \sigma + \nabla_x u, \tau + \nabla_x v\rangle_{L^2(Q)} \\ = \langle f, \partial_t v + \operatorname{div}_x \tau\rangle_{L^2(Q)} =: l(v,\tau) \end{aligned} \tag{4.12}$$

for all $(v,\tau) \in U$. For the variational formulation (4.11), (4.12), respectively, it was shown in [72, Lem. 4] that the bilinear form b is bounded and elliptic on U, i.e., there exist constants $c_1^S, c_2^{\mathcal{B}} > 0$ such that

$$b(u,\sigma;v,\tau) \le c_2^{\mathcal{B}} \|(u,\sigma)\|_U \|(v,\tau)\|_U, \quad b(u,\sigma;u,\sigma) \ge c_1^S \|(u,\sigma)\|_U^2$$

for all $(u,\sigma),(v,\tau) \in U$, and that the linear functional l is bounded. These properties ensure that the variational formulation (4.12) obtains a unique solution, see [72, Thm. 5] and hence the least-squares minimization problem is well-posed. The ellipticity of the bilinear form b gives that the corresponding operator $\mathcal{B}$ is injective since it holds

$$\begin{aligned} \sup_{0\neq(w,\varrho)\in V} \frac{\langle \mathcal{B}(u,\sigma),(w,\varrho)\rangle_V}{\|(w,\varrho)\|_V} &\ge \frac{\langle \mathcal{B}(u,\sigma),\mathcal{B}(u,\sigma)\rangle_V}{\|\mathcal{B}(u,\sigma)\|_V} \\ &= \frac{b(u,\sigma;u,\sigma)}{\|\mathcal{B}(u,\sigma)\|_V} \\ &\ge \frac{c_1^S}{c_2^{\mathcal{B}}} \|(u,\sigma)\|_U. \end{aligned}$$

Note that this is sufficient for a least-squares method in order to be well-posed. It was shown in [81, Thm. 2.3] that $\mathcal{B}$ is indeed surjective and hence an isomorphism. Therefore, solving the operator equation (4.9) is equivalent to solving the minimization problem, in particular (4.10). Finally, we note that any $f \in Y^*$ can be expressed as $f = f_1 - \operatorname{div}_x f_2$ for $f_1 \in L^2(Q)$, $f_2 \in L^2(Q)^d$. This equality has to be understood in the distributional sense, i.e.,

$$\langle f, q\rangle_Q = \langle f_1, q\rangle_{L^2(Q)} + \langle f_2, \nabla_x q\rangle_{L^2(Q)} \quad \text{for all } q \in Y. \tag{4.13}$$

The quantities f_1, f_2 can be obtained by setting $f_1 = w$, $f_2 = \nabla_x w$, where $w \in Y$ solves

$$\langle \nabla_x w, \nabla_x q\rangle_{L^2(Q)} + \langle w, q\rangle_{L^2(Q)} = \langle f, q\rangle_Q$$

for all $q \in Y$. Using this splitting the following holds.

LEMMA 4.6. *Let $f \in Y^*$ be given as in (4.13), where $f_1 \in L^2(Q)$ and $f_2 \in [L^2(Q)]^d$. Then $u \in X$ solves the variational problem of Bochner type (4.2) for the heat equation and $\sigma = -\nabla_x u + f_2$ if and only if $(u, \sigma) \in U$ solves (4.9) with right-hand side $\mathcal{F} = [f_1, f_2]^\top$.*

Proof. A proof can be found in [81, Proposition 2.5]. □

The result of Lemma 4.6 states that the first order formulation of the heat equation (4.9) applies whenever the space-time variational formulation (4.2) does, see [82]. Note that the right-hand side l in (4.12) changes to

$$l(v, \tau) = \langle f_1, \partial_t v + \operatorname{div}_x \tau \rangle_{L^2(Q)} + \langle f_2, \tau + \nabla_x v \rangle_{L^2(Q)},$$

when considering a splitting of $f \in Y^*$ as in (4.13).

In what follows we consider the finite dimensional subspace $U_h \subset U$, which is defined with respect to an admissible decomposition $\mathcal{T}_h$ of Q into simplicial elements, and we assume $f \in L^2(Q)$. The discrete variational problem of (4.12) then reads to find $(u_h, \sigma_h) \in U_h$ such that

$$b(u_h, \sigma_h; v_h, \tau_h) = l(v_h, \tau_h) \quad \text{for all } (v_h, \tau_h) \in U_h. \tag{4.14}$$

Due to the ellipticity of b the variational formulation is well-posed, and we have the best approximation result [72, Thm. 5]

$$\|(u - u_h, \sigma - \sigma_h)\|_U \leq \frac{c_2^{\mathcal{B}}}{c_1^S} \inf_{(v_h, \tau_h) \in U_h} \|(u - v_h, \sigma - \sigma_h)\|_U.$$

For the discrete space $U_h = S_h^1(\mathcal{T}_h) \cap X \times [S_h^1(\mathcal{T}_h)]^d$ there holds for a smooth solution $u \in L^2(0,T; H_0^1(\Omega)) \cap H^1(0,T; H^2(\Omega)) \cap H^2(0,T; L^2(\Omega)) \cap L^\infty(0,T; H^3(\Omega))$ the error behaviour, see [72, Cor. 15]

$$\|(u - u_h, \sigma - \sigma_h)\|_U = \mathcal{O}(h).$$

For the solution (u_h, σ_h) of (4.12) an a posteriori error indicator can be defined via the least-squares functional

$$\eta_h^2(u_h, \sigma_h) = J(u_h, \sigma_h) = \sum_{K \in \mathcal{T}_h} \eta_K^2(u_h, \sigma_h) \tag{4.15}$$

with local error indicators

$$\eta_K^2(u_h, \sigma_h) := \frac{1}{2} \left[\|\partial_t u_h + \operatorname{div}_x \sigma_h - f\|_{L^2(K)}^2 + \|\sigma_h + \nabla_x u_h\|_{L^2(K)}^2 \right].$$

The error indicator is efficient and reliable which can be seen from the following lemma.

LEMMA 4.7. *Let $f \in L^2(Q)$. Further, let $(u, \sigma) \in U$ be the unique solution* (4.12) *and $(u_h, \sigma_h) \in U_h$ be the unique solution to* (4.14). *Then the error indicator η_h defined in* (4.15) *is efficient and reliable, i.e., there exist constants $C_c, C_b > 0$ such that*

$$C_c \|(u - u_h, \sigma - \sigma_h)\|_U^2 \leq \eta_h^2(u_h, \sigma_h) \leq C_b \|(u - u_h, \sigma - \sigma_h)\|_U^2.$$

Proof. A proof is given in [72, Thm. 17] □

4.1.3 Numerical examples

In this section we consider numerical examples for the developed minimal residual framework for the heat equation. The inbuilt error indicator will be used to drive an adaptive refinement scheme and the results will be compared with those of a uniform refinement scheme. Furthermore, we provide a comparison to the FOSLS method proposed by Führer and Karkulik in [72]. The underlying finite element spaces used in our computations are defined with respect to an admissible and locally quasi-uniform decomposition $\mathcal{T}_H$ of Q into shape regular simplicial elements.

Minimal residual method

We consider (4.7) with the trial space $X_H = S_H^1(\mathcal{T}_H) \cap X$ of piecewise linear and globally continuous functions and the test space $Y_h = Y_H = S_H^2(\mathcal{T}_H) \cap Y$ of piecewise quadratic and globally continuous functions. Note that the latter is defined with respect to the mesh of local mesh size H. However, since we use piecewise quadratic basis functions the test space can be identified with a space of piecewise linear functions which are defined with respect to a refined mesh of local mesh size $h = H/2$, cf. [106]. In the adaptive refinement scheme we use the global error indicator $p_h \in Y_h$. Due to the choice of the test spaces it allows the representation

$$\eta_H^2 = \|p_h\|_Y^2 = \langle \nabla_x p_h, \nabla_x p_h \rangle_{L^2(Q)} = \sum_{\tau \in \mathcal{T}_H} \langle \nabla_x p_h, \nabla_x p_h \rangle_{L^2(\tau)} = \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2, \tag{4.16}$$

with the local error indicators

$$\eta_\tau^2 = \langle \nabla_x p_h, \nabla_x p_h \rangle_{L^2(\tau)} \quad \text{for } \tau \in \mathcal{T}_H.$$

If not other stated we use as a marking strategy the Dörfler criterion [58] with parameter $\theta = 0.5$. This criterion seeks for a minimal set of elements $\mathcal{M} \subset \mathcal{T}_H$ such that

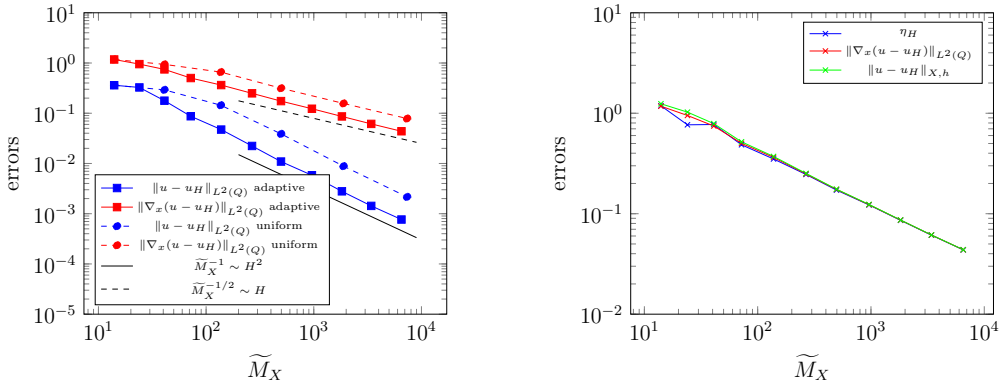
$$\sum_{\tau \in \mathcal{M}} \eta_\tau^2 \geq \theta \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2.$$

The selected space-time simplicial elements are refined using newest vertex bisection. All computations were done in the software `Netgen/NGSolve` [152], where we used the sparse direct solver `Pardiso` [148, 149, 150] to solve the resulting linear systems.

In the first example we use, as in [106], the one-dimensional spatial domain $\Omega = (0,3)$ and the time horizon $T = 6$, i.e., we have the space-time domain $Q := (0,3) \times (0,6) \subset \mathbb{R}^2$. As exact solution we consider the smooth function

$$u(x,t) := \begin{cases} \frac{1}{2}(t-x-2)^3(x-t)^3 \sin \frac{\pi}{3}x & \text{for } x \leq t \text{ and } t-x \leq 2, \\ 0 & \text{else,} \end{cases} \tag{4.17}$$

and we compute $f = \partial_t u - \Delta_x u$ accordingly. The exact solution u is smooth. Hence, we expect to see a rate of $\mathcal{O}(H^2)$ for the error in $L^2(Q)$ and a rate of $\mathcal{O}(H)$ for the error measured in the energy norm $\|\nabla_x \cdot\|_{L^2(Q)}$. These rates are confirmed in our numerical experiment for both a uniform and an adaptive refinement strategy, see Fig. 4.1a, where we plotted the errors against the number of vertices $\widetilde{M}_X = \dim(S^1_H(\mathcal{T}_H))$. Further, we see that for the same amount of degrees of freedom we have in the adaptive case a higher accuracy than in the uniform case, which is an expected behaviour. In Fig. 4.1b a comparison between the error estimator $\eta_H = \|\nabla_x p_h\|_{L^2(Q)}$ and the errors $\|\nabla_x(u-u_H)\|_{L^2(Q)}$, $\|u-u_H\|_{X,h}$, respectively, is provided. It demonstrates that the error indicator is effective. Finally, in Fig. 4.2 we present



a) Errors $\|\nabla_x(u-u_H)\|_{L^2(Q)}$ and $\|u-u_H\|_{L^2(Q)}$ for uniform and adaptive refinement strategies.

b) Comparison of the error estimator and true errors for an adaptive refinement strategy.

Figure 4.1: Convergence results in the case of a smooth solution for the heat equation.

the related space-time finite element meshes. The adaptive mesh reflects the fact that the solution behaves like a travelling wave.

As a second example we consider, as in [106], $\Omega = 1$, T=1, i.e., the unit square $Q = (0,1)^2 \subset \mathbb{R}^2$ as a space-time domain. For the right-hand side we consider the

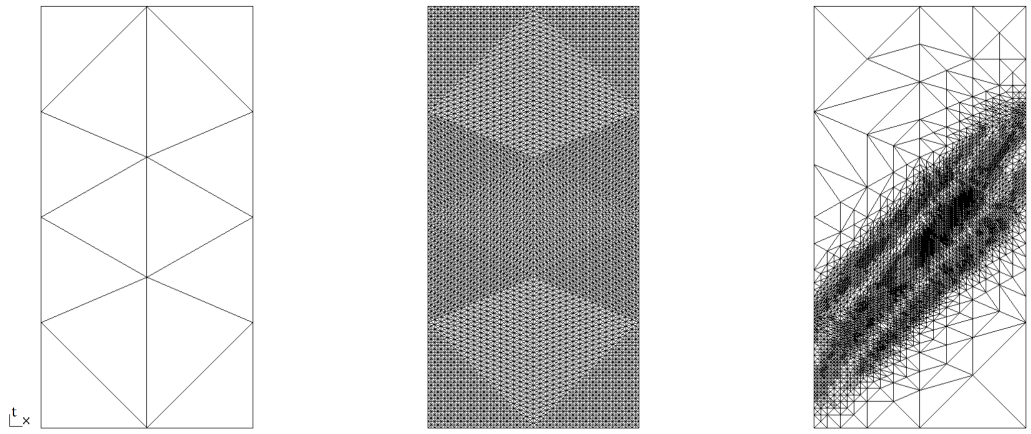


Figure 4.2: Space-time finite element meshes. Left: Initial mesh. Middle: Uniform refinement $L = 5$. Right: Adaptive refinement $L = 10$.

discontinuous function

$$f(x,t) = \begin{cases} 1 & (x,t) \in \left\{(x,t) \in (0,1) \times \left(\frac{1}{10}, \frac{1}{2}\right) : \; x - \frac{1}{10} \leq t \leq x - \frac{1}{20}\right\}, \\ 0 & \text{else}, \end{cases} \tag{4.18}$$

and we choose homogeneous initial and boundary conditions. Note that this example is also considered in [72, Sec. 5.2.3]. For this example the exact solution is unknown. Thus, we only provide error rates for the error estimator $\eta_H = \|\nabla_x p_h\|_{L^2(Q)}$. These rates can be seen in Fig. 4.3. In case of a uniform refinement strategy we obtain a reduced rate of $\mathcal{O}(H^{\frac{1}{2}})$, while in the adaptive case we have the optimal rate $\mathcal{O}(H)$. These rates are also observed in [72]. In Fig. 4.4 we depict the numerical solution u_H as well as the adaptive generated space-time finite element mesh. Here, stronger refinements around the support of the function f can be observed as in [72], see also Fig. 4.9 in case of our own implementation of the FOSLS method.

In the third example for the heat equation we consider a problem with incompatible initial datum similar as in [72, Sec. 5.2.4] and [106]. To be more precise, we consider $Q = (0,1)^2$, $f(x,t) = 2$ for $(x,t) \in (0,1)^2$ and an inhomogeneous inital datum $u_0(x) = 1$ for $x \in (0,1)$. Obviously, we have $u_0 \in L^2(\Omega)$, but $u_0 \notin H_0^1(\Omega)$. This means there is no compatibility with the homogeneous Dirichlet boundary condition for $t = 0$. Therefore, we expect to see a reduced rate. The numerical implementation of the inhomogeneous initial datum is done via a homogenization approach similar as for inhomogeneous Dirichlet boundary conditions. This is possible because in the space-time setting the initial condition acts like a Dirichlet condition on the initial boundary. In Fig. 4.5 the rates for the error indicators as well as the adaptive generated mesh are depicted. We observe a reduced rate of $\mathcal{O}(H^{\frac{1}{4}})$ in the uniform case which is improved to $\mathcal{O}(H^{0.8})$ in the adaptive case. These rates are better than

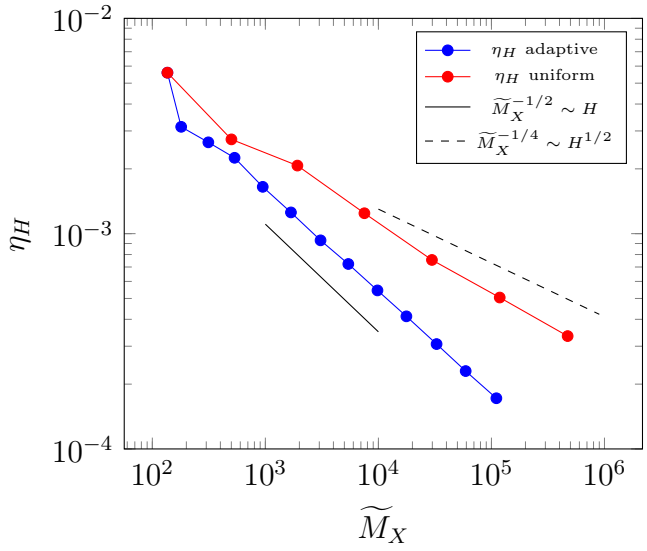
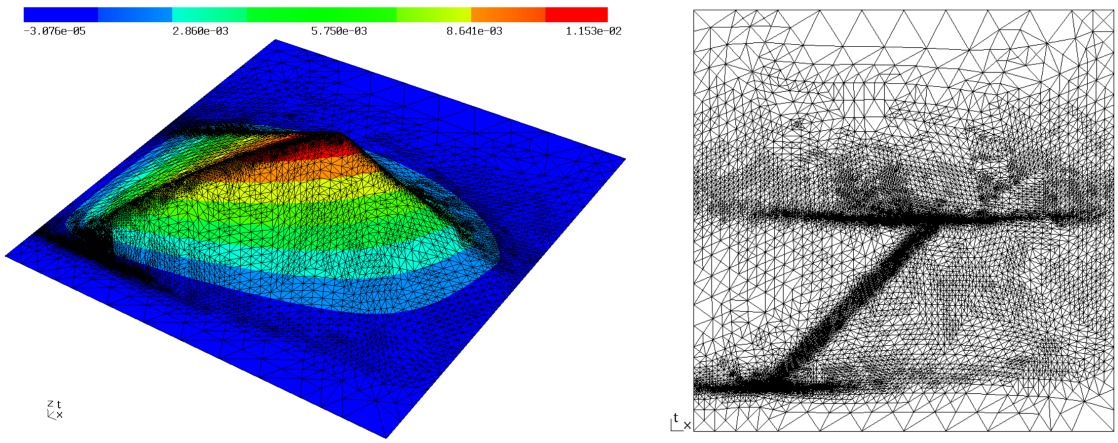


Figure 4.3: Error estimator η_H in the case of a discontinuous right-hand side.



a) Solution u_H b) Adaptive mesh with $\widetilde{M}_X = 17689$

Figure 4.4: Singular solution of the heat equation for the discontinuous right-hand side (4.18).

those observed in [72, Section 5.2.4] and similar as in [82, Section 4.2.1]. The adaptive space-time mesh shows stronger refinements around the incompatibility of the initial datum with the homogeneous Dirichlet datum.

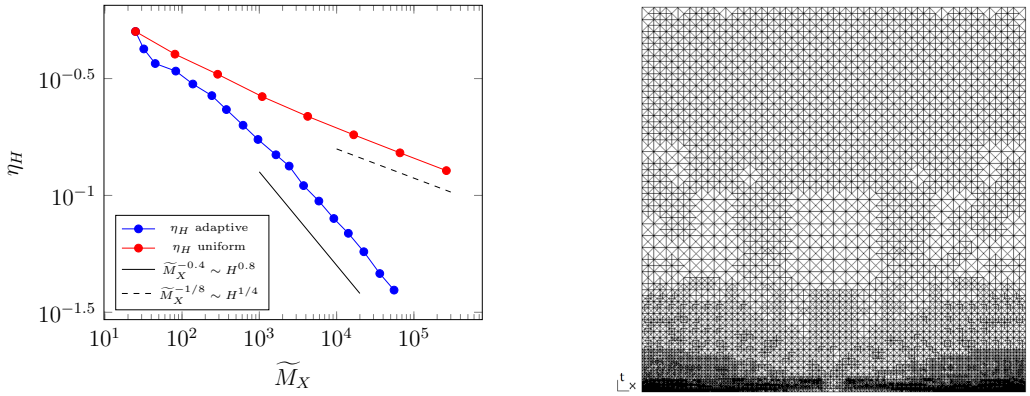
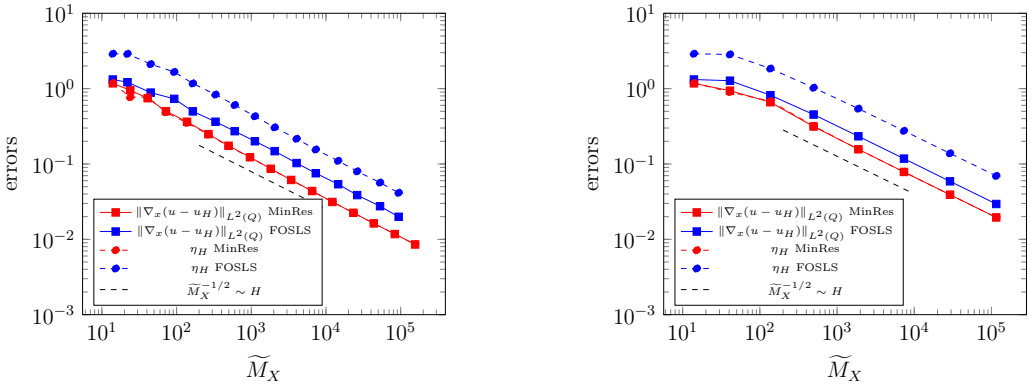


Figure 4.5: Numerical results in case of an incompatible initial condition $u_0 \notin H_0^1(\Omega)$.

Comparison with FOSLS method

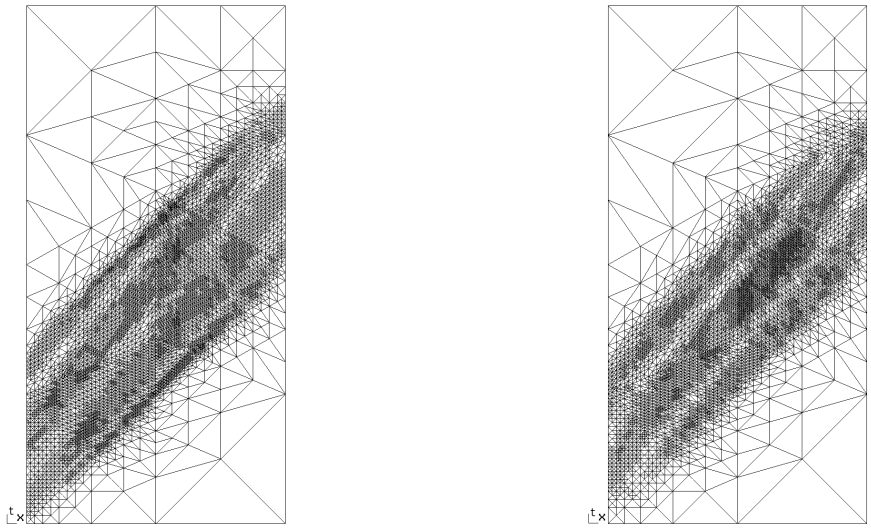
In the last part of this section we want to compare the results obtained from the minimal residual framework, see Section 4.1.1 with those of the FOSLS method, see 4.1.2 in more detail. For this reason we implemented the FOSLS method within the finite element software `Netgen/NGSolve` [152], where we used the sparse direct solver `Pardiso` [149]. For the discretization we use the space $U_H = S_H^1(\mathcal{T}_H) \cap X \times [S_H^1(\mathcal{T}_H)]^d$. The adaptive refinement scheme in the FOSLS method is driven by the error indicator η_H defined in (4.15), where we use the Dörfler criterion [58] with parameter $\theta = 0.5$. The marked elements are refined using newest vertex bisection.

In the first comparison we revisit the example with the smooth solution (4.17), which corresponds to a travelling wave in the space-time domain $Q = (0, 3) \times (0, 6)$. In Fig. 4.6 we present the related errors in the energy norm as well as the error indicators for both a uniform and an adaptive refinement scheme. In both cases we see a linear rate $\mathcal{O}(H)$ for the errors and the indicators. Further, we observe that in the minimal residual method the value of the true error $\|\nabla_x(u - u_H)\|_{L^2(Q)}$ is smaller compared to the value in the FOSLS method for the same amount of dofs. This means that the minimal residual method delivers a more accurate solution. In addition to that we see that the error estimator (4.16) used in the minimal residual method estimates the value of the true error better than the estimator (4.15) used in the FOSLS method. In Fig. 4.7 we present the related adaptive refined space-time finite element meshes. Both meshes are created from the same initial mesh, which can be seen in Fig. 4.2. The mesh obtained from the minimal residual method seems a bit more concentrated



a) adaptive b) uniform

Figure 4.6: Comparison of the numerical results from the minimal residual (MinRes) method and the FOSLS method in case of the smooth solution (4.17).



a) FOSLS, $L = 10$, $\widetilde{M}_X = 7301$ dofs b) MinRes, $L = 10$, $\widetilde{M}_X = 6524$ dofs

Figure 4.7: Comparison of the generated adaptive space-time finite element meshes for the smooth solution (4.17).

at the turning point of the travelling wave while the FOSLS method seems to resolve the interface where the solution has a steep gradient a bit better.

As a second comparison we revisit the example with the discontinuous right-hand side f given in (4.18) on the unit square $Q = (0,1)^2$. As mentioned this example is also considered in [72, Sec. 5.2.3]. The numerical results are depicted in Fig. 4.8. We can confirm the rates observed in [72] also for our own implementation of the FOSLS

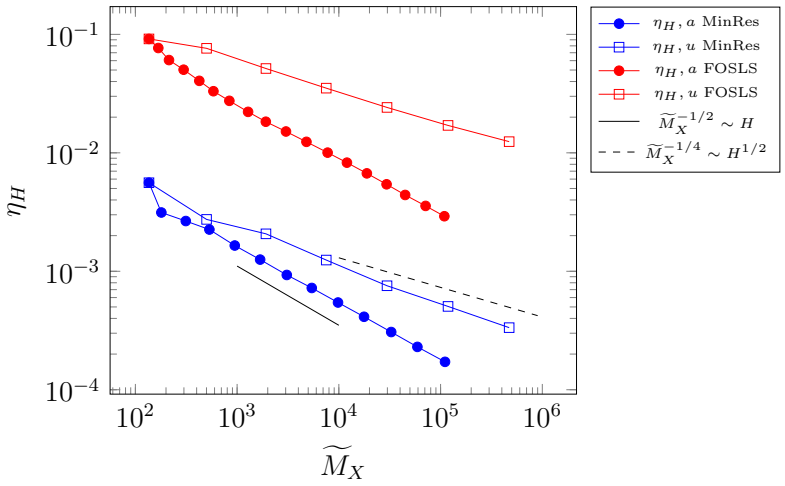
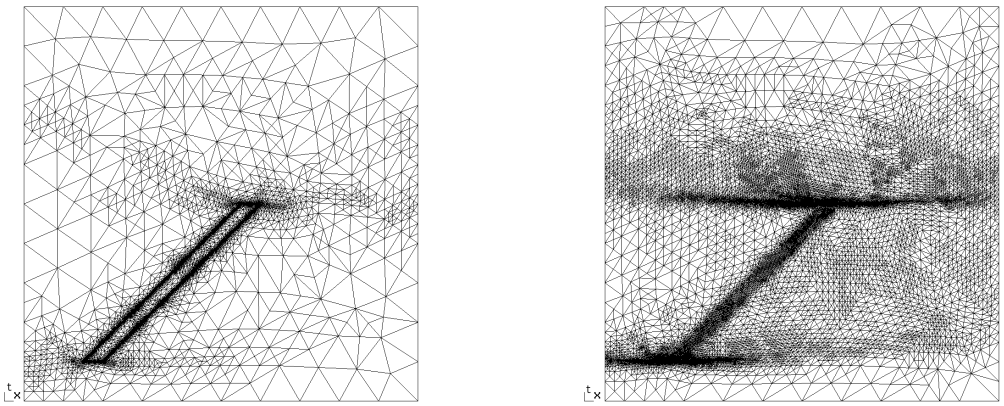


Figure 4.8: Numerical results for the piecewise constant right-hand side f given in (4.18) for the minimal residual method and the FOSLS method.



a) FOSLS, $L = 13$, $\widetilde{M}_X = 18856$ dofs

b) MinRes, $L = 10$, $\widetilde{M}_X = 17689$ dofs

Figure 4.9: Comparison of the generated adaptive space-time finite element meshes for the piecewise constant function (4.18).

method. Moreover, in Fig. 4.9 we provide a comparison of the generated adaptive space-time meshes. We started again with the same initial mesh in both cases, which had $\widetilde{M}_X = 136$ dofs. We can see that both meshes show stronger refinements around the support of f. However, in the case of the FOSLS method the refinement is more concentrated along the interface of the function f.

4.2 Convection-diffusion equation

In this section we consider the nonstationary advection-diffusion problem

$$\begin{aligned} \partial_t u(x,t) - \varepsilon \Delta_x u(x,t) + \beta(x,t) \cdot \nabla_x u(x,t) &= f(x,t) \quad \text{for } (x,t) \in Q, && \text{(4.19a)} \\ u(x,t) &= 0 \quad \text{for } (x,t) \in \Sigma, && \text{(4.19b)} \\ u(x,0) &= 0 \quad \text{for } x \in \Omega. && \text{(4.19c)} \end{aligned}$$

Here, $Q := \Omega \times (0,T)$ is the space-time cylinder, where $\Omega \subset \mathbb{R}^d$, $d = 1,2,3$ is a bounded Lipschitz domain with boundary $\Gamma = \partial\Omega$, and $T > 0$ is a given time horizon. We consider $\varepsilon > 0$ to be a given parameter, and $\beta \in L^\infty(0,T; W^{1,\infty}(\Omega, \mathbb{R}^d))$ some given velocity field. We assume that the velocity field satisfies $\operatorname{div}_x(\beta) = 0$. Note that this corresponds to the conservation of mass property for incompressible flows. As mentioned in [117, 141] this problem arises in many applications, e.g., in pollution simulations, in the modeling of heat and flow problems, in heat transfer problems with respect to thin domains or in semiconductor device simulations, just to name a few. In particular, we are interested in the case $\frac{\varepsilon}{\|\beta\|} \ll 1$, which means that this problem gets advection dominated. In this case the solutions are characterized by boundary layers and hence the numerical solution of this problem becomes difficult [141] as standard finite element discretization schemes lead to oscillatory solutions unless the mesh size is chosen in the order of ε, i.e., $h \sim \varepsilon$. To illustrate this effect we consider similar as in [141, Ex. 1.2] the boundary value problem

$$\begin{aligned} -\varepsilon u''(x) + u'(x) &= f \quad \text{on } (0,1) \\ u(0) = u(1) &= 0, \end{aligned} \tag{4.20}$$

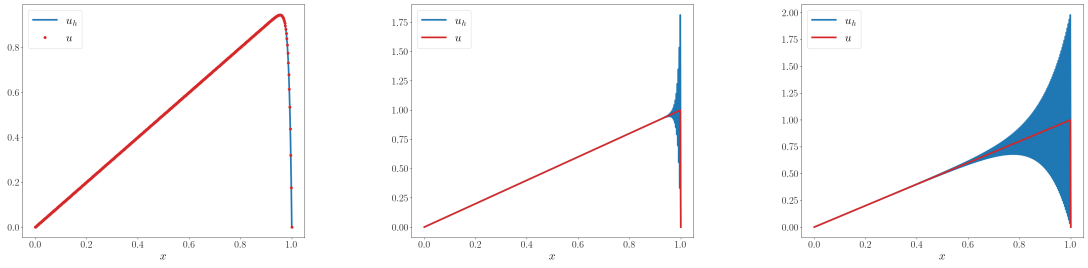
with the exact solution

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}. \tag{4.21}$$

The right-hand side f is computed accordingly and reads $f = 1$. Note that the exact solution has a boundary layer at $x = 1$ and that $\beta = 1$. A standard finite element discretization with piecewise linear finite elements for (4.20) reads to find $u_h \in V_h := S_h^1(0,1) \cap H_0^1(0,1)$ such that

$$b(u_h, v_h) := \langle \varepsilon u_h', v_h' \rangle_{L^2(0,1)} + \langle u_h', v_h \rangle_{L^2(0,1)} = \langle f, v_h \rangle_{L^2(0,1)} \quad \text{for all } v_h \in V_h. \tag{4.22}$$

The convergence behaviour of the error $\|\nabla(u - u_h)\|_{L^2(0,1)}$ is depicted in Tab. 4.1 and Fig. 4.11. We see that we obtain linear convergence if the mesh size h is of the order of the parameter ε. Before, the numerical solutions u_h obtain oscillations and are unphysical, see Fig. 4.10, where we plotted the numerical solution u_h for $N_V = 512$ elements and $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-5}\}$. In applications, choosing $h \sim \varepsilon$



a) $\varepsilon = 10^{-2}$ b) $\varepsilon = 10^{-4}$ c) $\varepsilon = 10^{-5}$

Figure 4.10: Numerical solution u_h for the variational formulation (4.22) in case of $N_V = 512$ elements.

$\varepsilon = 10^{-2}$			$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-5}$		
$\widetilde{M}_V$	$\|\nabla(u-u_h)\|_{L^2(0,1)}$	eoc	$\widetilde{M}_V$	$\|\nabla(u-u_h)\|_{L^2(0,1)}$	eoc	$\widetilde{M}_V$	$\|\nabla(u-u_h)\|_{L^2(0,1)}$	eoc
129	1.567e+00		4097	4.280e+01		65537	9.237e+01	
257	7.938e-01	0.981	8193	2.389e+01	0.841	131073	4.843e+01	0.931
513	3.982e-01	0.995	16385	1.233e+01	0.955	262145	2.452e+01	0.982
1025	1.993e-01	0.999	32769	6.213e+00	0.988	524289	1.230e+01	0.995
2049	9.966e-02	1.000	65537	3.113e+00	0.997	1048577	6.154e+00	0.999

Table 4.1: Numerical results for the variational formulation (4.22) in case of the function (4.21).

is impractical as this leads to unacceptably large numbers of mesh points. For this reason one is interested in robust methods that work for all values of the singular perturbation parameter ε and give physical correct solutions even before the mesh size h is in the order of ε. In literature different techniques have been proposed to achieve this, see e.g. [141]. There are methods like the Streamline-Diffusion (SD) [15, 141] or the Streamline-Upwind Petrov-Galerkin method (SUPG) [22, 29, 94, 141] which add some stabilizing terms to the bilinear form $b(u_h, v_h)$ and to the right-hand side. Other methods are based on the residual minimization idea. Here one can differ between methods which are proposed in a Petrov-Galerkin (PG) setting [28, 43, 56, 62], or in a first order least-squares setting (FOSLS) [44, 117] or the variational stabilization/saddle point least-squares technique [12, 42, 46]. We will focus on the latter approach and apply the abstract framework presented in Section 3.1 to (4.19).

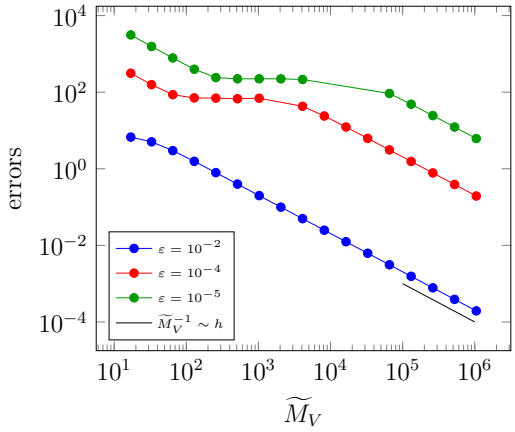


Figure 4.11: Convergence behaviour of the error $\|\nabla(u - u_h)\|_{L^2(0,1)}$ for a uniform refinement strategy.

4.2.1 Application of the abstract framework

We apply the abstract framework from Section 3.1 in the space-time setting considered in [157]. We consider the spaces

$$\begin{aligned} Y &:= L^2(0,T;H_0^1(\Omega)), \\ X &:= \{u \in Y : \ \partial_t u + \beta \cdot \nabla_x u \in Y^*, \ u(x,0) = 0, \ x \in \Omega\}, \\ H &:= L^2(Q), \end{aligned} \tag{4.23}$$

equipped with the norms

$$\|p\|_Y := \langle \varepsilon \nabla_x p, \nabla_x p \rangle_{L^2(Q)}, \ |||u|||_X = \sqrt{\|\partial_t u + \beta \cdot \nabla_x u\|_{Y^*}^2 + \|u\|_Y^2}.$$

Introducing $w_u \in Y$ as solution to

$$\langle \varepsilon \nabla_x w_u, \nabla_x q \rangle_{L^2(Q)} = \langle \partial_t u + \beta \cdot \nabla_x u, q \rangle_Q \quad \text{for all } q \in Y, \tag{4.24}$$

we have

$$\|\partial_t u + \beta \cdot \nabla_x u\|_{Y^*} = \|w_u\|_Y, \quad |||u|||_X = \sqrt{\|w_u\|_Y^2 + \|u\|_Y^2}.$$

Note that the most standard choice for the space X would be to consider, as for the heat equation, the space $L^2(0,T;H_0^1(\Omega)) \cap H_{0,}^1(0,T;H^{-1}(\Omega))$. The space X defined in (4.23) allows for the consideration of more general velocity fields β. However, in case of bounded velocity fields β, which we assume, we have

$$X = L^2(0,T;H_0^1(\Omega)) \cap H_{0,}^1(0,T;H^{-1}(\Omega)),$$

and we can provide the following result.

LEMMA 4.8. *In case of bounded velocity fields β there holds the norm equivalence*

$$\frac{1}{\max\{(1+2c(\varepsilon)^2),2)\}}\|u\|_X^2 \leq |||u|||_X^2 \leq \max\{(1+2c(\varepsilon)^2),2)\}\|u\|_X^2, \tag{4.25}$$

where $c(\varepsilon) = \frac{1}{\varepsilon}c_p\sqrt{d}\|\beta\|_{L^\infty(Q)}$ *and* $\|u\|_X^2 = \|\partial_t u\|_{Y^*}^2 + \|\nabla_x u\|_{L^2(Q)}^2$.

Proof. Using the Cauchy-Schwarz inequality and a spatial Poincaré inequality, see (2.6) we can estimate

$$\begin{aligned}
\langle \beta\cdot\nabla_x u, q\rangle_{L^2(Q)} &= \int_0^T\int_\Omega \sum_{i=1}^d \beta_i\partial_{x_i}uq\mathrm{d}x\mathrm{d}t\\
&\leq \|\beta\|_{L^\infty(Q)}\sum_{i=1}^d\int_0^T\int_\Omega |\partial_{x_i}uq|\mathrm{d}x\mathrm{d}t\\
&\leq \|\beta\|_{L^\infty(Q)}\sum_{i=1}^d \|\partial_{x_i}u\|_{L^2(Q)}\|q\|_{L^2(Q)}\\
&\leq \|\beta\|_{L^\infty(Q)}\|q\|_{L^2(Q)}\sqrt{d}\|\nabla_x u\|_{L^2(Q)}\\
&\leq \|\beta\|_{L^\infty(Q)}c_p\sqrt{d}\|\nabla_x u\|_{L^2(Q)}\|\nabla_x q\|_{L^2(Q)}\\
&= \frac{1}{\varepsilon}\|\beta\|_{L^\infty(Q)}c_p\sqrt{d}\|\varepsilon^{1/2}\nabla_x u\|_{L^2(Q)}\|\varepsilon^{1/2}\nabla_x q\|_{L^2(Q)} = c(\varepsilon)\|u\|_Y\|q\|_Y,
\end{aligned}$$

with $c(\varepsilon) = \frac{1}{\varepsilon}c_p\sqrt{d}\|\beta\|_{L^\infty(Q)}$. Thus, it holds

$$\|\beta\cdot\nabla_x u\|_{Y^*} = \sup_{0\neq q\in Y}\frac{\langle\beta\cdot\nabla_x u, q\rangle_{L^2(Q)}}{\|q\|_Y} \leq c(\varepsilon)\|u\|_Y.$$

Now, we can conclude using standard estimates the upper inequality

$$\begin{aligned}
|||u|||_X^2 &= \|u\|_Y^2 + \|\partial_t u + \beta\cdot\nabla_x u\|_{Y^*}^2\\
&\leq \|u\|_Y^2 + 2\|\partial_t u\|_{Y^*}^2 + 2\|\beta\cdot\nabla_x u\|_{Y^*}^2\\
&\leq (1+2c(\varepsilon)^2)\|u\|_Y^2 + 2\|\partial_t u\|_{Y^*}^2\\
&\leq \max\{(1+2c(\varepsilon)^2),2\}\|u\|_X^2.
\end{aligned}$$

On the other hand we have for the lower inequality

$$\begin{aligned}
\|u\|_X^2 &= \|u\|_Y^2 + \|\partial_t u\|_{Y^*}^2\\
&\leq \|u\|_Y^2 + 2\|\partial_t u + \beta\cdot\nabla_x u\|_{Y^*}^2 + 2\|\beta\cdot\nabla_x u\|_{Y^*}^2\\
&\leq (1+2c(\varepsilon)^2)\|u\|_Y^2 + 2\|\partial_t u + \beta\cdot\nabla_x u\|_{Y^*}^2\\
&\leq \max\{(1+2c(\varepsilon)^2),2)\}|||u|||_X^2
\end{aligned}$$

and thus the norm equivalence follows. □

The operators $A : Y \to Y^*$ and $B : X \to Y^*$ are defined in the variational sense satisfying

$$\begin{aligned} \langle Ap, q\rangle_Q &:= \langle \varepsilon \nabla_x p, \nabla_x q\rangle_{L^2(Q)}, \\ \langle Bu, q\rangle_Q &:= \langle \partial_t u + \beta \cdot \nabla_x u, q\rangle_Q + \langle \varepsilon \nabla_x u, \nabla_x q\rangle_{L^2(Q)}, \end{aligned} \tag{4.26}$$

for all $p, q \in Y$ and $u \in X$. The operator $A : Y \to Y^*$ is self-adjoint bounded and elliptic with constants $c_1^A = c_2^A = 1$. This can be proven in the same way as demonstrated in Lemma 4.1. For the operator $B : X \to Y^*$ we can show the following properties.

LEMMA 4.9. *The operator $B : X \to Y^*$ fulfills:*

(i) B is bounded with $c_2^B = \sqrt{2}$, i.e., it holds $\|Bv\|_{Y^*} \leq \sqrt{2}|||v|||_X$ for all $v \in X$,

(ii) B is inf-sup stable with $c_1^B = 1$, i.e., it satisfies

$$|||u|||_X \leq \sup_{0 \neq q \in Y} \frac{\langle Bu, q\rangle_Q}{\|q\|_Y}$$

for all $u \in X$,

(iii) B is surjective.

Proof. We start with the proof of (i). As in the proof of Lemma 4.2 we can estimate

$$\begin{aligned} \langle Bu, q\rangle_Q &:= \langle \partial_t u + \beta \cdot \nabla_x u, q\rangle_Q + \langle \varepsilon \nabla_x u, \nabla_x q\rangle_{L^2(Q)} \\ &\leq \|\partial_t u + \beta \cdot \nabla_x u\|_{Y^*} \|q\|_Y + \|u\|_Y \|q\|_Y \leq \sqrt{2}|||u|||_X \|q\|_Y, \end{aligned}$$

for all $u \in X$, $q \in Y$, which gives the boundedness of B. The proof of (ii) can also be done as in Lemma 4.2 with some small adaptations. We consider the specific test function $\bar{q} := u + w_u \in Y$, where $w_u \in Y$ solves (4.24). Then, we have

$$\begin{aligned} \langle Bu, \bar{q}\rangle_Q &= \langle \partial_t u + \beta \cdot \nabla_x u, \bar{q}\rangle_Q + \langle \varepsilon \nabla_x u, \nabla_x \bar{q}\rangle_{L^2(Q)} \\ &= \langle \varepsilon \nabla_x w_u, \nabla_x \bar{q}\rangle_{L^2(Q)} + \langle \varepsilon \nabla_x u, \nabla_x \bar{q}\rangle_{L^2(Q)} \\ &= \|\bar{q}\|_Y^2. \end{aligned}$$

Now, using integration by parts and that by assumption $\mathrm{div}_x(\beta) = 0$, we obtain

$$\begin{aligned} \langle \partial_t u + \beta \cdot \nabla_x u, u\rangle_Q &= \langle \partial_t u, u\rangle_Q + \langle \beta \cdot \nabla_x u, u\rangle_{L^2(Q)} \\ &= \langle \partial_t u, u\rangle_Q - \frac{1}{2}\big\langle \mathrm{div}_x(\beta), u^2\big\rangle_{L^2(Q)} \geq 0, \end{aligned}$$

which gives

$$\begin{aligned}\|\bar{q}\|_Y^2 &= \|u + w_u\|_Y^2 \\ &= \|u\|_Y^2 + \|w_u\|_Y^2 + 2\langle \varepsilon \nabla_x w_u, \nabla_x u\rangle_{L^2(Q)} \\ &= |||u|||_X^2 + 2\langle \partial_t u + \beta \cdot \nabla_x u, u\rangle_Q \\ &\geq |||u|||_X^2.\end{aligned}$$

This implies

$$\langle Bu, \bar{q}\rangle_Q \geq |||u|||_X \|\bar{q}\|_Y,$$

and therefore the inf-sup stability condition

$$|||u|||_X \leq \sup_{0\neq q\in Y} \frac{\langle Bu, q\rangle_Q}{\|q\|_Y} \quad \text{for all } u \in X$$

follows with $c_1^B = 1$. We continue with the proof of (iii). For the surjectivity of the operator B, let $w \in Y \setminus \{0\}$ and consider the auxiliary initial boundary value problem to find $u_w \in X$ such that

$$\langle Bu_w, q\rangle_Q = \langle \varepsilon \nabla_x w, \nabla_x q\rangle_{L^2(Q)} \quad \text{for all } q \in Y \tag{4.27}$$

is satisfied. The operator B allows the representation $B = \partial_t + \mathcal{A}$, with an operator $\mathcal{A} : Y \to Y^*$, defined in the weak sense via

$$\langle \mathcal{A}u_w, q\rangle_Q = \langle \varepsilon \nabla_x u_w, \nabla_x q\rangle_{L^2(Q)} + \langle \beta \cdot \nabla_x u_w, q\rangle_{L^2(Q)} \quad \text{for all } q \in Y,$$

Since the operator $\mathcal{A}$ is bounded

$$\langle \mathcal{A}p, q\rangle_Q \leq (1 + c(\varepsilon))\|p\|_Y \|q\|_Y, \quad c(\varepsilon) = \frac{1}{\varepsilon} c_p \sqrt{d} \|\beta\|_{L^\infty(Q)}$$

for all $p, q \in Y$ and elliptic with constant $c_1^{\mathcal{A}} = 1$, i.e.,

$$\langle \mathcal{A}p, p\rangle_Q \geq \|p\|_Y^2 \quad \text{for all } p \in Y,$$

an application of Theorem 2.10 yields a unique solution $u_w \in X$ to (4.27). Therefore, we have

$$\langle Bu_w, w\rangle_Q = \langle \varepsilon \nabla_x w, \nabla_x w\rangle_{L^2(Q)} > 0,$$

and hence surjectivity of B follows. □

Remark 4.10. *The proof of the inf-sup condition also stays true for velocity fields with* $-\frac{1}{2}\operatorname{div}_x(\beta) \geq 0$. *This is inline with other works concerning convection-diffusion equations, see e.g., [7, 33, 46].*

The properties of the operator B ensure that the variational formulation of the abstract operator equation $Bu = f$ which reads to find $u \in X$ such that

$$\langle \partial_t u + \beta \cdot \nabla_x u, q \rangle_Q + \langle \varepsilon \nabla_x u, \nabla_x q \rangle_{L^2(Q)} = \langle f, q \rangle_Q \quad \text{for all } q \in Y \tag{4.28}$$

admits a unique solution. Together with the properties of A we can apply the least-squares framework, which ends up in the mixed system to find $(u, p) \in X \times Y$ such that

$$\begin{aligned} \langle \varepsilon \nabla_x p, \nabla_x q \rangle_{L^2(Q)} + \langle \partial_t u + \beta \cdot \nabla_x u, q \rangle_Q + \langle \varepsilon \nabla_x u, \nabla_x q \rangle_{L^2(Q)} &= \langle f, q \rangle_Q, \\ \langle \partial_t v + \beta \cdot \nabla_x v, p \rangle_Q + \langle \varepsilon \nabla_x v, \nabla_x p \rangle_{L^2(Q)} &= 0, \end{aligned} \tag{4.29}$$

is satisfied for all $(v, q) \in X \times Y$.

For the discretization of (4.29) we consider finite dimensional subspaces $X_H \subset X$ and $Y_h \subset Y$, where we assume the inclusion $X_H \subset Y_h$. Similar as for the heat equation we define $w_{uh} \in Y_h$ as the unique solution of the variational formulation

$$\langle \varepsilon \nabla_x w_{uh}, \nabla_x q_h \rangle_{L^2(Q)} = \langle \partial_t u + \beta \cdot \nabla_x u, q_h \rangle_Q \quad \text{for all } q_h \in Y_h. \tag{4.30}$$

Note that (4.30) is the discrete variational formulation of (4.24). Hence, we have $\|w_{uh}\|_Y \leq \|w_u\|_Y$. Now we can define the mesh-dependent norm

$$|||u|||_{X,h} := \sqrt{\|u\|_Y^2 + \|w_{uh}\|_Y^2} \leq |||u|||_X. \tag{4.31}$$

This enables us to prove on the discrete level a refined boundedness estimate as well as a stability result for the operator B.

LEMMA 4.11. *Let $X_H \subset X$ and $Y_h \subset Y$ be finite dimensional subspaces of X, Y, respectively. Further, assume the inclusion $X_H \subset Y_h$. Then there holds*

(i)
$$\langle Bu, q_h \rangle_Q \leq \sqrt{2} |||u|||_{X,h} \|q_h\|_Y \quad \textit{for all } u \in X,\ q_h \in Y_h,$$

(ii)
$$|||u_H|||_{X,h} \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle Bu_H, q_h \rangle_Q}{\|q_h\|_Y} \quad \textit{for all } u_H \in X_H.$$

Proof. For the proof of (i) we use (4.30), the Cauchy-Schwarz inequality and the inequality $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \geq 0$. Thus, we obtain

$$\begin{aligned} \langle Bu, q_h \rangle_Q &= \langle \partial_t u + \beta \cdot \nabla_x u, q_h \rangle_Q + \langle \varepsilon \nabla_x u, \nabla_x q_h \rangle_{L^2(Q)} \\ &= \langle \varepsilon \nabla_x w_{uh}, \nabla_x q_h \rangle_{L^2(Q)} + \langle \varepsilon \nabla_x u, \nabla_x q_h \rangle_{L^2(Q)} \\ &\leq (\|w_{uh}\|_Y + \|u\|_Y) \, \|q_h\|_Y \\ &\leq \sqrt{2} |||u|||_{X,h} \|q_h\|_Y. \end{aligned}$$

The proof of (ii) can be done in the same way as in Lemma 4.9. We sketch the main ideas. We consider the specific test function $\bar{q}_h := u_H + w_{u_H h}$, where $w_{u_H h} \in Y_h$ solves (4.30). Due to the inclusion $X_H \subset Y_h$ we have $\bar{q}_h \in Y_h$. Then we can compute

$$\begin{aligned}\langle Bu_H, \bar{q}_h\rangle_Q &= \langle \partial_t u_H + \beta \cdot \nabla_x u_H, \bar{q}_h\rangle_Q + \langle \varepsilon \nabla_x u_H, \nabla_x \bar{q}_h\rangle_{L^2(Q)} \\ &= \langle \varepsilon \nabla_x w_{u_H h}, \nabla_x \bar{q}_h\rangle_{L^2(Q)} + \langle \varepsilon \nabla_x u_H, \nabla_x \bar{q}_h\rangle_{L^2(Q)} \\ &= \|\bar{q}_h\|_Y^2.\end{aligned}$$

Now, we have due to $\operatorname{div}_x(\beta) = 0$ that

$$\langle \varepsilon \nabla_x w_{u_H h}, \nabla_x u_H\rangle_{L^2(Q)} = \langle \partial_t u_H + \beta \cdot \nabla_x u_H, u_H\rangle_Q \geq 0.$$

This gives

$$\begin{aligned}\|\bar{q}_h\|_Y^2 &= \|u_H + w_{u_H h}\|_Y^2 \\ &= \|u_H\|_Y^2 + \|w_{u_H h}\|_Y^2 + 2\langle \varepsilon \nabla_x w_{u_H h}, \nabla_x u_H\rangle_{L^2(Q)} \\ &\geq |\!|\!| u_H |\!|\!|_{X,h}^2,\end{aligned}$$

which implies

$$\langle Bu_H, \bar{q}_h\rangle_Q \geq |\!|\!| u_H |\!|\!|_{X,h} \|\bar{q}_h\|_Y.$$

Hence, we conclude the discrete inf-sup stability condition

$$|\!|\!| u_H |\!|\!|_{X,h} \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle Bu_H, q_h\rangle_Q}{\|q_h\|_Y} \quad \text{for all } u_H \in X_H,$$

with $\tilde{c}_S = 1$. □

Now we are in a position to conclude unique solvability of the discrete variational formulation (3.13) which for the particular problem at hand reads to find $(u_H, p_h) \in X_H \times Y_h$ such that

$$\begin{aligned}\langle \varepsilon \nabla_x p_h, \nabla_x p_h\rangle_{L^2(Q)} + \langle \partial_t u_H + \beta \cdot \nabla_x u_H, q_h\rangle_Q + \langle \varepsilon \nabla_x u_H, \nabla_x q_h\rangle_{L^2(Q)} &= \langle f, q_h\rangle_Q, \\ \langle \partial_t v_H + \beta \cdot \nabla_x v_H, p_h\rangle_Q + \langle \varepsilon \nabla_x v_H, \nabla_x p_h\rangle_{L^2(Q)} &= 0\end{aligned} \tag{4.32}$$

is satisfied for all $(v_H, q_h) \in X_H \times Y_h$. An application of the abstract error estimate (3.36) gives the best approximation result

$$|\!|\!| u - u_H |\!|\!|_{X,h} \leq \sqrt{2} \inf_{v_H \in X_H} |\!|\!| u - v_H |\!|\!|_X. \tag{4.33}$$

REMARK 4.12. *The dependency on the parameter ε is hidden in the norm $|||\cdot|||_X$. Changing to the equivalent norm $\|\cdot\|_X$ will result in constants that depend on ε.*

In case of a piecewise linear finite element space for the trial space and a sufficient regular solution $u \in H^s(Q)$ for some $s \in [1,2]$ we can derive with similar techniques as in [157] the error estimate

$$\|u - u_H\|_Y \leq |||u - u_H|||_{X,h} \leq c(\varepsilon) H^{s-1} |u|_{H^s(Q)}. \tag{4.34}$$

Using $\bar{X}_H = Y_h \cap X$ we can show the discrete inf-sup condition (3.37) similar as in Lemma 4.11 with $\bar{c}_S = 1$. Now, an application of Lemma 3.21 gives for some $\eta \in (0,1)$ an efficiency and reliability estimate for the global error indicator $p_h \in Y_h$ which reads

$$\frac{1}{\sqrt{2}} \|p_h\|_Y \leq |||u - u_H|||_{X,h} \leq \frac{1}{1-\eta} \|p_h\|_Y. \tag{4.35}$$

4.2.2 Numerical examples for the nonstationary case

In this section we will consider the developed framework for the convection-diffusion equation in an adaptive refinement scheme and compare the results with those obtained from a uniform refinement scheme. For the discretization we choose the trial space $X_H = S^1_H(\mathcal{T}_H) \cap X$ of piecewise linear and globally continuous functions and the test space $Y_h = Y_H = S^2_H(\mathcal{T}_H) \cap Y$ of piecewise quadratic and globally continuous functions, which are defined with respect to an admissible and locally quasi-uniform decomposition $\mathcal{T}_H$ of Q into shape regular simplicial elements. In the adaptive refinement scheme we use the global error estimator p_h, which allows the representation

$$\eta_H^2 = \|p_h\|_Y^2 = \langle \varepsilon \nabla_x p_h, \nabla_x p_h \rangle_{L^2(Q)} = \sum_{\tau \in \mathcal{T}_H} \langle \varepsilon \nabla_x p_h, \nabla_x p_h \rangle_{L^2(\tau)} = \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2,$$

with the local error indicators

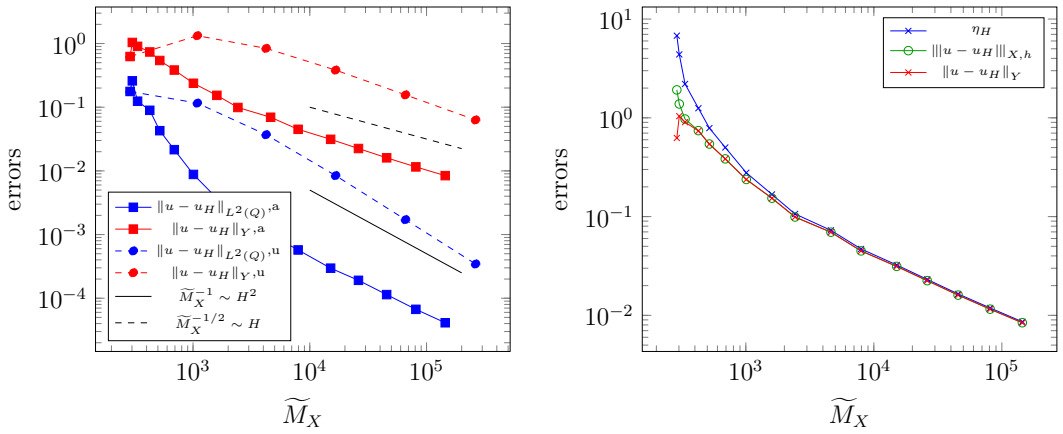
$$\eta_\tau^2 = \langle \varepsilon \nabla_x p_h, \nabla_x p_h \rangle_{L^2(\tau)} \quad \text{for } \tau \in \mathcal{T}_H.$$

As a marking strategy we use the Dörfler criterion [58] with parameter $\theta = 0.5$. The selected space-time simplicial elements are refined using newest vertex bisection. All computations were done in the software `Netgen/NGSolve` [152], where we used the sparse direct solver `Pardiso` [149] to solve the resulting linear systems.

In the first numerical example we consider the one dimensional spatial domain $\Omega = (0,1)$ and the time horizon $T = 1$, i.e., we have the space-time domain $Q = (0,1)^2$. As exact solution we choose the smooth function

$$u(x,t) := \left(1 - e^{-\frac{t}{\varepsilon}}\right) \left(\frac{e^{\frac{x-1}{\varepsilon}} - 1}{e^{-\frac{1}{\varepsilon}} - 1} + x - 1 \right), \tag{4.36}$$

and we compute $f = \partial_t u - \varepsilon \Delta_x u + \beta \cdot \nabla_x u$ accordingly for $\varepsilon = 10^{-2}$ and $\beta = 1$. Note that a similar function in a 2d-1d setting is considered in [33]. The smooth function (4.36) exhibits a spatial ($x = 1$) and a temporal ($t = 0$) boundary layer. The numerical results for both a uniform and an adaptive refinement strategy are shown in Fig. 4.12a. We observe a rate of $\mathcal{O}(H)$ for the error in the energy norm and $\mathcal{O}(H^2)$ for the L^2-error as expected. In Fig. 4.12b we present a comparison between the errors $\|u - u_H\|_Y$, $|||u - u_H|||_{X,h}$ and the error estimator $\eta_H = \|p_h\|_Y$. One can see that the error indicator is effective and that the error in the norm $|||\cdot|||_{X,h}$ is mainly driven by the spatial part of the norm. Finally, in Fig. 4.13 we present the related finite element mesh and the numerical solution u_H.



a) Errors $\|u - u_H\|_Y$ and $\|u - u_H\|_{L^2(Q)}$ for uniform and adaptive refinement strategies.

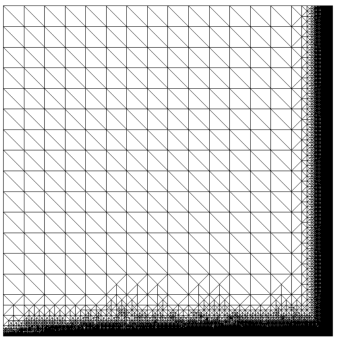
b) Comparison between estimator and true error for an adaptive refinement strategy.

Figure 4.12: Convergence results in the case of a smooth solution for a nonstationary convection-diffusion equation.

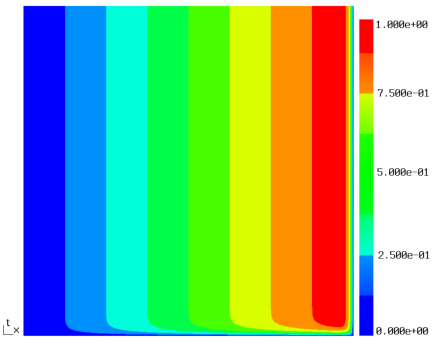
As a second numerical example we use the two-dimensional spatial domain $\Omega = (0,1)^2$ and the time horizon $T = 1$, i.e., we have the space-time domain $Q = (0,1)^3$. As initial state u_0 we consider similar as in [117] the function

$$u_0(x) := \psi(10\|x - c\|_2), \quad \psi(r) := \begin{cases} (1 - r^2)^2, & \text{for } r \leq 1, \\ 0, & \text{for } r > 1, \end{cases}$$

with $c = (0.5, 0.5)^T$. We compute numerical solutions to (4.19) for the velocity field $\beta = (0,1)^T$ and without a source term, i.e. $f \equiv 0$. Furthermore, we consider the diffusion coefficient to be $\varepsilon \in \{10^{-3}, 10^{-5}, 10^{-6}\}$. The results for a mesh with $32 \times 32 \times 32$ elements (35937 dofs) can be seen in Fig. 4.14. In the top row one can see the numerical solution u_H computed by solving (4.28) with the space-time finite element



a) Adaptive mesh on $L = 15$, 144563 dofs

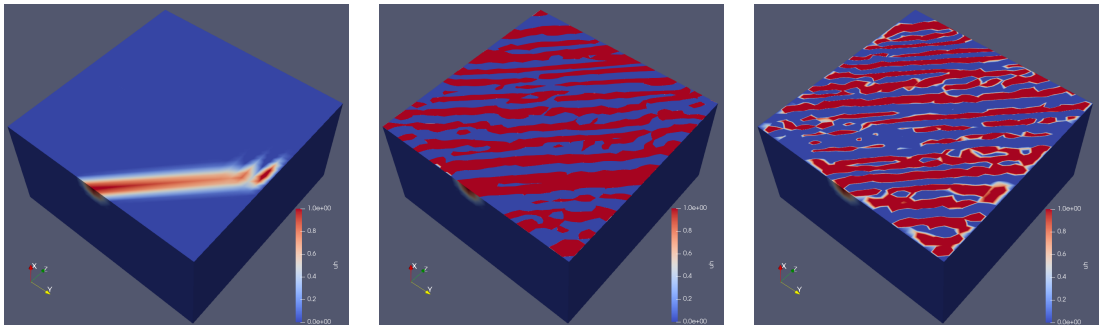


b) Solution u_H on adaptive mesh

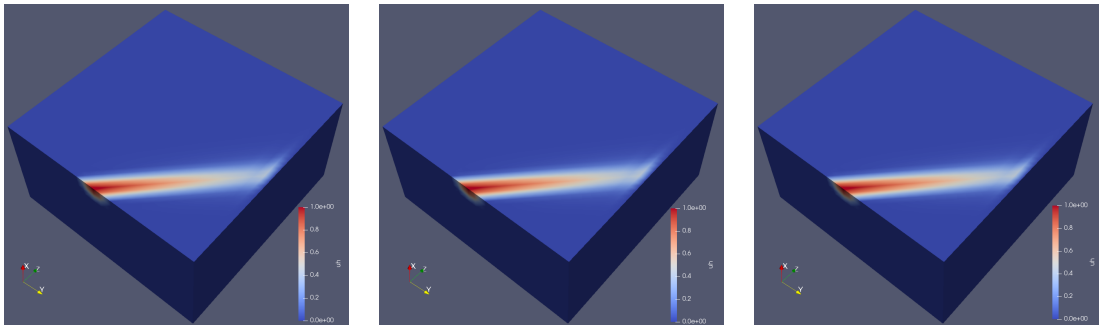
Figure 4.13: Simulation results for the adaptive refinement process.

method described in [157]. This leads to oscillations in the solution as the mesh size is not sufficiently small. In the bottom row one can see the solution using the developed least-squares formulation with piecewise linear trial and piecewise quadratic test functions. This formulation leads to stable results. We want to remark that the stabilization method considered in [117, Fig. 3] still has some minor oscillations, while in our case it seems that these oscillations are gone. In a further step we use the inbuilt error estimator to drive an adaptive refinement scheme for the parameters $\varepsilon = 10^{-3}$, $\beta = (0, 0.3)^T$ and $\varepsilon = 10^{-6}$, $\beta = (0, 1)^T$. In Fig. 4.15 the convergence rate of the error estimator in case of a uniform and an adaptive refinement strategy for both sets of parameters is depicted. In the case $\varepsilon = 10^{-3}$, $\beta = (0, 0.3)^T$, we observe a linear rate $\mathcal{O}(H)$ for both refinement strategies. In the case $\varepsilon = 10^{-6}$, $\beta = (0, 1)^T$ we observe a reduced rate of $\mathcal{O}(H^{0.4})$ in the uniform case. However, we can recover the full rate of $\mathcal{O}(H)$ in the adaptive case. The obtained adaptive meshes as well as the corresponding numerical solutions are depicted in Fig. 4.16 and 4.17. In every case we obtain a mesh which is fully unstructured in space and time.

As a third example we consider again the unit cube in the space-time domain, i.e., $Q = (0, 1)^3$. We choose $u_0 = 0$, $\varepsilon = 10^{-2}$ and the source term to be $f = 1$. The advection vector β is given in terms of a time dependent function with $\beta(t) = (\sin(2\pi t), \cos(2\pi t))^T$. Thus, the solution u has a boundary layer whose location depends on time. Note that a similar example is considered in [33, Ex. 4]. In Fig. 4.18 a comparison of the error estimator in case of an adaptive and uniform refinement strategy is depicted. We observe a convergence rate of $\mathcal{O}(H)$ in both cases. The generated grids in the adaptive case at different fixed times can be seen in Fig. 4.19. The circular movement of the boundary layer in time is visible.

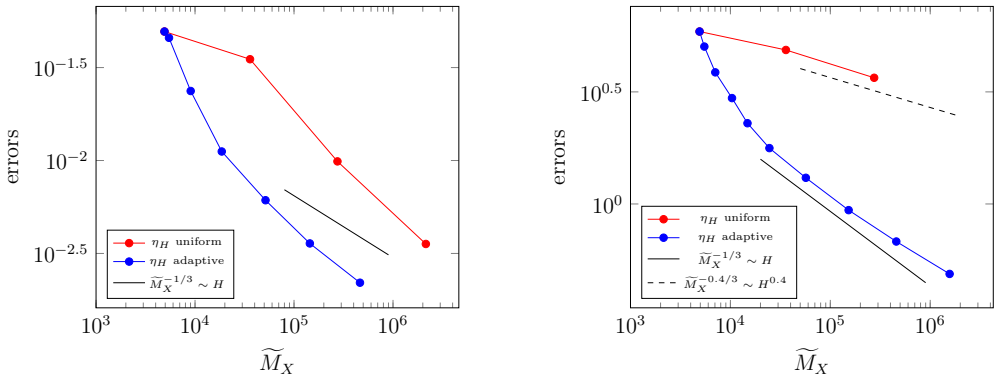


a) $\varepsilon = 10^{-3}$ b) $\varepsilon = 10^{-5}$ c) $\varepsilon = 10^{-6}$



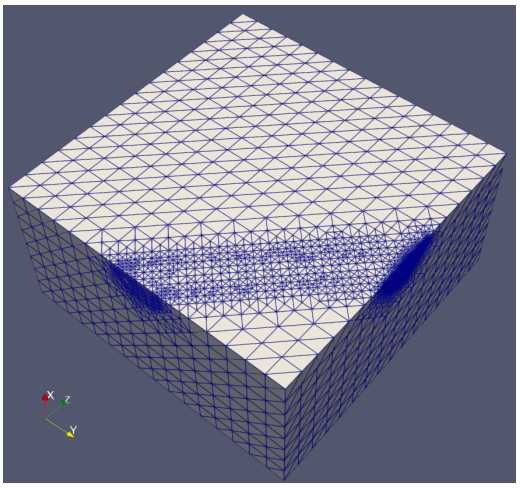
d) $\varepsilon = 10^{-3}$ e) $\varepsilon = 10^{-5}$ f) $\varepsilon = 10^{-6}$

Figure 4.14: Numerical results for $\beta = (0,1)^T$ on a mesh with $32 \times 32 \times 32$ elements. Top: no stabilization via direct formulation [157], bottom: stabilization via developed least-squares formulation.

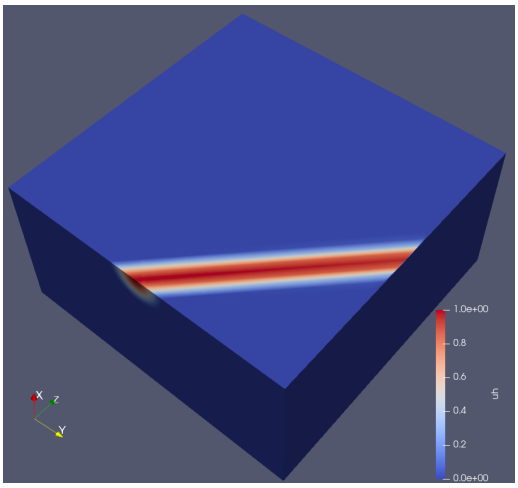


a) $\varepsilon = 10^{-3}$, $\beta = (0, 0.3)^T$ b) $\varepsilon = 10^{-6}$, $\beta = (0,1)^T$

Figure 4.15: Error estimator $\eta_H = \|p_h\|_Y$ in case of an adaptive and uniform refinement strategy for the second example.

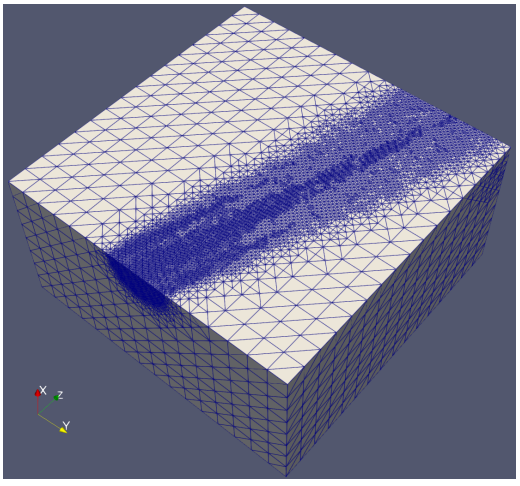


a) Adaptive mesh on $L = 7$, 152513 dofs

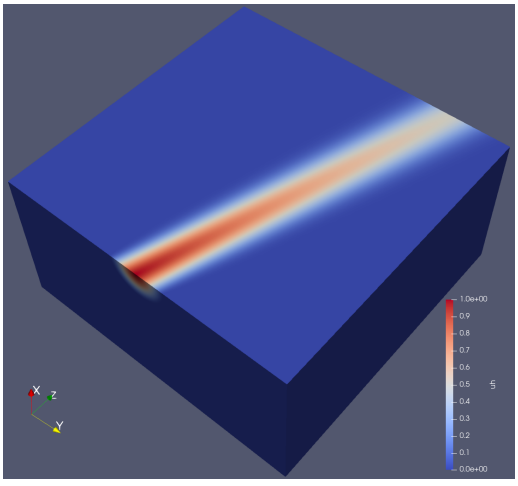


b) Solution u_H on the adaptive mesh

Figure 4.16: Obtained results for $\varepsilon = 10^{-6}$ and $\beta = (0, 1)^T$ after the adaptive refinement process.



a) Adaptive mesh on $L = 6$, 463696 dofs



b) Solution u_H on adaptive the mesh

Figure 4.17: Obtained results for $\varepsilon = 10^{-3}$ and $\beta = (0, 0.3)^T$ after the adaptive refinement process.

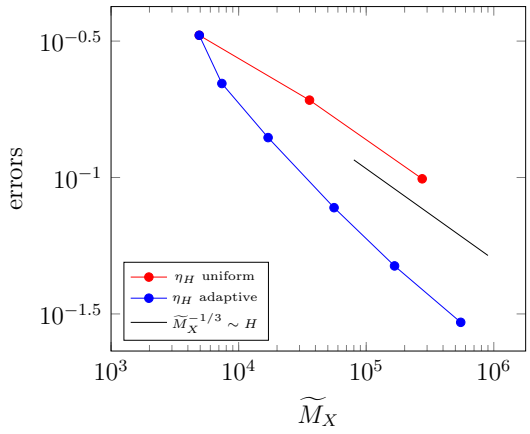


Figure 4.18: Error estimator $\eta_H = \|p_h\|_Y$ in case of an adaptive and uniform refinement strategy for the third example.

4.2.3 Numerical examples for the stationary case

Although the scope of Section 4.2 lies on the application of the minimal residual method to the nonstationary convection-diffusion equation within a space-time framework, we also want to demonstrate the approach in case of stationary convection-diffusion problems. The reason for this is that there are benchmark problems available, which allow for a comparison with reference values. We will quickly state the variational setting involving the spaces and the operators. Afterwards we show some numerical examples. For the sake of brevity we omit a detailed analysis of the operators, but we mention that this can be done similarly as in the instationary case.

In the stationary case we have the spaces

$$Y = H_0^1(\Omega), \quad X = \{u \in Y : \ \beta \cdot \nabla u \in Y^*\},$$

and consider the norms

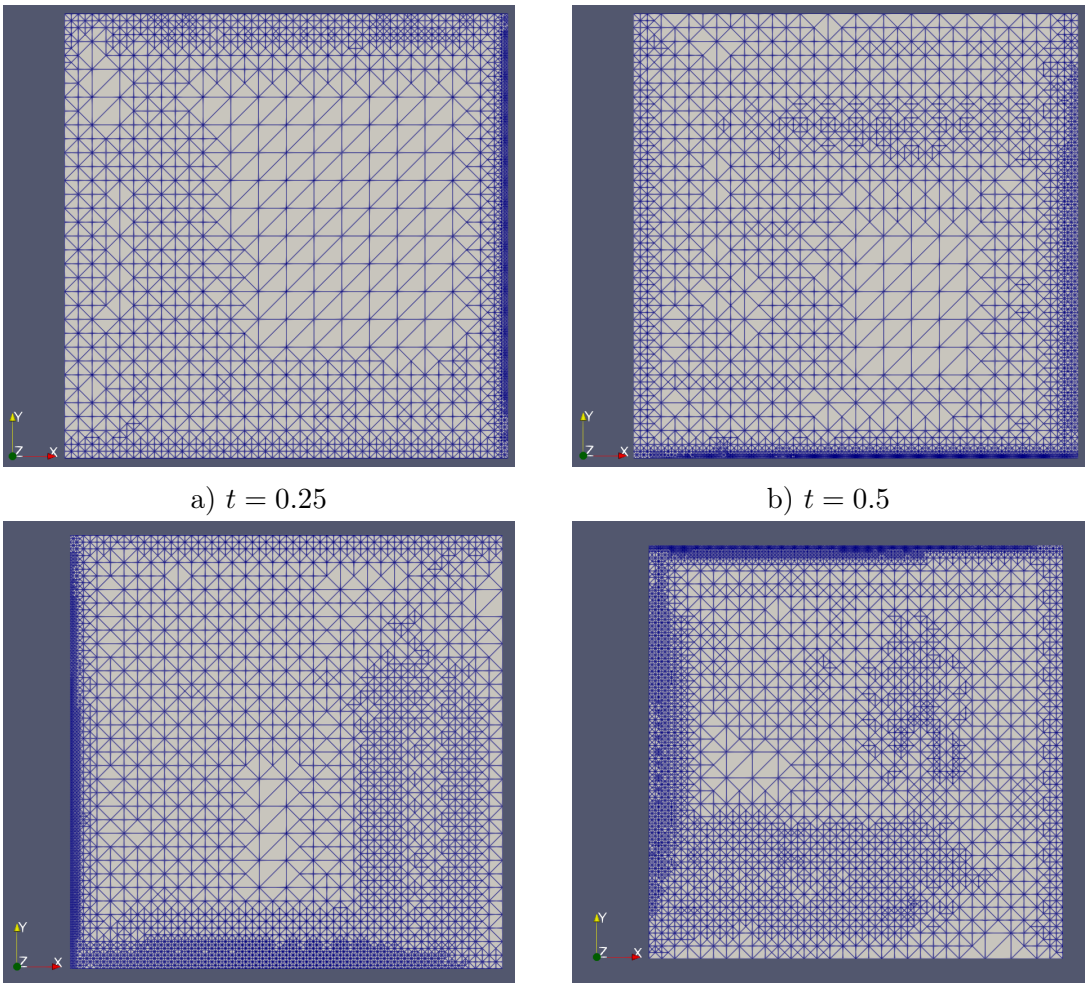
$$\|p\|_Y = \sqrt{\langle \varepsilon \nabla p, \nabla p \rangle_{L^2(\Omega)}}, \ \|u\|_X = \sqrt{\|u\|_Y^2 + \|\beta \cdot \nabla u\|_{Y^*}^2}, \ \|\beta \cdot \nabla u\|_{Y^*} = \|w_u\|_Y,$$

where the Riesz representative $w_u \in Y$ is the unique solution of the variational problem

$$\langle \varepsilon \nabla w_u, \nabla q \rangle_{L^2(\Omega)} = \langle \beta \cdot \nabla u, q \rangle_\Omega.$$

Since we assume bounded velocity fields β the norm $\|\cdot\|_X$ defines an equivalent norm in $H_0^1(\Omega)$ and the space X coincides with $Y = H_0^1(\Omega)$. This can be shown similar as in Lemma 4.8. The operators $A : Y \to Y^*$ and $B : X \to Y^*$ are given as

$$\langle Ap, q \rangle_\Omega := \langle \varepsilon \nabla p, \nabla q \rangle_{L^2(\Omega)}, \quad \langle Bu, q \rangle_\Omega := \langle \varepsilon \nabla u, \nabla q \rangle_{L^2(\Omega)} + \langle \beta \cdot \nabla u, q \rangle_\Omega$$



a) $t = 0.25$ b) $t = 0.5$

c) $t = 0.75$ d) $t = 1$

Figure 4.19: Generated mesh on refinement level $L = 5$ at different times t in case of the time dependent velocity field $\beta(t) = (\sin(2\pi t), \cos(2\pi t))^\top$.

for all $u \in X$, $p, q \in Y$. We remark that the operator $A : Y \to Y^*$ is self-adjoint bounded and elliptic with constants $c_1^A = c_2^A = 1$. For the operator $B : X \to Y^*$ we conclude

$$\|Bu\|_{Y^*} \leq \sqrt{2}\|u\|_X, \quad \|u\|_X \leq \sup_{0 \neq q \in Y} \frac{\langle Bu, q\rangle_\Omega}{\|q\|_Y},$$

i.e., boundedness and an inf-sup stability condition. This can be shown as demonstrated in Lemma 4.9. For the surjectivity we consider $0 \neq q \in Y = H_0^1(\Omega)$. Since we assume bounded velocity fields β we have $q \in X$, i.e., using $u_q = q \in X$ we obtain

$$\langle Bu_q, q\rangle_\Omega = \langle \varepsilon \nabla q, \nabla q\rangle_{L^2(\Omega)} + \langle \beta \cdot \nabla q, q\rangle_{L^2(\Omega)} = \|q\|_Y^2 > 0,$$

where we use $\langle \beta \cdot \nabla q, q\rangle_{L^2(\Omega)} = 0$ since we assume $\operatorname{div} \beta = 0$. Summing up all conditions of Section 3.1 are satisfied. The variational formulation of the mixed system is to find $(u, p) \in X \times Y$ such that

$$\begin{aligned} \langle \varepsilon \nabla p, \nabla q\rangle_{L^2(\Omega)} + \langle \varepsilon \nabla u, \nabla q\rangle_{L^2(\Omega)} + \langle \beta \cdot \nabla u, q\rangle_\Omega &= \langle f, q\rangle_\Omega, \\ \langle \varepsilon \nabla v, \nabla p\rangle_{L^2(\Omega)} + \langle \beta \cdot \nabla v, p\rangle_\Omega &= 0 \end{aligned}$$

is satisfied for all $(v, q) \in X \times Y$. For the discretization we choose $X_H = S_H^1(\mathcal{T}_H) \cap X$ and $Y_h = Y_H = S_H^2(\mathcal{T}_H)$, which are defined with respect to some admissible and locally quasi-uniform decomposition $\mathcal{T}_H$ of Ω into shape regular simplicial finite elements. Note that it holds $X_H \subset Y_h$. Similar as in the instationary case, cf. Lem. 4.11, one can prove a discrete inf-sup stability condition

$$\|u_H\|_{X,h} \leq \sup_{0 \neq q \in Y_h} \frac{\langle Bu_H, q_h\rangle_\Omega}{\|q_h\|_Y} \quad \text{for all } u_H \in X_H, \tag{4.37}$$

with the discrete norm

$$\|u_H\|_{X,h} := \sqrt{\|u\|_Y^2 + \|w_{uh}\|_Y^2} \leq \|u\|_X,$$

where $w_{uh} \in Y$ solves the variational problem

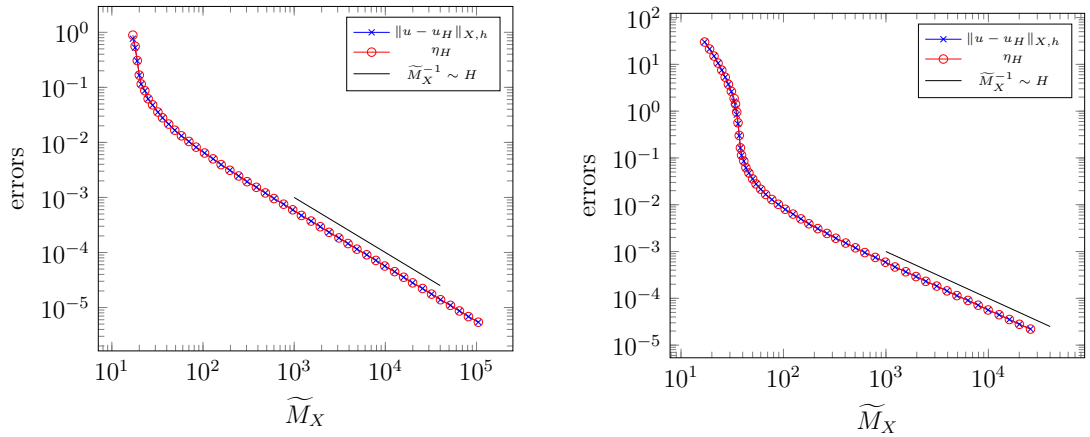
$$\langle \varepsilon \nabla w_{uh}, \nabla q_h\rangle_{L^2(\Omega)} = \langle \beta \cdot \nabla u, q_h\rangle_\Omega \quad \text{for all } q_h \in Y_h.$$

The stability condition (4.37) ensures that the discrete mixed system which reads to find $(u_H, p_h) \in X_H \times Y_h$ such that

$$\begin{aligned} \langle \varepsilon \nabla p_h, \nabla q_h\rangle_{L^2(\Omega)} + \langle \varepsilon \nabla u_H, \nabla q_h\rangle_{L^2(\Omega)} + \langle \beta \cdot \nabla u_H, q_h\rangle_\Omega &= \langle f, q_h\rangle_\Omega, \\ \langle \varepsilon \nabla v_H, \nabla p_h\rangle_{L^2(\Omega)} + \langle \beta \cdot \nabla v_H, p_h\rangle_\Omega &= 0 \end{aligned} \tag{4.38}$$

is satisfied for all $(v_H, q_h) \in X_H \times Y_h$ admits a unique solution. In our numerical experiments we solve (4.38) and use the global error estimator

$$\eta_H^2 = \|p_h\|_Y^2 = \langle \varepsilon \nabla p_h, \nabla p_h\rangle_{L^2(\Omega)} = \sum_{\tau \in \mathcal{T}_H} \langle \varepsilon \nabla p_h, \nabla p_h\rangle_{L^2(\tau)} = \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2,$$



a) $\varepsilon = 10^{-2}$ b) $\varepsilon = 10^{-5}$

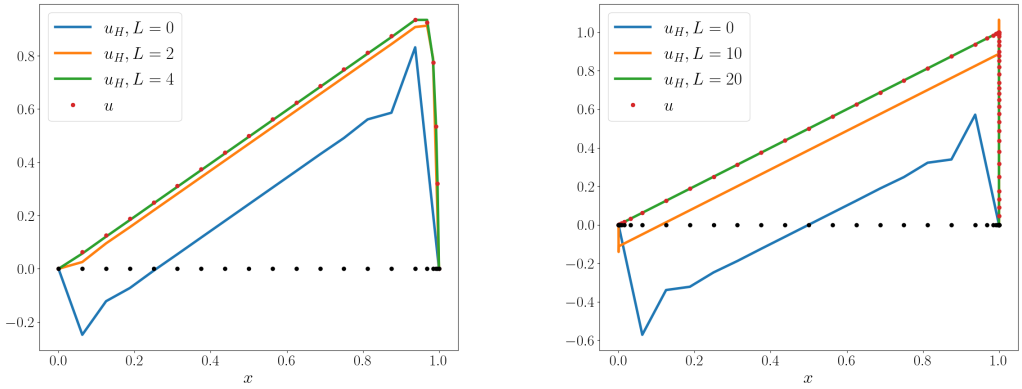
Figure 4.20: Convergence behaviour of the errors and the estimators for the problem (4.20) with $\varepsilon \in \{10^{-2}, 10^{-5}\}$.

with the local indicators

$$\eta_\tau^2 = \langle \varepsilon \nabla p_h, \nabla p_h \rangle_{L^2(\tau)} \quad \text{for } \tau \in \mathcal{T}_H$$

to drive an adaptive refinement scheme. As a marking strategy we use the Dörfler criterion [58] with parameter $\theta = 0.5$. All linear systems were solved using the sparse direct solver `Pardiso` [149].

As a first example we revisit the singularly perturbed boundary value problem (4.20), see also [141, Ex. 1.2], and apply the least-squares framework. The convergence behaviour of the error and the estimator in case of $\varepsilon \in \{10^{-2}, 10^{-5}\}$ are given in Fig. 4.20. As expected, we observe a linear rate. The numerical solutions obtained on different refinement levels are provided in Fig. 4.21. We see that the sequence of iterates from the adaptive refinement process converges to the physical true solution (4.21). However, the first few iterates obtain a constant shift from the true solution and some minor oscillations at $x = 0$ and at $x = 1$. Note that this behaviour is also observed in [12, 46]. Further, we see that the numerical solutions on the first few refinement levels have negative values even though the true solution is not negative. However, the inbuilt error estimator detects how many dofs need to be added in order to obtain a physical correct solution, which fulfills a discrete maximum principle (DMP). Finally, we want to mention that in case of $\varepsilon = 10^{-5}$ the numerical solution u_H on $L = 20$ was computed on a mesh with $\widetilde{M}_X = 60$ vertices (indicated in black) and 179 dofs for the corresponding saddle point system. This already lead to a satisfactory result. In comparison, the direct approach needed about 100000 dofs to give a satisfactory approximation to the solution.



a) $\varepsilon = 10^{-2}$ b) $\varepsilon = 10^{-5}$

Figure 4.21: Numerical solutions u_H and exact solution u (4.21) to the problem (4.20). The black dots indicate the grid points on $L = 4$ for $\varepsilon = 10^{-2}$ and on $L = 20$ for $\varepsilon = 10^{-5}$.

As a second example we consider the problem in [98], which deals with a non-constant convection field β. In particular, we consider $\Omega = (0,1)^2$ with the velocity field $\beta = (-y, x)^\top$, $\varepsilon = 10^{-5}$ and right-hand side $f = 0$. We prescribe homogeneous Dirichlet boundary conditions on the boundaries $\{1\} \times [0,1]$ and $[0,1] \times \{1\}$, i.e., on the right and top boundary. At the inlet boundary $[0,1] \times \{0\}$ we consider the inhomogeneous boundary condition given by

$$u(x,0) = \begin{cases} 1 - \frac{1}{4}\left(1 - \cos\left(\frac{1/3+\xi-x}{2\xi}\pi\right)\right)^2 & \text{for } x \in \left[\frac{1}{3}-\xi, \frac{1}{3}+\xi\right], \\ 1 & \text{for } x \in \left(\frac{1}{3}+\xi, \frac{2}{3}-\xi\right), \\ 1 - \frac{1}{4}\left(1 - \cos\left(\frac{x-2/3+\xi}{2\xi}\pi\right)\right)^2 & \text{for } x \in \left[\frac{2}{3}-\xi, \frac{2}{3}+\xi\right], \\ 0 & \text{else} \end{cases} \tag{4.39}$$

with $\xi = 10^{-3}$. On the remaining outlet boundary $\{0\} \times (0,1)$ we prescribe a homogeneous Neumann boundary condition. In order to study the satisfaction of the global DMP we evaluate as in [97] the quantity

$$\mathrm{osc}_{max}(u_H) = \max_{(x,y)\in\overline{\Omega}} u_H(x,y) - 1 - \min_{(x,y)\in\overline{\Omega}} u_H(x,y). \tag{4.40}$$

In order to assess the accuracy of the numerical solution three characteristic values of the solution at the outflow boundary are provided in [98]. The reference values read:

- width of the lower layer: 0.01439869,

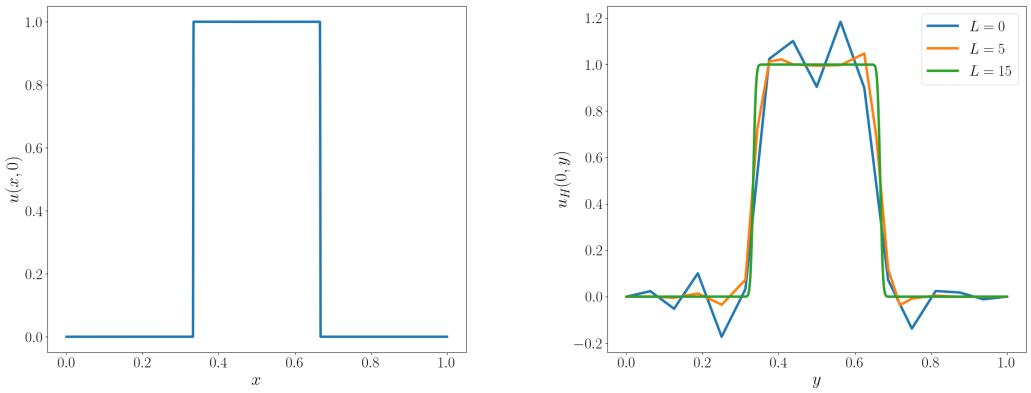


Figure 4.22: Left: Inhomogeneous boundary condition (4.39) prescribed on inlet boundary. Right: Numerical solution u_H at the outlet boundary for different refinement levels.

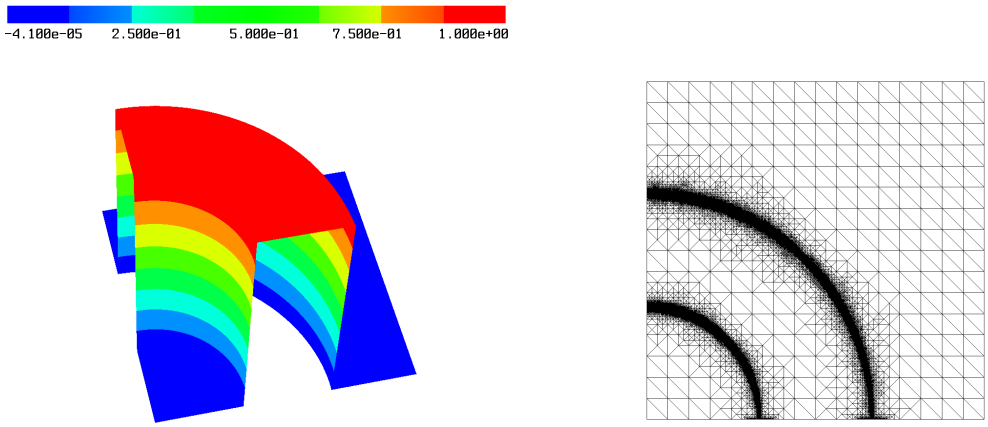


Figure 4.23: Left: Numerical solution u_H on $L = 15$. Right: Adaptive mesh on $L = 15$ with 148968 dofs.

- width of the upper layer: 0.01439637,
- outflow profile width: 0.3482541.

In Fig. 4.22 we provide a plot of the inhomogeneous boundary condition (4.39) on the inlet boundary as well as the numerical solution u_H at the outlet boundary for different refinement levels. In Fig. 4.23 the numerical solution u_H as well as the adaptive mesh generated on refinement level $L = 15$ are depicted. The adaptive mesh was generated from a structured initial mesh with 16×16 elements. Further, in Tab. 4.2 and Fig. 4.24 we provide a comparison of the computed characteristic values to the reference values and the satisfaction of the global DMP. We see that

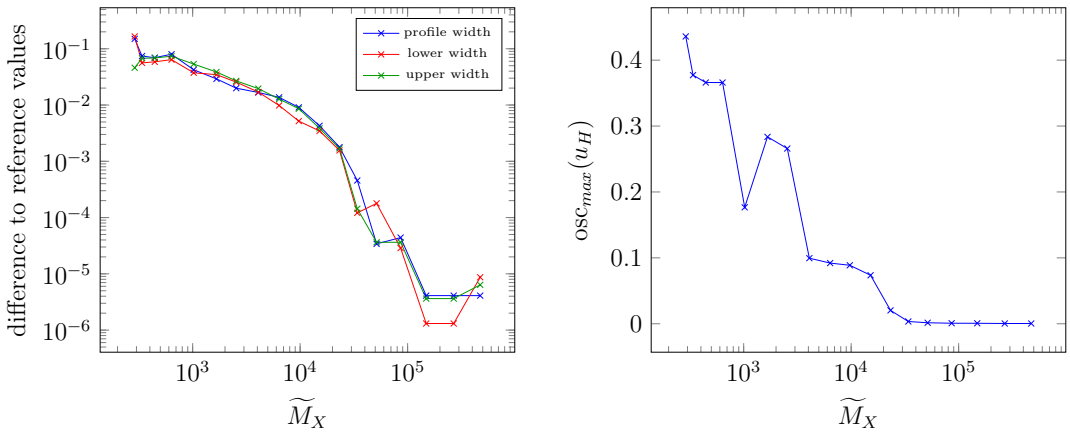


Figure 4.24: Left: Comparison of the computed characteristic values to the reference values. Right: Satisfaction of the global DMP on different refinement levels.

L	$\widetilde{M}_X$	lower layer	upper layer	profile	$\mathrm{osc}_{max}(u_H)$
0	289	0.17989000	0.06040999	0.49811000	0.43594821
5	1666	0.04909000	0.05312000	0.37743000	0.28357896
10	15160	0.01785999	0.01824999	0.35253999	0.07359097
13	51480	0.01422000	0.01436000	0.34822000	1.523e-03
15	148968	0.01440000	0.01440000	0.34825000	7.373e-04
17	473779	0.01439000	0.01439000	0.34825000	4.033e-04

Table 4.2: Computed characteristic values and evaluation of (4.40).

for earlier refinement levels over- and undershoots are visible, but they are almost vanishing on higher refinement levels. This can be also seen from the quantity (4.40) in Table 4.2 and Fig. 4.24, which shows a good satisfaction of the global DMP for higher refinements. In addition to that Fig. 4.24 shows that the difference of the computed characteristic quantities to the reference values is in the order of floating point precision for higher refinement levels, i.e., we have a good agreement of the computed reference values with the characteristic values.

As a third example we consider the Hemker problem [91], which is a standard benchmark problem for steady state convection-diffusion problems. The domain is given by $\Omega = (-3,9)\times(-3,3)\setminus\{(x,y) : x^2+y^2 \leq 1\}$. The velocity is given by $\beta = (1,0)^\top$, and the right-hand side f is set equal to zero. Further we have the boundary conditions

$$u(x,y) = \begin{cases} 0 & x = -3, \\ 1 & x^2 + y^2 = 1, \\ \varepsilon \nabla u \cdot \mathbf{n} = 0 & x = 9 \ \vee \ y = -3 \ \vee \ y = 3 \end{cases} .$$

As in [10, 97, 98] we consider the diffusion coefficient to be $\varepsilon = 10^{-4}$. In order to assess the accuracy of the numerical solution a value for the width of the interior layer at $x = 4$ was provided in [10]. This width is defined to be the length of the interval, where $u(4,y) \in [0.1, 0.9]$. In [10] the reference value 0.0723 is provided for the upper layer, i.e., where $y \geq 0$. Furthermore, we evaluate the quantity (4.40) to measure the satisfaction of the global DMP. The values of the reference solution provided in [10] are contained in the interval $[0,1]$. In Fig. 4.25 the initial mesh and the adaptive mesh otained on $L = 17$ are depicted. We see stronger refinements at the boundary layer around the circle, which at the top and bottom of the circle passes into an interior layer that spread into the direction of the convection. Moreover, we observe some refinements at the left boundary. This may be explained in terms of the over- and undershoots that occur for lower refinements. In Fig. 4.27 we provide a plot of the numerical solution u_H for $y = 1$ and $x = 4$, i.e., we consider the cut lines $u_H(x,1)$ and $u_H(4,y)$. For ealier refinement levels we see an oscillatory behaviour of the cut line $u_H(x,1)$ and some over- and undershoots in both cut lines. However, for higher refinements these oscillations are reduced, and we see cut lines which are in good accordance with the lines of the physically correct reference solution, see [10, Fig. 6]. In Tab. 4.3 and Fig. 4.28 we provide a comparison of the computed characteristic value with the reference value as well as a plot of the satisfaction of the DMP. We see that we converge to the reference characteristic value, while the quantity (4.40) gets reduced. For example on level $L = 18$ the difference of the computed upper width to the reference upper width is about $3.5 \cdot 10^{-4}$ and the quantitiy (4.40) is approximately $6.8 \cdot 10^{-4}$. This means we have an accurate numerical solution which shows a good satisfaction of the DMP.

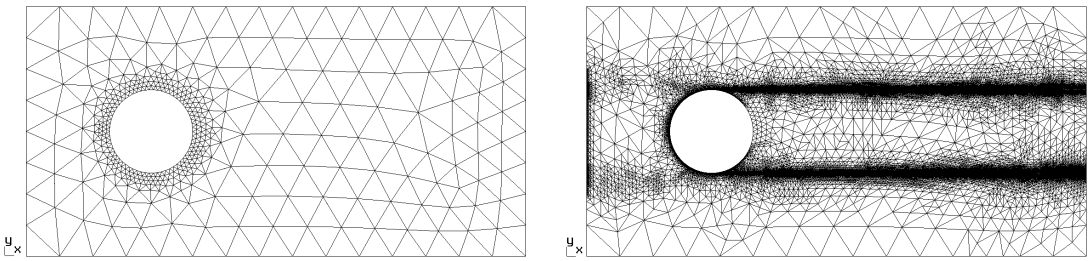


Figure 4.25: Left: Initial mesh with 438 dofs. Right: Adaptive mesh on $L = 17$ with 676072 dofs.

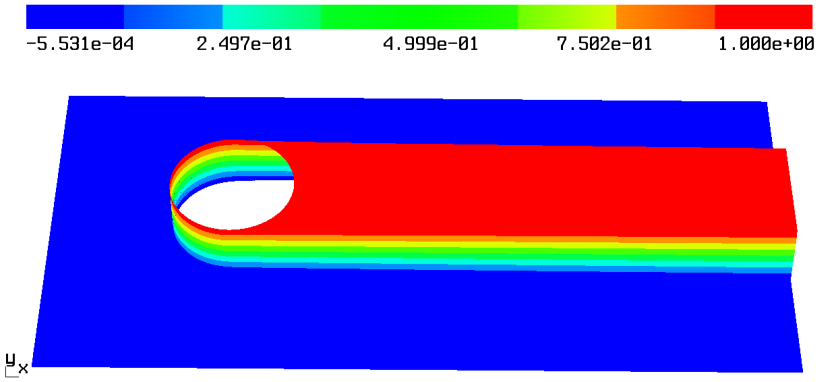


Figure 4.26: Numerical solution u_H on $L = 17$.

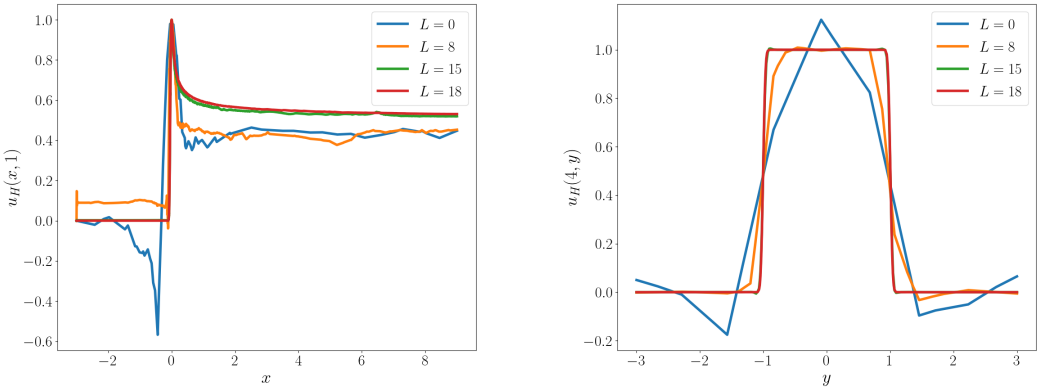


Figure 4.27: Cut lines of the numerical solution on different refinement levels. Left: $u_H(x, 1)$. Right: $u_H(4, y)$.

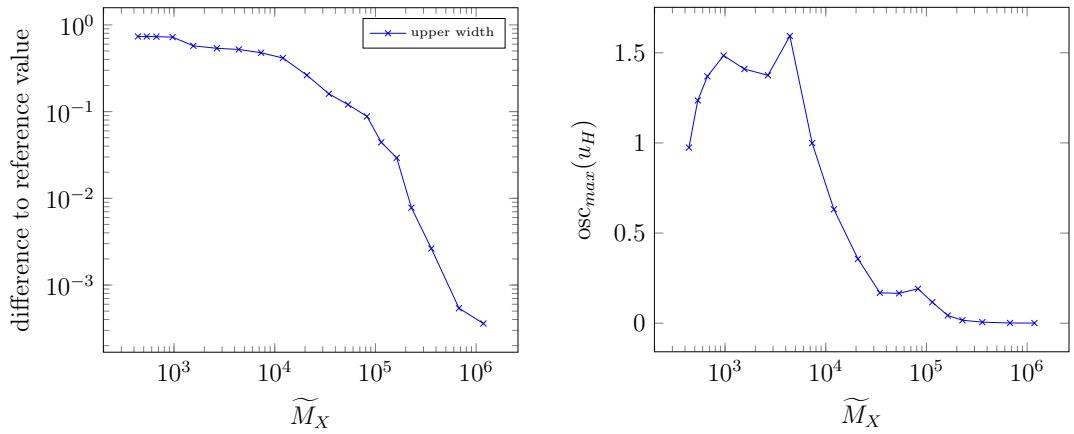


Figure 4.28: Left: Comparison of the computed characteristic values to the reference values. Right: Satisfaction of the global DMP on different refinement levels.

L	$\widetilde{M}_X$	upper layer	$\mathrm{osc}_{max}(u_H)$
0	438	0.8087	0.9742
8	12048	0.4873	0.6315
12	82453	0.1603	0.1909
15	227728	0.0801	0.0165
17	676072	0.0728	1.187e-03
18	1181853	0.0727	6.823e-04

Table 4.3: Computed characteristic value and evaluation of (4.40) for the Hemker problem.

4.3 A semilinear model problem

In this section we will apply the minimal residual framework presented in Section 3.2 to a semilinear equation. In particular, we consider the problem

$$\partial_t u(x,t) - \Delta_x u(x,t) + u^3(x,t) = f(x,t) \quad \text{for } (x,t) \in Q := \Omega \times (0,T), \tag{4.41a}$$
$$u(x,t) = 0 \quad \text{for } (x,t) \in \Sigma := \Gamma \times (0,T), \tag{4.41b}$$
$$u(x,0) = 0 \quad \text{for } x \in \Omega, \tag{4.41c}$$

where $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ is a bounded Lipschitz domain with boundary $\Gamma = \partial\Omega$, and $T > 0$ is a given time horizon. This model problem is motivated from the Schlögl model [39, 151], which models electrical currents on the human heart. It can be also related to the FitzHugh-Nagumo equations [155], which describe reaction-diffusion systems in biology.

4.3.1 Minimal residual formulation of the nonlinear problem

In view of the abstract setting we have the same spaces as for the heat equation, i.e.,

$$X := L^2(0,T;H_0^1(\Omega)) \cap H_{0,}^1(0,T;H^{-1}(\Omega)),\ Y := L^2(0,T;H_0^1(\Omega)),\ H := L^2(Q),$$

with the corresponding norms

$$\|p\|_Y := \|\nabla_x p\|_{L^2(Q)}, \quad \|u\|_X := \sqrt{\|\partial_t u\|_{Y^*}^2 + \|\nabla_x u\|_{L^2(Q)}^2}.$$

The operators $A : Y \to Y^*$ and $B : X \to Y^*$ are defined in the variational sense satisfying

$$\begin{aligned} \langle Ap, q\rangle_Q &:= \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)}, \\ \langle B(u), q\rangle_Q &:= \langle \partial_t u, q\rangle_Q + \langle \nabla_x u, \nabla_x q\rangle_{L^2(Q)} + \big\langle u^3, q\big\rangle_{L^2(Q)}, \end{aligned} \tag{4.42}$$

for all $p, q \in Y$ and $u \in X$. Note that the expression

$$\big\langle u^3, q\big\rangle_{L^2(Q)} = \int_0^T \big\langle u^3(t), q(t)\big\rangle_{L^2(\Omega)} \mathrm{d}t$$

is well-defined due to the embeddings $H_0^1(\Omega) \hookrightarrow C(\overline{\Omega})$ for $\Omega \subset \mathbb{R}$, $H_0^1(\Omega) \hookrightarrow L^q(\Omega)$ with $1 \le q < \infty$ for $\Omega \subset \mathbb{R}^2$ and $H_0^1(\Omega) \hookrightarrow L^6(\Omega)$ for $\Omega \subset \mathbb{R}^3$, see [168, Satz 7.1]. The first and second directional derivative of the operator B are then given by $B'(u) : X \to Y^*$, $B''(u) : X \times X \to Y^*$ satisfying

$$\begin{aligned} \langle B'(u)\varphi, q\rangle_Q &= \langle \partial_t \varphi, q\rangle_Q + \langle \nabla_x \varphi, \nabla_x q\rangle_{L^2(Q)} + 3\big\langle u^2\varphi, q\big\rangle_{L^2(Q)}, \\ \langle B''(u)(\psi,\varphi), q\rangle_Q &= 6\langle u\psi\varphi, q\rangle_{L^2(Q)}, \end{aligned}$$

for all $u, \varphi, \psi \in X$, $q \in Y$. The abstract mixed Galerkin variational formulation of the optimality system (3.47) in this case reads to find $(u, p) \in X \times Y$ such that

$$\begin{aligned} \langle \nabla_x p, \nabla_x q \rangle_{L^2(Q)} + \langle \partial_t u, q \rangle_Q + \langle \nabla_x u, \nabla_x q \rangle + \big\langle u^3, q \big\rangle_{L^2(Q)} &= \langle f, q \rangle_Q, \\ \langle \partial_t v, p \rangle_{L^2(Q)} + \langle \nabla_x v, \nabla_x p \rangle_{L^2(Q)} + 3\big\langle u^2 v, p \big\rangle_{L^2(Q)} &= 0, \end{aligned} \tag{4.43}$$

holds for all $(v, q) \in X \times Y$. The application of Newton's method to (4.43) gives the following algorithm.

ALGORITHM 4.13. *Choose an initial guess* $(p^0, u^0) \in Y \times X$.

For $k = 0, 1, 2, ...,$ *until convergence do*

(i) Find $(w_p^k, w_u^k) \in Y \times X$ *such that*

$$\begin{aligned} \big\langle \nabla_x w_p^k, \nabla_x q \big\rangle_{L^2(Q)} + \big\langle \partial_t w_u^k, q \big\rangle_Q \\ + \big\langle \nabla_x w_u^k, \nabla_x q \big\rangle_{L^2(Q)} + 3\big\langle (u^k)^2 w_u^k, q \big\rangle_{L^2(Q)} &= -\big\langle G_1(p^k, u^k), q \big\rangle_Q \\ \big\langle \partial_t v, w_p^k \big\rangle_Q + \big\langle \nabla_x v, \nabla_x w_p^k \big\rangle_{L^2(Q)} \\ + 3\big\langle (u^k)^2 v, w_p^k \big\rangle_{L^2(Q)} + 6\big\langle u^k w_u^k v, p^k \big\rangle_{L^2(Q)} &= -\big\langle G_2(p^k, u^k), v \big\rangle_Q \end{aligned}$$

is satisfied for all $(q, v) \in Y \times X$, *where* $G_1 : Y \times X \to Y^*$ *and* $G_2 : Y \times X \to X^*$ *are given as*

$$\begin{aligned} \big\langle G_1(p^k, u^k), q \big\rangle_Q &= \big\langle \nabla_x p^k, \nabla_x q \big\rangle_{L^2(Q)} + \big\langle \partial_t u^k, q \big\rangle_Q \\ &\quad + \big\langle \nabla_x u^k, \nabla_x q \big\rangle + \big\langle (u^k)^3, q \big\rangle - \langle f, q \rangle_Q \end{aligned}$$

$$\big\langle G_2(p^k, u^k), v \big\rangle_Q = \big\langle \partial_t v, p^k \big\rangle_{L^2(Q)} + \big\langle \nabla_x v, \nabla_x p^k \big\rangle_{L^2(Q)} + 3\big\langle (u^k)^2 v, p^k \big\rangle_{L^2(Q)},$$

for all $v \in X$, $q \in Y$.

(ii) Set $(p^{k+1}, u^{k+1}) = (p^k, u^k) + (w_p^k, w_u^k)$.

The Gauß-Newton algorithm for this particular problem is given via the following algorithm.

ALGORITHM 4.14. *Choose an inital guess* $u^0 \in X$.

For $k = 0, 1, 2, ...,$ *until convergence do*

(i) Find $p^k \in Y$ *such that*

$$\big\langle \nabla_x p^k, \nabla_x q \big\rangle_{L^2(Q)} = \big\langle \partial_t u^k, q \big\rangle_Q + \big\langle \nabla_x u^k, \nabla_x q \big\rangle_{L^2(Q)} + \big\langle (u^k)^3, q \big\rangle_{L^2(Q)} - \langle f, q \rangle_Q$$

holds for all $q \in Y$.

(ii) Solve the mixed system to find $(w_p^k, w_u^k) \in Y \times X$ *such that*

$$\begin{aligned}
\big\langle \nabla_x w_p^k, \nabla_x q \big\rangle_{L^2(Q)} + \big\langle \partial_t w_u^k, q \big\rangle_Q & \\
+ \big\langle \nabla_x w_u^k, \nabla_x q \big\rangle_{L^2(Q)} + 3\big\langle (u^k)^2 w_u^k, q \big\rangle_{L^2(Q)} &= 0, \\
\big\langle \partial_t v, w_p^k \big\rangle_Q + \big\langle \nabla_x v, \nabla_x w_p^k \big\rangle_{L^2(Q)} & \\
+ 3\big\langle (u^k)^2 v, w_p^k \big\rangle_{L^2(Q)} &= \big\langle \partial_t v, p^k \big\rangle_Q + \big\langle \nabla_x v, \nabla_x p^k \big\rangle_{L^2(Q)} \\
+ 3\big\langle (u^k)^2 v, p^k \big\rangle_{L^2(Q)} &
\end{aligned}$$

is satisfied for all $(q, v) \in Y \times X$.

(iii) Set $u^{k+1} = u^k + w_u^k$.

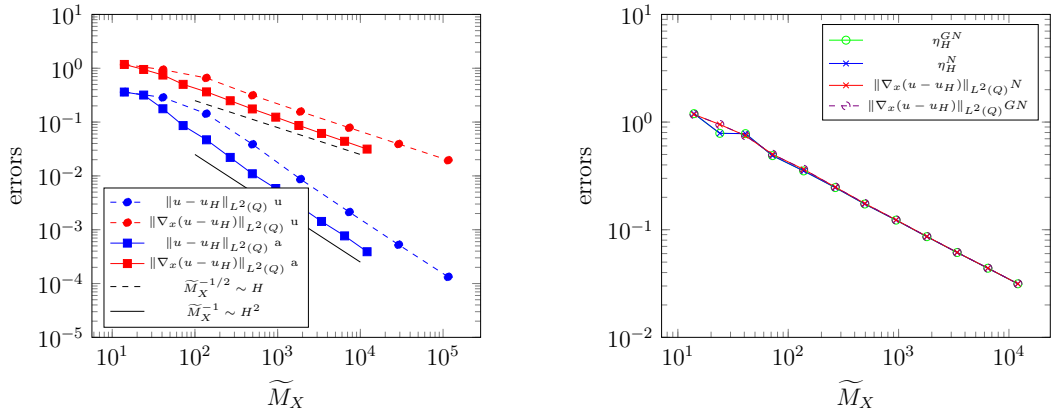
For the discretization we consider a trial space $X_H = \operatorname{span}\{\varphi_i\}_{i=1}^{M_X} \subset X$ and a test space $Y_h = \operatorname{span}\{\psi_i\}_{i=1}^{M_Y} \subset Y$. The discrete version of the Algorithms 4.13 and 4.14 are given by the Algorithms 3.30, 3.31 together with the matrices and vectors

$$\begin{aligned}
A_h[i,j] &= \langle \nabla_x \psi_j, \nabla_x \psi_i \rangle_{L^2(Q)}, \quad i,j = 1, ..., M_Y, \\
\underline{B}(\underline{u})[i] &= \langle \partial_t u_H, \psi_i \rangle_Q + \langle \nabla_x u_H, \nabla_x \psi_i \rangle + \big\langle u_H^3, \psi_i \big\rangle_{L^2(Q)}, \quad i = 1, ..., M_Y, \\
B_h'(\underline{u})[i,j] &= \langle \partial_t \varphi_j, \psi_i \rangle_Q + \langle \nabla_x \varphi_j, \nabla_x \psi_i \rangle_{L^2(Q)} + 3\big\langle u_H^2 \varphi_j, \psi_i \big\rangle_{L^2(Q)}, \\
& \qquad i = 1, ..., M_Y, \ j = 1, ..., M_X \\
B_h''(\underline{u}, \underline{p})[i,j] &= 6\langle u_H p_h \varphi_j, \varphi_i \rangle_{L^2(Q)}, \quad i,j = 1, ..., M_X \\
\underline{f}[i] &= \langle f, \psi_i \rangle_Q \quad i = 1, ..., M_Y,
\end{aligned}$$

where we use the identifications $u_H \in X_H \leftrightarrow \underline{u} \in \mathbb{R}^{M_X}$, $p_h \in Y_h \leftrightarrow \underline{p} \in \mathbb{R}^{M_Y}$.

4.3.2 Numerical example

In this section we apply Algorithm 3.30 and Algorithm 3.31 in their damped version with a backtracking line search strategy for the solution of the semilinear heat equation (4.41). For the discretization we choose the trial space $X_H = S_H^1(\mathcal{T}_H) \cap X$ of piecewise linear and continuous basis functions and the test space $Y_h = Y_H = S_H^2(\mathcal{T}_H) \cap Y$ of piecewise quadratic basis functions. In both algorithms we compute quantities η_H^N and η_H^{GN}, which will be used as an error indicator to drive an adaptive refinement scheme. They allow for a represenation similar as in (4.16). As a marking strategy we will use the Dörfler criterion [58] with parameter $\theta = 0.5$. The marked elements are then refined using newest vertex bisection. For the implementation we used the finite element software `Netgen/NGSolve` [152]. All systems of algebraic equations were solved using the sparse direct solver `Pardiso` [148, 149, 150].



a) Errors in Newton's method for adaptive and uniform refinement.

b) Comparison of errors and estimators in Newton's and Gauß-Newton's method.

Figure 4.29: Convergence behavior for the approximation of the smooth function (4.17).

In the numerical example we consider the space-time domain $Q = (0,3) \times (0,6) \subset \mathbb{R}^2$. As exact solution $u : Q \to \mathbb{R}$ we consider the smooth function (4.17), and we compute $f = \partial_t u - \Delta_x u$ accordingly. On refinement level $L = 0$ we choose the initial guesses $(\underline{p}^{L,0}, \underline{u}^{L,0}) = (\underline{0}, \underline{0})$, $\underline{u}^{L,0} = \underline{0}$, for Newtons's, Gauß-Newton's method, respectively. For $L > 0$ we prolongate the solution $\underline{u}^{L-1}$ from the mesh level $L-1$ to the current mesh level L and take this as an initial guess in both methods, i.e., $\underline{u}^{L,0} = P_h^{L,L-1}\underline{u}^{L-1}$. Additionally, in Newton's method we choose $\underline{p}^{L,0} = 0$ for $L > 0$. The numerical results for both a uniform and an adaptive refinement scheme in case of Newton's method can be seen in Fig. 4.29a. We observe a rate of $\mathcal{O}(H)$ for the error $\|\nabla_x(u-u_H)\|_{L^2(Q)}$ and a rate $\mathcal{O}(H^2)$ for the $L^2(Q)$ error. This is expected as we consider the smooth function (4.17) for the true solution. In Fig. 4.29b we present a comparison between the results of Newton's and Gauß-Newton's method. On the one hand it shows that we get the same energy errors in both methods. On the other hand one can also see that in both cases the error estimators η_H^N, η_H^{GN} are effective. In Tab. 4.4 and 4.5 we present the number of iterations (iter), the step size τ in the final step and the final error (err) for Newton's and Gauß-Newton's method, respectively. As a stopping criterion we chose a maximal error of 10^{-10}. We can see that Newton's method takes 3 to 4 iterations on each refinement level until convergence. The Gauß-Newton method needs slightly more iterations namely 3 to 8. This is expected since in the Gauß-Newton method we neglect the information of the second order derivative of the operator B for computing a descent direction. However, we can observe from the number of nodes $\widetilde{M}_X = \dim(S_h^1(Q))$ that both algorithms lead to the same number of marked elements which are refined in every step.

L	$\widetilde{M}_X$	iter	err	τ
0	14	4	8.872e-14	1.000
1	24	4	3.295e-16	1.000
2	41	4	6.023e-11	1.000
3	72	4	1.182e-12	1.000
4	138	4	3.408e-16	1.000
5	267	4	3.616e-16	0.500
6	494	3	3.839e-11	1.000
7	949	3	5.593e-12	1.000
8	1804	3	1.702e-13	1.000
9	3395	3	4.266e-14	1.000
10	6448	3	1.582e-15	1.000
11	12057	3	1.405e-15	1.000

Table 4.4: Number of iterations and final errors of Newton's method during adaptive refinement.

L	$\widetilde{M}_X$	iter	err	τ
0	14	7	3.434e-11	1.000
1	24	8	6.246e-12	1.000
2	41	8	8.577e-12	1.000
3	72	6	3.275e-11	1.000
4	138	5	2.032e-11	1.000
5	267	5	1.547e-12	1.000
6	494	4	3.392e-11	1.000
7	949	4	4.121e-12	1.000
8	1804	4	2.272e-13	1.000
9	3395	4	2.361e-14	1.000
10	6448	3	1.042e-11	1.000
11	12057	3	2.862e-12	1.000

Table 4.5: Number of iterations and final errors of Gauß-Newton's method during adaptive refinement.

5 Application to the simulation of electric machines

This chapter is dedicated to the application of the minimal residual method to the simulation of electric machines. Electric machines are part of our everyday life, ranging from applications in industry, public services or in households. They are also used to power a variety of equipment including wind blowers, water pumps or compressors, see e.g., [145]. Further, it is mentionend in [115, 145] that electric machines consume about 2/3 of the industrial power in each nation and about 46% of the total electricity worldwide, which result in about 6040 Mega-tonnes of CO_2 emission. Increasing the efficiency of electric machines is therefore of great interest, not only from an economic point of view, since the most efficient machines will dominate the market, but also from the point of view of sustainability and environmental protection. The latter two concerns have become more and more important over the last years due to the fight against climate change.

The design of efficient electric machines relies on powerful simulation tools. The creation of such tools is a challenging task as electric machines are multiphysical objects, meaning that underlying models include electromagnetic, thermodynamic as well as structural mechanics partial differential equations, which are accomplished by nonlinear and possibly hysteretic material models. In this thesis we restrict ourselves to magnetic field computations like in [76, 78, 89, 90, 136]. Magnetic field computations are fundamental in the design process of an electric machine as performance criteria like torque [144] or losses can be derived from the magnetic flux density. Therefore, an application of a numerical method in combination with an adaptive mesh refinement scheme to determine the magnetic flux density accurately is of high interest.

5.1 A brief introduction into electric machines and the electromagnetic model

Electric machines can be distinguished between motors and generators, see [19, 76, 104]. An electric motor converts electrical into mechancial energy, while a generator does the opposite. Further, one can classify electric motors in direct current motors (DC-motors) and in alternating current motors (AC-motors). Two important

representatives of AC-motors are the induction motors or asynchronous motors and the synchronous motors. The latter can be divided into synchronous reluctance machines and permanent synchronous motors. In general, electric motors consist of a fixed part called stator and a moving part called rotor, which are separated via an air gap, cf. [76]. For more details on electric machines we refer the interested reader to the book of Binder [19].

Starting point for the derivation of the physical model are Maxwell's equations, see e.g., [95, 96, 102, 103, 109, 122]. The complete set of equations in its differential form reads

$$\operatorname{curl} \mathbf{H} = \mathbf{J} + \frac{\partial}{\partial t}\mathbf{D}, \tag{5.1a}$$

$$\operatorname{curl} \mathbf{E} = -\frac{\partial}{\partial t}\mathbf{B}, \tag{5.1b}$$

$$\operatorname{div} \mathbf{B} = 0, \tag{5.1c}$$

$$\operatorname{div} \mathbf{D} = \varrho. \tag{5.1d}$$

The electromagnetic quantities are connected via the constitutive equations, see e.g., [40, 100, 102, 103]

$$\mathbf{J} = \mathbf{J}_i + \sigma(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{5.2}$$

$$\mathbf{D} = \varepsilon\mathbf{E} + \mathbf{P}, \tag{5.3}$$

$$\mathbf{B} = \mu\left(\mathbf{H} + \mathbf{M}\right). \tag{5.4}$$

Note that the physical quantities are vector valued functions mapping from $\mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}^3$. Further, we want to mention that the differential operators like curl, div, ∇ are acting on the spatial coordinates. A detailed description of the underlying physical quantities in Maxwell's equations and the related material laws together with their units are given in Table 5.1. The units are given in terms of SI units like metre m, seconds s, Ampere A and derived units like Volt V, Tesla T, Ohm $\Omega = V/A$ and its reciprocal Siemens $S = 1/\Omega$. It has to be stated, see e.g., [102, 174] that the fundamental equations (5.1) are based on experiments and concluding empirical laws stated by Ampère, Faraday and Gauß. In particular, Ampère's law (5.1a) states how electric current generates a magnetic field intensity. The law of Faraday (5.1b) describes how a time varying magnetic flux density induces a voltage. On the one hand it is based on the observation that a uniform and constant magnetic flux density induces a voltage in an electrically conductive slab which moves inside the magnetic field $\mathbf{B}$, cf. [102]. On the other hand it represents the fact that a time varying magnetic flux density induces a voltage in an open conductive loop, cf. [102]. The relation (5.1c) states that there exists no magnetic monopols, i.e., no magnetic charges exist, and the magnetic field is solenoidal, i.e., the field lines are closed, see [102]. The law of Gauß (5.1d) postulates that the amount of electric flux density

Quantity	Unit	Description
$\mathbf{E}$	$\frac{V}{m}$	Electric field intensity
$\mathbf{D}$	$\frac{As}{m^2}$	Electric flux density or electric induction
$\mathbf{H}$	$\frac{A}{m}$	Magnetic field intensity
$\mathbf{B}$	$T = \frac{Vs}{m^2}$	Magnetic flux density or magnetic induction
$\mathbf{J}$	$\frac{A}{m^2}$	Current density
ϱ	$\frac{As}{m^3}$	Electrical charge density
$\mathbf{M}$	$\frac{A}{m}$	Magnetization
$\mathbf{P}$	$\frac{As}{m^2}$	Electric polarization
σ	$\frac{1}{\Omega m} = \frac{S}{m}$	Electric conductivity
ε	$\frac{As}{Vm}$	Electric permittivity
μ	$\frac{Vs}{Am}$	Magnetic permeability

Table 5.1: Description of the electromagnetic quantities and material parameters.

crossing a closed surface Γ is equal to the total electric charge ϱ within the volume Ω. This means that the sources of the electric field are the electric charges and therefore the electric field is irrotational, see [102]. For a more detailed discussion of the fundamental equations (5.1) we refer the interested reader to [95, 96, 102]. It is important to note that the set of equations (5.1) imply the continuity equation. In fact, taking the divergence of (5.1a) and combining it with (5.1d) gives

$$\partial_t \varrho + \operatorname{div} \mathbf{J} = 0. \tag{5.5}$$

This corresponds to the conservation of charges, which can be seen from

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\omega(t)} \varrho(y,t) \mathrm{d}y = \int_{\omega(t)} \partial_t \varrho(y,t) + \operatorname{div}(\varrho(y,t)\mathbf{v}(y,t)) \mathrm{d}y = 0,$$

where we used Reynold's transport theorem [61, Satz 5.4] and the relation $\mathbf{J} = \varrho\mathbf{v}$.

The material parameters ε, μ, σ are in general tensors of rank 2 which depend on space, time and the field quantities $\mathbf{E}, \mathbf{H}$. We assume isotropic material in all our computations, which means that the material parameters become scalar quantities, see e.g., [76, 136]. Further, we assume that the material parameters are constant in time. In so-called linear materials the material parameters are independent of the field quantities, cf. [174]. Of special interest in the framework of electric machine simulation is the magnetic reluctivity ν, which is defined as the reciprocal of the magnetic permeability, i.e., $\nu := \mu^{-1}$. In ferromagnetic material it will be a nonlinear function depending on the field quantity $\mathbf{B}$, i.e., $\nu = \nu(y, |\mathbf{B}|)$. Note that we will neglect in our simulation any kind of hysteresis effects [95]. For electric machine simulations including hysteresis effects we refer to [83].

The constitutive relation (5.2) consists of an impressed current density $\mathbf{J}_i$ and a conduction current density $J_c = \sigma(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, which is due to the force relation by Lorentz [102]. In terms of electric machine simulation the impressed current density is given via the excitation of the coils. The conduction currents are used to model so-called eddy currents in electric machines. Note that the second term in the conduction currents is due to the rotor, which is moving within the magnetic field $\mathbf{B}$. The eddy currents arise in metalic bodies if excited by time varying magnetic fields [174]. This is for instance the case in the permanent magnets of the machine, but they do not occur in the ferromagnetic part of the stator and rotor since they are made of laminated steel sheets which prevent the appearance of eddy currents.

In what follows we will derive, see e.g., [100], the common physical model to describe electric machines, which is the eddy current or magneto-quasistatic approximation of Maxwell's equations in its vector potential formulation. The eddy current approximation is well suited for low frequency applications which is the case for electric machines, transformers or relays [174]. In this case the displacement currents $\frac{\partial D}{\partial t}$ can be neglected and hence (5.1d) decouples from (5.1a)-(5.1c), and we end up with the system

$$\operatorname{curl} \mathbf{H} = \mathbf{J}, \ \operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \ \operatorname{div} \mathbf{B} = 0.$$

Since $\mathbf{B}$ is divergence free there exists a vector potential $\tilde{\mathbf{A}}$ which is unique up to a gradient field such that

$$\mathbf{B} = \operatorname{curl} \tilde{\mathbf{A}} \tag{5.6}$$

Inserting (5.6) into Faraday's law (5.1b) gives

$$\operatorname{curl} \left(\mathbf{E} + \frac{\partial \tilde{\mathbf{A}}}{\partial t} \right) = 0.$$

Due to the identity $\operatorname{curl} \nabla(.) = 0$ we conclude that there exists a scalar field φ such that

$$-\nabla \varphi = \mathbf{E} + \partial_t \tilde{\mathbf{A}}.$$

Now, we introduce a modified vector potential via

$$\mathbf{A}(y,t) = \tilde{\mathbf{A}}(y,t) + \int_0^t \nabla \varphi(y,s) \mathrm{d}s.$$

For this modified vector potential we have $\mathbf{B} = \operatorname{curl} \mathbf{A}$ and $\partial_t A = \partial_t \tilde{\mathbf{A}} + \nabla \varphi$. Thus, it holds

$$\mathbf{E} = -\partial_t \tilde{\mathbf{A}} - \nabla \varphi = -\partial_t \mathbf{A}.$$

Using this and that by (5.4) it holds that $\mathbf{H} = \nu \mathbf{B} - \mathbf{M}$ we obtain from Ampère's law (5.1a)

$$\sigma \left[\partial_t \mathbf{A} + \operatorname{curl} \mathbf{A} \times \mathbf{v} \right] + \operatorname{curl}(\nu \operatorname{curl} \mathbf{A}) = \mathbf{J}_i + \operatorname{curl} \mathbf{M}, \tag{5.7}$$

which is the vector potential formulation of the eddy current problem. For the solution of (5.7) one needs to impose boundary, initial, and interface conditions. For this reason let $D \subset \mathbb{R}^3$, $T > 0$ and $\Gamma := \partial D$. Further we denote with Γ_I the interface where the reluctivity ν and the electrical conductivity σ jumps. We introduce n as the outer unit vector on $\Gamma \times (0,T)$, $\Gamma_I \times (0,T)$, respectively. A possible choice in terms of electric machines is to consider

$$\begin{aligned}
\mathbf{A} \times n &= 0 \quad \text{on } \Gamma \times (0,T),\\
\mathbf{A}(.,0) &= 0 \quad \text{in } \Omega,\\
[\![\mathbf{A} \times n]\!] &= 0 \quad \text{on } \Gamma_I \times (0,T),\\
[\![\nu n \times \operatorname{curl} \mathbf{A}]\!] &= 0 \quad \text{on } \Gamma_I \times (0,T),\\
[\![\sigma n \cdot (\partial_t \mathbf{A} + \operatorname{curl} \mathbf{A} \times \mathbf{v})]\!] &= 0 \quad \text{on } \Gamma_I \times (0,T),
\end{aligned}$$

where $[\![v]\!]$ denotes the jump of a function v along the interface, i.e., cf. [76, 104],

$$[\![v]\!] = v^+\big|_{\Gamma_I \times (0,T)} - v^-\big|_{\Gamma_I \times (0,T)}.$$

Here v^+, v^- are the restrictions of v to the corresponding subdomains, cf. [76]. The boundary condition $\mathbf{A} \times n = 0$ implies that $\mathbf{B} \cdot n = 0$ on $\Gamma \times (0,T)$, see [136]. This means that no magnetic flux leaves the computational domain [40, 104]. This is also called induction boundary condition, see e.g., [76].

It is common practice to reduce (5.7) for the simulation of electric machines to a two-dimensional model in space, see e.g., [40, 76, 77, 78, 136]. For this we assume that the spatial computational domain $D \subset \mathbb{R}^3$ is of the form

$$D = \Omega \times (-\ell, \ell) \text{ with } \ell \gg \operatorname{diam}(\Omega),$$

i.e., one spatial component is much larger than the other two components and that the current density $\mathbf{J}_i$, the magnetization $\mathbf{M}$, the magnetic field intensity $\mathbf{H}$ and the velocity $\mathbf{v}$ are of the form

$$\mathbf{J}_i = \begin{bmatrix} 0 \\ 0 \\ f(y_1, y_2, t) \end{bmatrix}, \mathbf{M} = \begin{bmatrix} M_1(y_1, y_2, t) \\ M_2(y_1, y_2, t) \\ 0 \end{bmatrix},$$
$$\mathbf{H} = \begin{bmatrix} H_1(y_1, y_2, t) \\ H_2(y_1, y_2, t) \\ 0 \end{bmatrix}, \ \mathbf{v} = \begin{bmatrix} v_1(y_1, y_2, t) \\ v_2(y_1, y_2, t) \\ 0 \end{bmatrix}.$$

From the constitutive relation (5.4) we immediately obtain that the magnetic flux density is of the same form as $\mathbf{H}$ and $\mathbf{M}$, i.e,

$$\mathbf{B} = \begin{bmatrix} B_1(y_1, y_2, t) \\ B_2(y_1, y_2, t) \\ 0 \end{bmatrix}.$$

This can be achieved by the ansatz

$$\mathbf{A} = \begin{bmatrix} 0 \\ 0 \\ u(y_1, y_2, t) \end{bmatrix}, \tag{5.8}$$

since then it holds

$$\mathbf{B} = \operatorname{curl} \mathbf{A} = \begin{bmatrix} \partial_{y_2} u \\ -\partial_{y_1} u \\ 0 \end{bmatrix}.$$

Note that the ansatz of the vector potential (5.8) also ensures the Coloumb gauge condition $\operatorname{div} \mathbf{A} = 0$. Plugging (5.8) into (5.7) yields a two-dimensional partial differential equation in space for the scalar field u which is posed on the cross-section $\Omega(t) \subset \mathbb{R}^2$ of the electric machine. The cross-section $\Omega(t)$ depends on the time t as the rotor is moving. This can be described via a bijective and sufficiently smooth deformation mapping $\varphi : \Omega \times (0,T) \to \mathbb{R}^2$, $(x,t) \mapsto y(t) := \varphi_t(x) := \varphi(x,t)$ which satisfies $v(y,t) = \frac{\mathrm{d}}{\mathrm{d}t} y(t)$ and $\varphi(x,0) = x$. The domain $\Omega \subset \mathbb{R}^2$ describes the cross-section of the electric motor at initial time $t = 0$, and we will also refer to it as reference domain. This domain is then transported via the mapping φ to the deformed domain $\Omega(t)$, i.e., $\Omega(t) = \varphi_t(\Omega)$. Introducing the space-time domain

$$Q := \{(y,t) \in \mathbb{R}^3 : \ y = \varphi(x,t), x \in \Omega, \ t \in (0,T)\}$$

together with its lateral boundary

$$\Sigma := \{(y,t) : \ y = \varphi(x,t), \ x \in \partial\Omega, \ t \in (0,T)\}$$

the resulting initial boundary value problem together with the simplified boundary, initial and transmisson conditions reads

$$\begin{aligned}
\sigma\left[\partial_t u + \nabla_y u \cdot v\right] - \operatorname{div}_y(\nu \nabla_y u) &= f - \operatorname{div}_y M^{\perp} && \text{in } Q, && \text{(5.9a)}\\
u &= 0 && \text{on } \Sigma, && \text{(5.9b)}\\
u(.,0) &= 0 && \text{in } \Omega, && \text{(5.9c)}\\
[\![u]\!] &= 0 && \text{on } \Gamma_I(t) \times (0,T), && \text{(5.9d)}\\
[\![\nu \nabla_y u \cdot n]\!] &= 0 && \text{on } \Gamma_I(t) \times (0,T). && \text{(5.9e)}
\end{aligned}$$

Note that for the reluctivity we have $\nu = \nu(y, |\nabla_y u|)$, and we denote with $M^{\perp} = (-M_2, M_1)^{\top}$ the counterclockwise rotation of the vector $M = (M_1, M_2)^{\top}$.

5.2 Physical properties of B-H-curves

In this section we will discuss some properties of B-H-curves. We follow the presentation in [76, 136, 137]. These properties play a key role in the analysis of the related

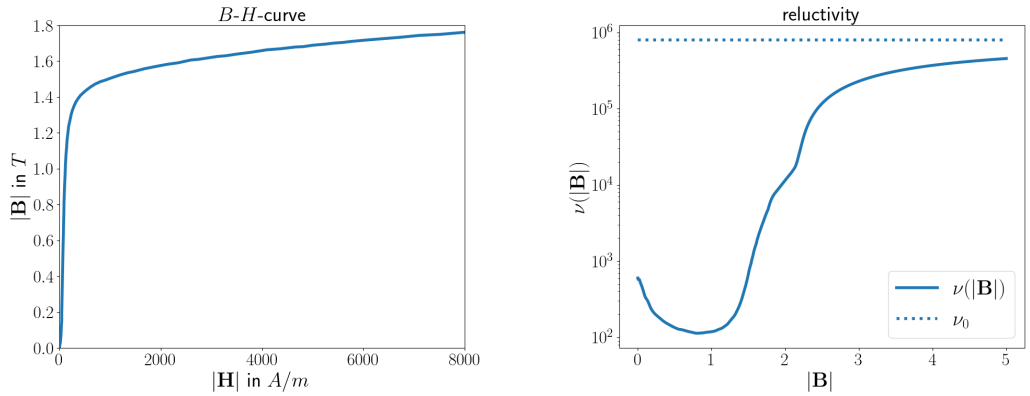


Figure 5.1: Left: B-H-curve f. Right: Corresponding reluctivity in semilogarithmic plot.

nonlinear partial differential equations occurring in electromagnetic field computations. In particular, they ensure that the resulting differential operators admit some monotonicity property. As pointed out we restrict ourselves to the case of isotropic materials and neglect any kind of hysteresis effect [18], cf. [76, 136].

The magnetic permeability μ as well as its reciprocal the magnetic reluctivity describe the connection between the magnetic flux density $\mathbf{B}$ and the magnetic field strength $\mathbf{H}$ via the constitutive relation (5.4). In many materials this can be described via a linear relation $|\mathbf{B}| = \mu|\mathbf{H}|$, where μ is just a constant value [76]. However, electric machines are also made of laminated steel sheets, which behave as ferromagnetic material. In ferromagnetic material the relation between $\mathbf{B}$ and $\mathbf{H}$ is nonlinear and described by a B-H-curve, see e.g., [101, 137],

$$f : \mathbb{R}_0^+ \to \mathbb{R}_0^+, \; H \mapsto B = f(H), \tag{5.10}$$

where $H = |\mathbf{H}|$, $B = |\mathbf{B}|$ denotes the intensity of the field quantities and $\mathbb{R}_0^+$ represents the set of nonnegative real numbers. Based on (5.10) one defines the permeability and reluctivity via

$$\mu(s) = \frac{f(s)}{s}, \quad \nu(s) = \frac{f^{-1}(s)}{s},$$

which gives the relations

$$\mathbf{B} = \mu(|\mathbf{H}|)\mathbf{H}, \quad \mathbf{H} = \nu(|\mathbf{B}|)\mathbf{B}.$$

In Fig. 5.1 we depict a typical example of a B-H-curve f. As pointed out in [137] we see for small values of H a strong amplification in B, whereas for high values of H this amplification gets more to the one of vacuum. These observations lead to the following assumptions on a B-H-curve.

ASSUMPTION 5.1 ([76, 136, 137]). *Let $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ be a B-H-curve. Then we assume that f fulfills*

(i) *f is continuously differentiable on $\mathbb{R}_0^+$,*

(ii) *$f(0) = 0$,*

(iii) *$f'(s) \geq \mu_0$ for all $s \geq 0$,*

(iv) *$\lim_{s\to\infty} f'(s) = \mu_0$.*

Here, $\mu_0 = 4\pi \cdot 10^{-7}\ \frac{Vs}{Am}$ denotes the permeability of vacuum, and we denote with $\nu_0 = 1/\mu_0$ the corresponding reluctivity. The following Lemma is a consequence of Assumption 5.1.

LEMMA 5.2 ([76, 105, 136]). *Let Assumption 5.1 be satisfied. Then there holds:*

(i) *ν is continuously differentiable on $(0, \infty)$ and $\nu'(s) \to 0$ for $s \to \infty$.*

(ii) *There exists a constant $\underline{\nu} > 0$ such that for all $s \in \mathbb{R}_0^+$ we have*

$$\begin{aligned} \underline{\nu} \leq \nu(s) \leq \nu_0, \\ \underline{\nu} \leq (\nu(s)s)' \leq \nu_0. \end{aligned}$$

(iii) *The mapping $s \mapsto \nu(s)s$ is strongly monotone with monotonicity constant $\underline{\nu}$, i.e.,*

$$(\nu(s)s - \nu(t)t)(s - t) \geq \underline{\nu}(s - t)^2 \quad \forall s, t \in \mathbb{R}_0^+,$$

and Lipschitz continuous with Lipschitz constant ν_0, i.e.,

$$|\nu(s)s - \nu(t)t| \leq \nu_0|s - t| \quad \forall s, t \in \mathbb{R}_0^+.$$

REMARK 5.3. *The properties of the reluctivity ν shown in Lemma 5.2 ensure well-posedness of the eddy current problem* (5.9)*, see [40, 83].*

In practice the B-H-curve has to be approximated from measurement data, see [76]. It is important that the interpolation of such real life measurements preserves the monotonicity. In terms of B-H-curves we want to mention the work by Heise [90] which is based on a cubic interpolation [69] and the more recent works of Pechstein, Jüttler [137] and Kaltenbacher B., Kaltenbacher M., Reitzinger [101].

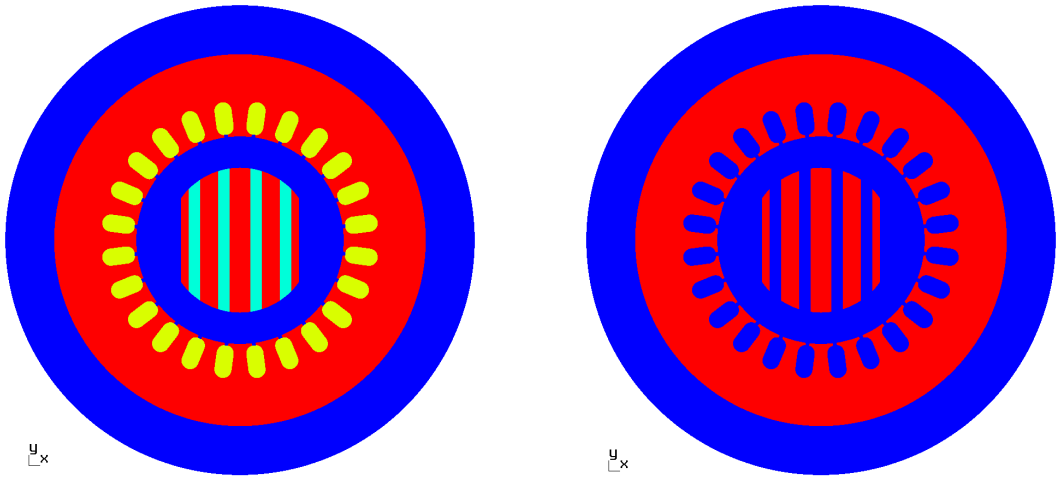


Figure 5.2: Left: Computational domain Ω indicating different materials of the machine. Red: iron, blue: air, yellow: copper, cyan: damping material. Right: Domain Ω divided into ferromagnetic material Ω_f and nonferromagnetic material $\Omega \setminus \Omega_f$.

5.3 Synchronous reluctance motor

In this section we consider a synchronous reluctance machine, which is intended for the use in an X-ray tube for medical applications, see [79, 124, 125]. We apply the minimal residual framework from Chapter 3 in a linear as well as nonlinear setting. Moreover, we demonstrate the application of the method for a fixed rotor position in a 2d spatial setting and for a moving rotor in the spirit of a 2d-1d space-time computation.

The basic computational domain Ω indicating different subdomains of the machine can be seen in Fig. 5.2. It consists of a stator and rotor, which are separated via an airgap indicated in blue. The ferromagnetic regions in stator and rotor are highlighted in red. The stator has 24 coils marked in yellow which correspond to copper and are of nonferromagnetic type. Moreover, the rotor has a layered design whose nonferromagnetic layers are depicted in cyan. Note that Fig. 5.2 also includes an additional air layer around the stator. In the static regime we will include such a layer, while for the quasistatic and eddy current approximation we will neglect this layer in order to safe some dofs since in the space-time setting we have to deal with a three-dimensional geometry. In what follows we denote with Ω_f the ferromagnetic subdomain of the computational domain Ω as depicted in the right plot of Fig. 5.2.

5.3.1 Magnetostatic problem for a fixed rotor position

In the following we consider a special regime of the eddy current problem (5.9), namely the magnetostatic approximation. In this setting it is assumed that the rotor moves at a constant speed, see [76]. Furthermore, we consider a fixed position of the rotor, where the impressed current density $f = f(x_1, x_2)$ is chosen such that the motor generates its maximal torque. Hence, all electromagnetic quantities are independent of time and (5.9) boils down to

$$
\begin{aligned}
-\operatorname{div}(\nu(x, |\nabla u(x)|)\nabla u(x)) &= f(x) && \text{for } x \in \Omega, && \text{(5.11a)}\\
u(x) &= 0 && \text{for } x \in \Gamma := \partial\Omega, && \text{(5.11b)}\\
[\![u(x)]\!] &= 0 && \text{for } x \in \Gamma_I, && \text{(5.11c)}\\
[\![\nu(x, |\nabla u(x)|)\nabla u(x) \cdot n(x)]\!] &= 0 && \text{for } x \in \Gamma_I, && \text{(5.11d)}
\end{aligned}
$$

where Γ_I denotes the interface between ferromagnetic and nonferromagnetic material. Note that the magnetization M vanishes since the motor does not have permanent magnets. We consider a linear as well as a nonlinear computation of the electric motor. For this we introduce the linear reluctivity $\nu_l(x)$ and the nonlinear reluctivity $\nu(x, |\nabla u(x)|)$ via

$$
\begin{aligned}
\nu_l(x) &:= \begin{cases} \nu_f = \frac{1}{\mu_0 \cdot \mu_{r,Fe}} & \text{for } x \in \Omega_f \\ \nu_0 & \text{for } x \in \Omega \setminus \Omega_f \end{cases}, \\
\nu(x, |\nabla u(x)|) &:= \begin{cases} \hat{\nu}(|\nabla u(x)|) & \text{for } x \in \Omega_f \\ \nu_0 & \text{for } x \in \Omega \setminus \Omega_f \end{cases}.
\end{aligned}
\qquad (5.12)
$$

Here $\nu_0 = 10^7/4\pi \, \frac{Am}{Vs}$ denotes the reluctivity of vacuum, air, respectively, which is the reciprocal of the permeability of vacuum $\mu_0 = 4\pi/10^7 \, \frac{Vs}{Am}$. Note that the coils in the stator of the machine obtain the value ν_0 as the ferromagnetic behaviour of copper is the same as of air. The value ν_f denotes the reluctivity of the ferromagnetic material if we assume it to behave linear. It is given as the reciprocal of the product between the permeability of vacuum and the relative permeability of iron which we choose to be $\mu_{r,Fe} = 3978$. This is a realistic appproximation when saturation of the material does not occur, see [78]. The function $s \mapsto \hat{\nu}(s)$ denotes a nonlinear reluctivity curve stemming from a B-H-curve f fulfilling the Assumption 5.1.

Linear case

In the linear setting we have in view of the abstract setting in Section 3.1 the spaces

$$X = Y = H_0^1(\Omega),$$

together with the norm

$$\|v\|_X := \sqrt{\langle \nu_l \nabla v, \nabla v \rangle_{L^2(\Omega)}}.$$

The operators $A = B : H_0^1(\Omega) \to H^{-1}(\Omega)$ are defined in the variational sense satisfying

$$\langle Bu, q \rangle_\Omega := \langle \nu_l \nabla u, \nabla q \rangle_{L^2(\Omega)}$$

for all $(u, q) \in H_0^1(\Omega) \times H_0^1(\Omega)$, where $\langle \cdot, \cdot \rangle_\Omega$ denotes the duality pairing. We obviously have that the operators A, B fulfill the assumptions of Section 3.1 with the constants $c_1^A = c_1^B = c_2^A = c_2^B = 1$. The abstract variational formulation (3.12) then reads to find $(u, p) \in H_0^1(\Omega) \times H_0^1(\Omega)$ such that

$$\langle \nu_l \nabla p, \nabla q \rangle_{L^2(\Omega)} + \langle \nu_l \nabla u, \nabla q \rangle_{L^2(\Omega)} = \langle f, q \rangle_\Omega, \quad \langle \nu_l \nabla v, \nabla p \rangle_{L^2(\Omega)} = 0 \tag{5.13}$$

is satisfied for all $(v, q) \in H_0^1(\Omega) \times H_0^1(\Omega)$. For the discretization we choose the finite dimensional subspaces $X_H = S_H^1(\Omega) \cap X = \text{span}\{\varphi_i\}_{i=1}^{M_X}$ and $Y_h = Y_H = S_H^2(\Omega) \cap Y = \text{span}\{\psi_j\}_{j=1}^{M_Y}$ which are defined with respect to some admissible and locally quasi-uniform decomposition $\mathcal{T}_H$ of Ω into shape regular simplicial finite elements. Obviously we have $X_H \subset Y_h$ and the discrete inf-sup condition

$$\|u_H\|_X = \frac{\langle \nu_l \nabla u_H, \nabla u_H \rangle_{L^2(\Omega)}}{\|u_H\|_Y} \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle \nu_l \nabla u_H, \nabla q_h \rangle_{L^2(\Omega)}}{\|q_h\|_Y} = \sup_{0 \neq q_h \in Y_h} \frac{\langle Bu_H, q_h \rangle_\Omega}{\|q_h\|_Y},$$

i.e., (3.15). This ensures unique solvability of the discrete variational formulation (3.13), which in this case reads to find $(u_H, p_h) \in X_H \times Y_h$ such that

$$\langle \nu_l \nabla p_h, \nabla q_h \rangle_{L^2(\Omega)} + \langle \nu_l \nabla u_H, \nabla q_h \rangle_{L^2(\Omega)} = \langle f, q_h \rangle_\Omega, \quad \langle \nu_l \nabla v_H, \nabla p_h \rangle_{L^2(\Omega)} = 0 \tag{5.14}$$

is satisfied for all $(v_H, q_h) \in X_H \times Y_h$. In the numerical experiment we solve the related linear system to (5.14), i.e., (3.14) with the matrices and vector

$$A_h[j,i] = \langle \nu_l \nabla \psi_i, \nabla \psi_j \rangle_{L^2(\Omega)}, \; B_h[j,k] = \langle \nu_l \nabla \varphi_k, \nabla \psi_j \rangle_{L^2(\Omega)}, \; f_j = \langle f, \psi_j \rangle_\Omega,$$

where $i, j = 1, ..., M_Y$ and $k = 1, ..., M_X$. Furthermore, we use the global error indicator

$$\eta_H^2 = \|p_h\|_Y^2 = \langle \nu_l \nabla p_h, \nabla p_h \rangle_{L^2(\Omega)} = \sum_{\tau \in \mathcal{T}_H} \langle \nu_l \nabla p_h, \nabla p_h \rangle_{L^2(\tau)} = \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2,$$

with the local indicators

$$\eta_\tau^2 = \langle \nu_l \nabla p_h, \nabla p_h \rangle_{L^2(\tau)} \quad \text{for } \tau \in \mathcal{T}_H,$$

to drive an adaptive refinement scheme. As a marking strategy we use the Dörfler criterion [58] with parameter $\theta = 0.5$. The marked elements are refined using

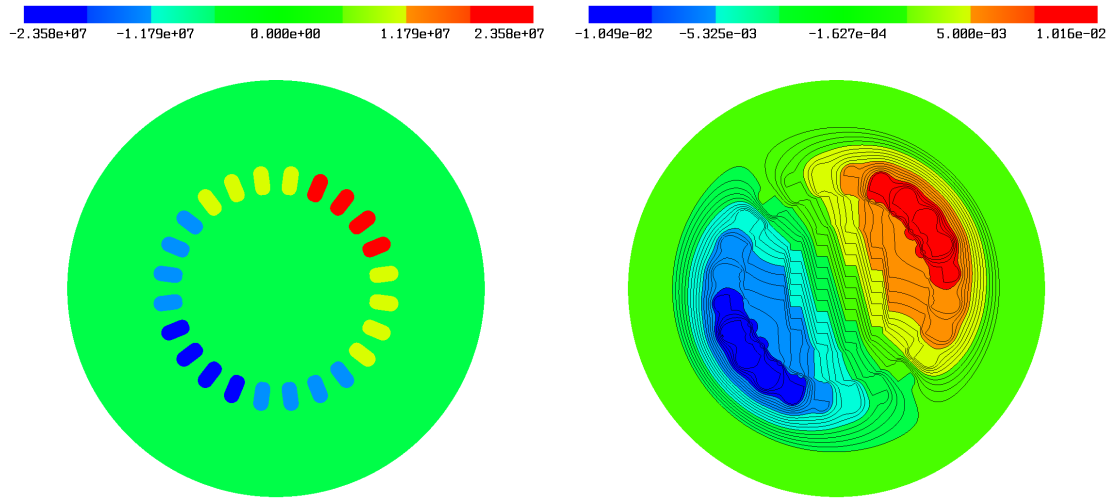


Figure 5.3: Left: Impressed current density f in the coils. Right: Numerical solution u_H obtained on refinement level $L = 8$.

newest vertex bisection. For the implementation we used the finite element software `Netgen/NGSolve` [152], where we used the sparse direct solver `Pardiso` [149] to solve the resulting linear systems.

In Fig. 5.3 we depict the impressed current density f used in our simulations and the corresponding numerical solution u_H obtained on refinement level $L = 8$ on a mesh with $\widetilde{M}_X = 120091$ vertices. The current density was provided from a reference simulation of C. Mellak using JMAG [99], see [79]. The maximal value of the current density indicated in red is given by $j_{max} = 2357.72 \cdot 10^4 \frac{A}{m^2}$. The coils indicated in yellow obtain the value $j_{max}/2$, light blue correspond to $-j_{max}/2$ and coils in dark blue admit the value $-j_{max}$. In Fig. 5.4 we see the generated adaptive mesh obtained from the adaptive refinement process. We started with an initial mesh of 13485 dofs. The resulting adaptive mesh shows stronger refinements around the coils, the air gap as well as the interfaces between ferromagnetic and nonferromagnetic material. In Fig. 5.5 we show the corresponding convergence behaviour of the error indicator η_H. For higher refinement levels we see a linear convergence rate of the estimator, which is expected as the current density satisfies $f \in L^2(\Omega)$.

Nonlinear case

In the nonlinear setting we have the same spaces $X = Y = H_0^1(\Omega)$ as in the linear case, but we use the norms

$$\|u\|_X := \|\nabla u\|_{L^2(\Omega)}, \quad \|p\|_Y := \sqrt{\langle \nu_l \nabla p, \nabla p \rangle_{L^2(\Omega)}}.$$

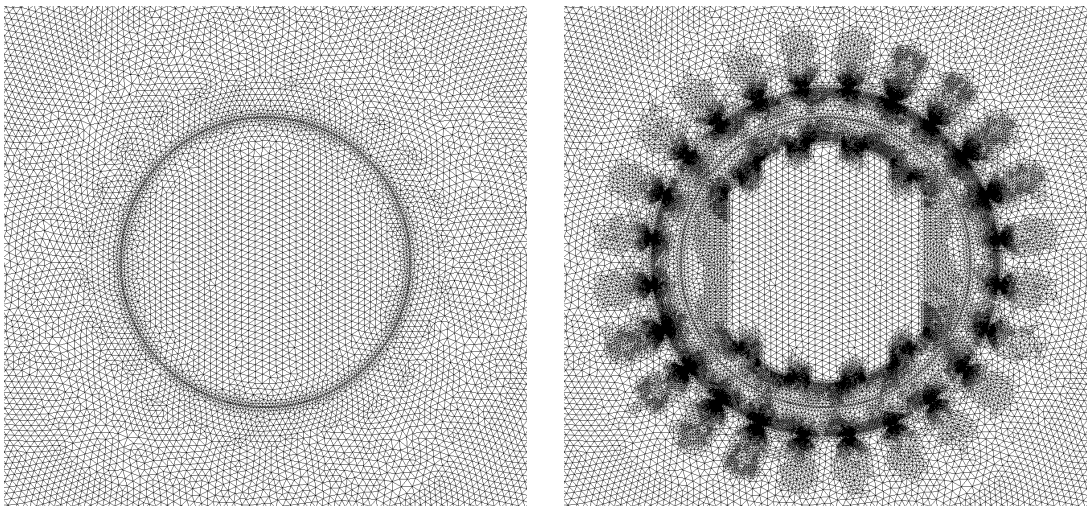


Figure 5.4: Left: Initial mesh with 13485 dofs. Right: Adaptive refined mesh on level $L = 7$ with 68425 dofs obtained for a linear material behaviour.

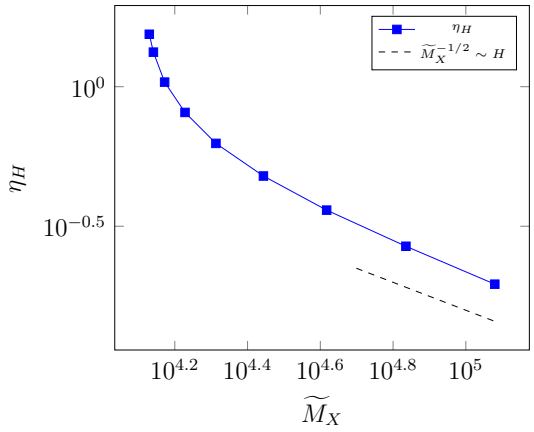


Figure 5.5: Convergence behaviour of the error estimator η_H during the adaptive refinement process in case of a linear material behaviour.

The operators $A, B : H_0^1(\Omega) \to H^{-1}(\Omega)$ are defined in the variational sense and read

$$\begin{aligned} \langle Ap, q\rangle_\Omega &:= \langle \nu_l \nabla p, \nabla p\rangle_{L^2(\Omega)}, \\ \langle B(u), q\rangle_\Omega &:= \int_\Omega \nu(x, |\nabla u(x)|)\nabla u(x) \cdot \nabla q(x)\ \mathrm{d}x \end{aligned} \tag{5.15}$$

for all $u, p, q \in H_0^1(\Omega)$. As in the linear case we have that A is a bounded, self-adjoint, and elliptic opertator. For the properties of the operator B we state the following result.

LEMMA 5.4. *Let $s \mapsto \hat{\nu}(s)$ from $\mathbb{R}_0^+ \to \mathbb{R}_0^+$ denote a nonlinear reluctivity curve stemming from a B-H-curve fulfilling Assumption 5.1 and let the global reluctivity curve $(x, s) \mapsto \nu(x, s)$ be given as in (5.12). Then the operator B is strongly monotone and Lipschitz continuous, i.e., it holds*

$$\langle B(u) - B(v), u - v\rangle_\Omega \geq \underline{\nu} \|\nabla(u - v)\|_{L^2(\Omega)}^2$$

and

$$\|B(u) - B(v)\|_{H^{-1}(\Omega)} \leq 3\nu_0 \|\nabla(u - v)\|_{L^2(\Omega)}.$$

Proof. First, we remark that the assumptions on the mapping $s \mapsto \hat{\nu}(s)$ for $s \in \mathbb{R}_0^+$ ensure that the map $s \mapsto \nu(x, s)s$ is strongly monotone with constant $\underline{\nu}$ and Lipschitz continuous with constant ν_0 for all $x \in \Omega$. Indeed, by Lemma 5.2 it follows that the mapping $s \mapsto \hat{\nu}(s)s$ is strongly monotone with constant $\underline{\nu}$ and Lipschitz continuous with constant ν_0. Hence, for $x \in \Omega_f$ the assertion is fulfilled. For $x \in \Omega \setminus \Omega_f$ we have $s \mapsto \nu(x, s)s = \nu_0 s$ which clearly is strongly monotone and Lipschitz continuous. Thus, we conclude the assertion on the mapping $s \mapsto \nu(x, s)s$ involving the global reluctivity $\nu(x, s)$. Now, an application of [136, Lem. 2.8, Lem. 2.9], compare also [182, pp. 130-131] gives the desired properties for the operator B. □

The properties of Lemma 5.4 ensure that the nonlinear operator equation

$$B(u) = f \quad \text{in } H^{-1}(\Omega) \tag{5.16}$$

admits a unique solution, see Thm. 2.11. For the application of the minimal residual framework we also need the derivative of the operator B. In order to compute the derivative we remark that the mapping $W \to \hat{\nu}(|W|)$ for $W \in \mathbb{R}^2$ is not differentiable in $W = (0, 0)^\top$, but $W \to \hat{\nu}(|W|)W$ is. Therefore, we introduce similar as in [76] the operator $\hat{T} : \mathbb{R}^2 \to \mathbb{R}^2$

$$\hat{T}(W) := \hat{\nu}(|W|)W$$

and the operator $T : \Omega \times \mathbb{R}^2 \to \mathbb{R}^2$

$$T(x, W) := \nu(x, |W|)W = \begin{cases} \hat{T}(W) & \text{for } x \in \Omega_f \\ \nu_0 W & \text{for } x \in \Omega \setminus \Omega_f \end{cases}$$

involving the global reluctivity. Thus, the operator B defined in (5.15) allows the representation

$$\langle B(u), q\rangle_\Omega = \int_\Omega T(x, \nabla u) \cdot \nabla q \,\mathrm{d}x.$$

The Fréchet derivative of B in $u \in H_0^1(\Omega)$ is then given by, see e.g., [76, 105, 136]

$$\begin{aligned} B' : H_0^1(\Omega) &\to \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega)), \\ \langle B'(u)w, q\rangle_\Omega &= \int_\Omega \mathrm{D}T(x, \nabla u)(\nabla w) \cdot \nabla q \,\mathrm{d}x \end{aligned} \tag{5.17}$$

for all $w, q \in H_0^1(\Omega)$, where we have for $W, V \in \mathbb{R}^2$

$$\mathrm{D}T(x, W)(V) = \begin{cases} \mathrm{D}\hat{T}(W)(V) & \text{for } x \in \Omega_f \\ \nu_0 V & \text{for } x \in \Omega \setminus \Omega_f \end{cases} \tag{5.18}$$

with the directional derivative $\mathrm{D}\hat{T}(W)(V)$ given as

$$\mathrm{D}\hat{T}(W)(V) = \begin{cases} \hat{\nu}(|W|)V + \frac{\hat{\nu}'(|W|)}{|W|}(W \cdot V)W & \text{for } W \neq (0,0)^\top \\ \hat{\nu}(0)V & \text{for } W = (0,0)^\top \end{cases}. \tag{5.19}$$

Plugging in (5.18) and (5.19) into (5.17) we obtain for $|\nabla u| \neq 0$ in more detail

$$\begin{aligned} \langle B'(u)w, q\rangle_\Omega = &\int_\Omega \nu(x, |\nabla u(x)|)\nabla w(x) \cdot \nabla q(x) \,\mathrm{d}x \\ &+ \int_{\Omega_f} \frac{\hat{\nu}'(|\nabla u(x)|)}{|\nabla u(x)|}(\nabla u(x) \cdot \nabla w(x))(\nabla u(x) \cdot \nabla q(x)) \,\mathrm{d}x, \end{aligned}$$

and for $|\nabla u| = 0$ we have more precisely

$$\langle B'(u)w, q\rangle_\Omega = \int_{\Omega_f} \hat{\nu}(0)\nabla w(x) \cdot \nabla q(x) \,\mathrm{d}x + \int_{\Omega \setminus \Omega_f} \nu_0 \nabla w(x) \cdot \nabla q(x) \,\mathrm{d}x.$$

In the following we state some properties of the operator B'.

LEMMA 5.5. *Let $u \in H_0^1(\Omega)$. The Fréchet derivative of the operator B given in (5.17), i.e., $B'(u) : H_0^1(\Omega) \to H^{-1}(\Omega)$, is bounded with constant ν_0 and elliptic with constant $\underline{\nu}$ uniformly for any fixed $u \in H_0^1(\Omega)$, i.e., it holds*

$$\langle B'(u)w, q\rangle_\Omega \leq \nu_0 \|\nabla w\|_{L^2(\Omega)} \|\nabla q\|_{L^2(\Omega)}, \quad \langle B'(u)w, w\rangle_\Omega \geq \underline{\nu} \|\nabla w\|_{L^2(\Omega)}^2 \tag{5.20}$$

for all $w, q \in H_0^1(\Omega)$.

Proof. First note that (5.18) can be written as,

$$\mathrm{D}T(x,W)(V) = JT(x,W)V, \quad JT(x,W) = \begin{cases} J\hat{T}(W) & \text{for } x \in \Omega_f \\ \nu_0 I & \text{for } x \in \Omega \setminus \Omega_f \end{cases},$$

where $JT(x,W) \in \mathbb{R}^{2\times 2}$ denotes the Jacobian of T, and $I \in \mathbb{R}^{2\times 2}$ the identity matrix. The Jacobian $J\hat{T}(W) \in \mathbb{R}^{2\times 2}$ of the mapping $\hat{T}$ is given by means of the directional derivative (5.19) and reads

$$J\hat{T}(W) = \begin{cases} \hat{\nu}(|W|)I + \frac{\hat{\nu}'(|W|)}{|W|} WW^\top & \text{for } W \neq (0,0)^\top \\ \hat{\nu}(0)I & \text{for } W = (0,0)^\top \end{cases}.$$

The eigenvalues and eigenvectors of $J\hat{T}(W)$ are, see e.g., [76, 136, 89]

$$\begin{aligned} \lambda_1 &= \hat{\nu}(|W|), & v_1 &= W^\perp = (-w_2, w_1)^\top, \\ \lambda_2 &= \hat{\nu}(|W|) + \hat{\nu}'(|W|)|W|, & v_2 &= W = (w_1, w_2)^\top \end{aligned} \tag{5.21}$$

We have due to the properties of $\hat{\nu}$, see Lem. 5.2, that

$$\lambda_{min} := \min\{\lambda_1, \lambda_2\} \geq \underline{\nu}, \quad \lambda_{max} := \max\{\lambda_1, \lambda_2\} \leq \nu_0 \tag{5.22}$$

Since $J\hat{T}(W)$ is symmetric we conclude

$$\underline{\nu}|Z|^2 \leq (J\hat{T}(W)Z) \cdot Z \leq \nu_0 |Z|^2, \tag{5.23}$$

which is independent of $W \in \mathbb{R}^2$. For the global Jacobian $JT(x,W)$ we either have the eigenvalues given in (5.21) if $x \in \Omega$ or $\lambda_1 = \lambda_2 = \nu_0$ if $x \in \Omega \setminus \Omega_f$. In any case we obtain the same estimates for the minimal and maximal eigenvalue as in (5.22) and since $JT(x,W)$ is also symmetric the inequalities (5.23) remain valid also for the global Jacobian $JT(x,W)$, i.e.,

$$\underline{\nu}|Z|^2 \leq JT(x,W)Z \cdot Z \leq \nu_0 |Z|^2 \tag{5.24}$$

for any $x \in \Omega$ and $W \in \mathbb{R}^2$. Now, using (5.24) we can estimate

$$\langle B'(u)w, w\rangle_\Omega = \int_\Omega JT(x, \nabla u)\nabla w \cdot \nabla w \,\mathrm{d}x \geq \underline{\nu} \int_\Omega |\nabla w|^2 \,\mathrm{d}x = \underline{\nu}\|\nabla w\|^2_{L^2(\Omega)}$$

which gives ellipticity uniformly for any $u \in H^1_0(\Omega)$. Further we conclude

$$\begin{aligned} \langle B'(u)w, q\rangle_\Omega &= \int_\Omega JT(x, \nabla u)\nabla w \cdot \nabla q \,\mathrm{d}x \leq \nu_0 \int_\Omega |\nabla w||\nabla q| \,\mathrm{d}x \\ &\leq \nu_0 \|\nabla w\|_{L^2(\Omega)} \|\nabla q\|_{L^2(\Omega)}, \end{aligned}$$

which is the desired result regarding the boundedness. □

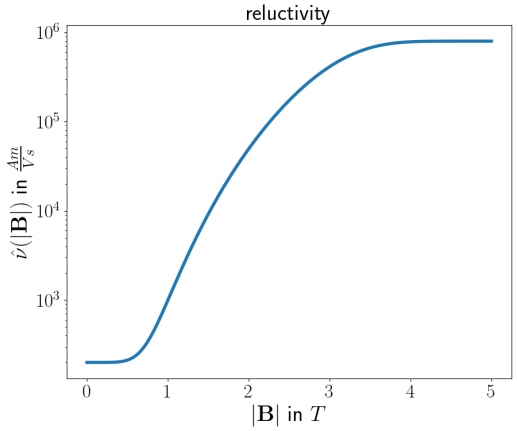


Figure 5.6: Magnetic reluctivity $\hat{\nu}$ defined in (5.26) in semilogarithmic plot.

The results of Lemma 5.5 ensure that the operator $S_u = B'(u)^* A^{-1} B'(u)$ used in Gauß-Newton's method is elliptic and bounded, see Lem. 3.27. Furthermore, by Cor. 3.28 we obtain that the search direction $w_u^k \in H_0^1(\Omega)$ computed from (3.50) is a descent direction.

For the discretization we choose $X_H = S_H^1(\mathcal{T}_H) \cap H_0^1(\Omega) = \operatorname{span}\{\varphi_i\}_{i=1}^{M_X}$ and $Y_h = Y_H = S_H^2(\mathcal{T}_H) \cap H_0^1(\Omega) = \operatorname{span}\{\psi_i\}_{i=1}^{M_Y}$, which are defined with respect to some admissible and locally quasi-uniform decomposition $\mathcal{T}_H$ of Ω into shape regular simplicial finite elements. In order to determine the numerical solution u_H we apply Gauß-Newton's method, i.e., Algorithm 3.31 with the matrices

$$\begin{aligned} A_h[i,j] &= \langle \nu_l \psi_j, \psi_i \rangle_{L^2(\Omega)} \quad i,j = 1,...,M_Y, \\ B_h'(\underline{u})[i,j] &= \int_\Omega \mathrm{D}T(x, \nabla u_H)(\nabla \varphi_j) \cdot \nabla \psi_i \,\mathrm{d}x \quad i = 1,...,M_Y,\ j = 1,...,M_X. \end{aligned} \tag{5.25}$$

The right-hand side f is given as in the linear case. We use an analytic expression for the reluctivity $\hat{\nu}(s)$ defined by

$$\hat{\nu}(s) = \nu_0 - (\nu_0 - c_1) \exp\left(-c_2 s^{c_3}\right) \tag{5.26}$$

with the parameters $c_1 = 200$, $c_2 = 0.001$, $c_3 = 6$, which can be seen in Fig. 5.6. As an initial guess we choose $u_H^{L,0} = 0$ on refinement level $L = 0$ and $u_H^{L,0} = P_H^{L,L-1} u_H^{L-1}$ for $L > 0$, i.e., we prolongate the numerical solution u_H^{L-1} on level $L-1$ via the prolongation operator $P_H^{L,L-1}$ to the latest mesh level. We let the algorithm run until the error is below 10^{-7} and use a backtracking line search strategy. As in the linear case we use the global error indicator

$$\eta_H^2 = \|p_h\|_Y^2 = \langle \nu_l \nabla p_h, \nabla p_h \rangle_{L^2(\Omega)} = \sum_{\tau \in \mathcal{T}_H} \langle \nu_l \nabla p_h, \nabla p_h \rangle_{L^2(\tau)} = \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2,$$

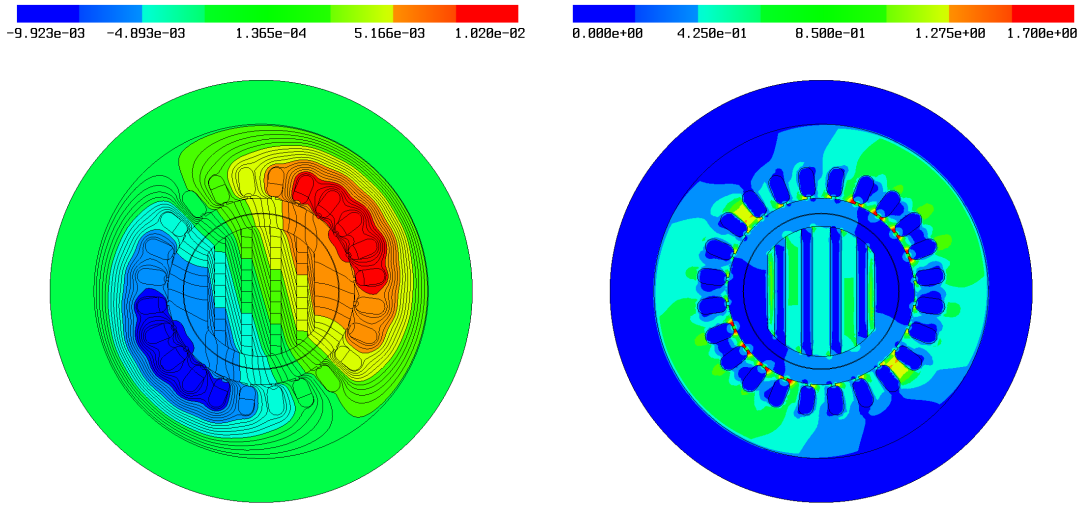


Figure 5.7: Left: Numerical solution u_H obtained on refinement level $L = 6$ with 53013 dofs. Right: Contour plot of the magnetic flux density $B = |\mathbf{B}|$ on $L = 6$.

with the local indicators

$$\eta_\tau^2 = \langle \nu_l \nabla p_h, \nabla p_h \rangle_{L^2(\tau)} \quad \text{for } \tau \in \mathcal{T}_H$$

to drive an adaptive refinement scheme and the Dörfler criterion [58] with parameter $\theta = 0.5$ as marking strategy. All linear systems were solved using the sparse direct solver `Pardiso` [149].

The numerical solution u_H and the corresponding magnetic flux density $B = |\mathbf{B}|$ can be seen in Fig. 5.7. We see that the material is in saturation where the coils admit the maximal and minimal value of the impressed current density. We remark that the maximal value in the contour plot is $B = 1.9\ T$, but for a better visualization of the areas in saturation we scaled the maximal value to $B = 1.7\ T$. The resulting adaptive mesh on $L = 6$ with 53013 dofs can be seen in Fig. 5.8. We started with the same initial mesh on $L = 0$ as in linear case. The adaptive mesh shows stronger refinements at the interface between ferromagnetic and nonferromagnetic material. Furthermore, we see that the refinement of the coils which admit the values j_{max} and $-j_{max}$ is prioritized at least until level $L = 6$. We remark that the adaptive mesh obtained on $L = 7$ consisting of 93999 dofs also shows stronger refinements for the other coils similar as in the linear case, see Fig. 5.4. In contrast to the mesh obtained from a linear calculation the mesh in the nonlinear case exhibits refinements in the stator where the fieldlines of the poles join. In Fig. 5.9 we present the related convergence behaviour of the error estimator η_H. As in the linear case we see a linear convergence rate. In Tab. 5.2 we present the number iterations (iter), the final error (err), and the step size τ in the final iteration of the Gauß-Newton solver during the

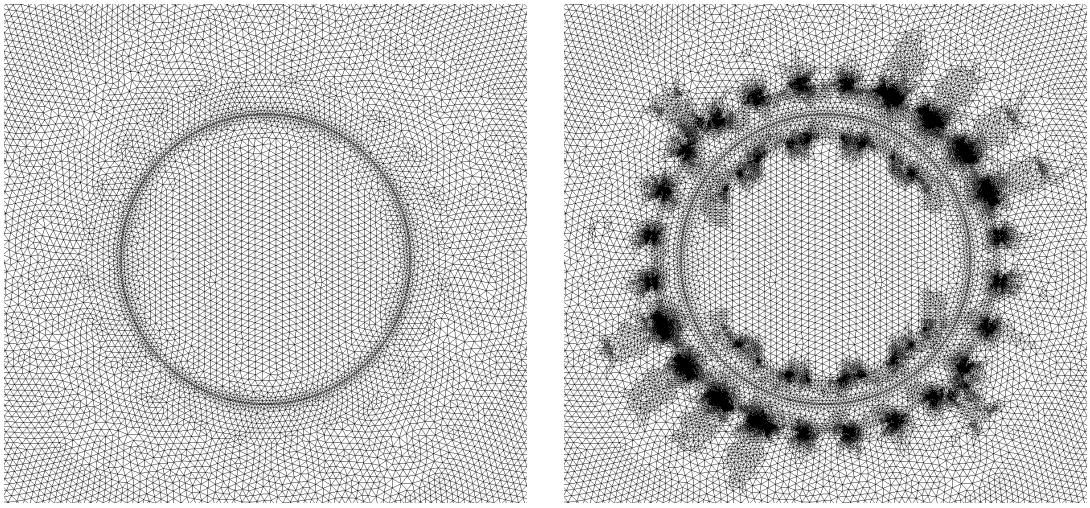


Figure 5.8: Left: Initial mesh with 13485 dofs. Right: Adaptive refined mesh on level $L = 6$ with 53013 dofs obtained for a nonlinear material behaviour.

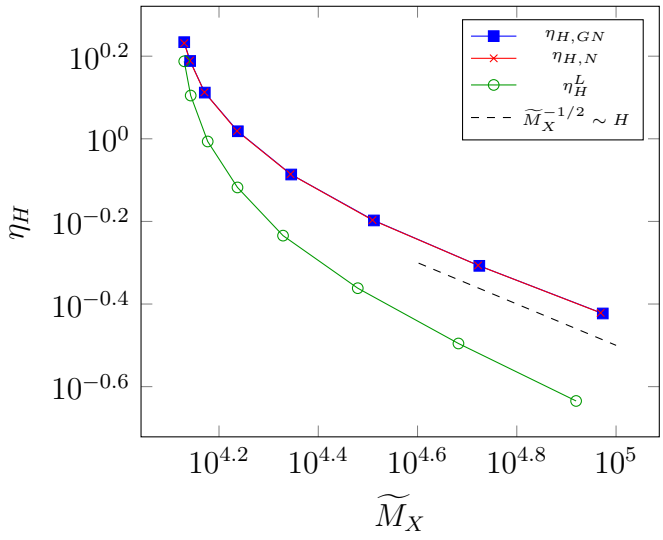


Figure 5.9: Convergence behaviour of the error estimator η_H during the adaptive refinement process in case of a nonlinear material behaviour.

L	$\widetilde{M_X}$	iter	err	τ
0	13485	86	9.794e-08	1.000e+00
1	13861	71	8.971e-08	1.000e+00
2	14836	18	9.969e-08	1.000e+00
3	17294	19	7.874e-08	1.000e+00
4	22153	30	9.462e-08	1.000e+00
5	32513	24	5.611e-08	1.000e+00
6	53013	20	6.196e-08	1.000e+00
7	93999	12	7.786e-08	1.000e+00

Table 5.2: Information of the Gauß-Newton solver during the adaptive refinement process.

adaptive refinement process. We see that after the first two refinement levels the number of iterations does not exceed 30 and that the step size in the final iteration on each level is 1.

In a next step we also applied Newton's method, see Algorithm 3.30 in order to solve (5.16) via a minimal residual approach. However, we remark that the application of this method is a bit speculative since we do not know any properties of the second order derivative of the nonlinear operator B. We use a backtracking line search strategy and an error tolerance of 10^{-10}. The second order derivative is computed via the automatic differentiability capabilities of `Netgen/NGSolve`. The convergence behaviour of the error estimator η_H can be seen in Fig. 5.9. We see linear convergence as in the case of Gauß-Newton's method. In Tab. 5.3 we provide some information of the Newton solver during the adaptive refinement process. On the first two refinement levels the Newton solver got stuck as it could not achieve a descent of the objective until the step size was below 10^{-11}. Hence, we broke up the algorithm. Interestingly, the value of the error estimator is nearly the same as in Gauß-Newton's method even though the method failed to converge. For the other refinement levels Newton's method converged in at most 15 iterations, which is faster than in Gauß-Newton's method.

The results obtained so far are not fully satisfactory. On the one hand we have Gauß-Newton's method which converges, but takes a lot of iterations. On the other hand if we try to accelerate the convergence with Newton's method we cannot ensure convergence in every iteration. The idea is now to use a different operator A in Gauß-Newton's method, which leads to a better performance of the method. From the abstract theory in Chapter 3 we know that we have to ensure that A is bounded, self-adjoint and elliptic. Apart from that there are no restrictions on A in the least-squares

L	$\widetilde{M_X}$	iter	err	τ
0	13485	28	1.410e+00	7.276e-12
1	13845	24	4.266e+00	7.276e-12
2	14815	8	7.333e-14	1.000e+00
3	17232	8	7.225e-14	5.000e-01
4	22083	15	1.970e-13	1.000e+00
5	32300	10	1.119e-13	1.000e+00
6	52722	8	5.212e-11	1.000e+00
7	93083	8	3.379e-13	1.000e+00

Table 5.3: Information of the Newton solver during the adaptive refinement process.

approach. We have seen in Lemma 5.5 that the operator $B'(u) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is uniformly bounded and elliptic. Furthermore, it is self-adjoint since the Jacobian $JT(x, W)$ is symmetric. Thus, in the following we apply Gauß-Newton's method with the matrices

$$\begin{aligned} A_h(\underline{u})[i,j] &= \int_\Omega JT(x, \nabla u_H)\nabla\psi_j \cdot \nabla\psi_i \,\mathrm{d}x \quad i,j = 1,...,M_Y, \\ B_h'(\underline{u})[i,j] &= \int_\Omega JT(x, \nabla u_H)\nabla\varphi_j \cdot \nabla\psi_i \,\mathrm{d}x \quad i = 1,...,M_Y,\ j = 1,...,M_X \end{aligned} \tag{5.27}$$

We use a backtracking line search startegy and for the stopping criterion we consider an error tolerance of 10^{-10}. We start with the zero initial guess on level $L = 0$, i.e., $u_H^{0,0} = 0$ and for $L > 0$ we use $u_H^{L,0} = P_H^{L,L-1} u_H^{L-1}$. For the global error estimator we apply on each refinement level L

$$\left(\eta_H^L\right)^2 = \int_\Omega JT(x, \nabla u_H^L)\nabla p_h^L \cdot \nabla p_h^L \,\mathrm{d}x = \sum_{\tau \in \mathcal{T}_H} \left(\eta_\tau^L\right)^2,$$

with the local indicators

$$\left(\eta_\tau^L\right)^2 = \int_\tau JT(x, \nabla u_H^L)\nabla p_h^L \cdot \nabla p_h^L \,\mathrm{d}x \quad \text{for } \tau \in \mathcal{T}_H.$$

The numerical results for the estimator η_H^L can be seen in Fig. 5.9 and detailed information about the solver is provided in Tab. 5.4. The convergence of the error estimator η_H^L is at the beginning faster than the convergence of the error indicator η_H from Gauß-Newton's and Newton's method with the matrix A_h defined in (5.25). For higher refinement levels it attemps to approach linear convergence. Further the number of iterations in each step reduces drastically. It takes 6 to 7 iterations until the solver is converged except for the first refinement level where it takes 13 iterations. This can be explained with the initial guess which is for level $L > 0$ much better than for $L = 0$.

L	$\widetilde{M_X}$	iter	err	τ
0	13485	13	3.455e-11	1.000e+00
1	13896	6	4.314e-11	1.000e+00
2	15037	6	6.771e-14	1.000e+00
3	17265	6	7.509e-13	1.000e+00
4	21327	7	7.933e-14	5.000e-01
5	30165	7	2.190e-11	1.000e+00
6	48123	7	1.359e-13	1.000e+00
7	83065	7	4.013e-13	1.000e+00

Table 5.4: Information of the Gauss-Newton solver with the matrices (5.27) during the adaptive refinement process.

5.3.2 Magnetostatic problem on a moving domain

In the previous section we considered the magnetostatic regime for a fixed rotor position. However, in practice one usually performs a sequence of magnetostatic computations for different rotor positions, see [76], as one is interested in e.g., the average torque for one round. In literature there exists different stator/rotor coupling techniques to incorporate the motion of the rotor, see e.g., the lock-step method [63, 138], the moving band/sliding surface method [144, 169], the Lagrange multiplier method [107, 140], Nitsche-type mortar methods [30] or DG methods [5]. We will go for a different approach. We consider a space-time setting, where the movement of the rotor is resolved within the mesh, see e.g., [40, 77, 78, 80]. This is possibly because in the magnetostatic approximation it is assumed that the rotor moves at a constant speed and hence the motion is a priori known. The impressed current density $f = f(y_1, y_2, t)$ is now a function dependent on time, which is given via a three-phase alternating current. Hence, also the third component of the magnetic vector potential $u = u(y_1, y_2, t)$ is time dependent and (5.9) boils down to

$$-\operatorname{div}_y(\nu(y, |\nabla_y u(y,t)|)\nabla_y u(y,t)) = f(y,t) \quad \text{for } (y,t) \in Q, \tag{5.28a}$$
$$u(y,t) = 0 \quad \text{for } (y,t) \in \Sigma, \tag{5.28b}$$
$$[\![u(y,t)]\!] = 0 \quad \text{for } (y,t) \in \Gamma_I(t) \times (0,T), \tag{5.28c}$$
$$[\![\nu(y, |\nabla_y u(y,t)|)\nabla_y u(y,t) \cdot n(y,t)]\!] = 0 \quad \text{for } (y,t) \in \Gamma_I(t) \times (0,T). \tag{5.28d}$$

The reference configuration Ω is depicted in Fig. 5.10. Instead of an axially-layered rotor as in Fig. 5.2 we consider a soild salient design of the rotor including a hole for the shaft. Furthermore, we omit an additional air layer around the stator. The domain Ω allows the representation

$$\Omega = \Omega_s \cup \Omega_r \cup \Omega_a,$$

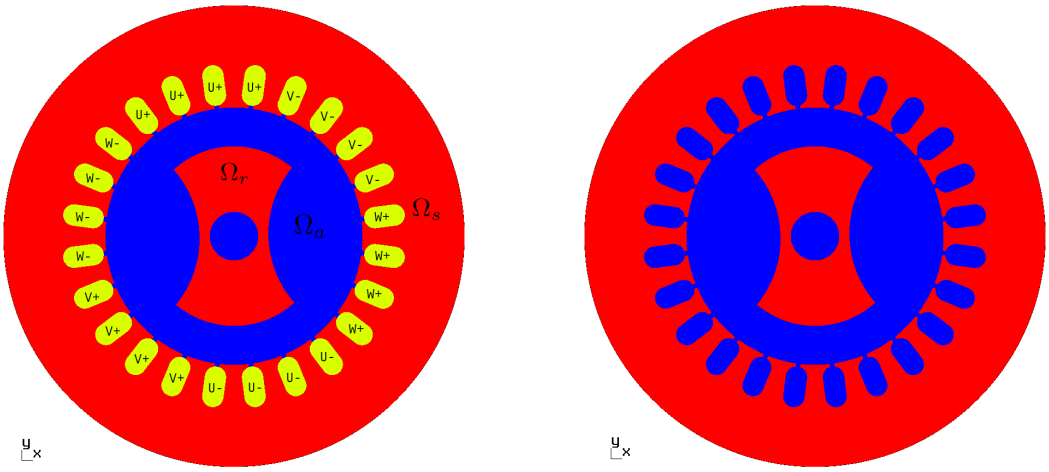


Figure 5.10: Left: Computational domain Ω indicating different materials and subdomains of the machine. Red: iron, blue: air, yellow: copper. Right: Domain Ω divided into ferromagnetic material Ω_f (red) and nonferromagnetic material $\Omega \setminus \Omega_f$ (blue).

where Ω_s denotes the stator including the coils, Ω_r the solid rotor and Ω_a the air region. The stator as well as the air domain are fixed while the rotor domain is rotationally moving over time. Therefore, the deformation to describe the evolution $\Omega(t)$ of Ω is given as

$$\begin{aligned} y(t) = \varphi(x,t) &= \begin{cases} R(\alpha(t))x & \text{for } x \in \Omega_r, \\ x & \text{for } x \in \Omega \setminus \Omega_r \end{cases}, \\ R(\alpha(t)) &= \begin{bmatrix} \cos(\alpha(t)) & -\sin(\alpha(t)) \\ \sin(\alpha(t)) & \cos(\alpha(t)) \end{bmatrix}. \end{aligned} \tag{5.29}$$

Here $\alpha(t)$ describes the rotation angle. We choose a 180-degree rotation within the final time horizon $T = 0.012\ s$, i.e.,

$$\alpha(t) = \pi \frac{t}{T}, \tag{5.30}$$

which corresponds to 2500 number of rounds per minute (nrpm). Using (5.29) we can define the space-time cylinder Q

$$Q := \{(y,t) \in \mathbb{R}^3 : \ y = \varphi(x,t), \ x \in \Omega, \ t \in (0,T)\}, \tag{5.31}$$

as well as the lateral boundary Σ

$$\Sigma := \{(y,t) \in \mathbb{R}^3 : \ y = \varphi(x,t), \ x \in \partial\Omega, \ t \in (0,T)\}. \tag{5.32}$$

We apply the minimal residual framework to (5.28) solely in a linear material setting, however we mention that an extension to the nonlinear case is possible, similar as it was done in Section 5.3.1. The linear material coefficient $v_l(y)$ reads

$$\nu_l(y,t) = \begin{cases} \nu_f & \text{for } (y,t) \in Q_f, \\ \nu_0 & \text{for } (y,t) \in Q \setminus Q_f \end{cases}, \tag{5.33}$$

where the space-time cylinder Q_f is defined as

$$Q_f := \{(y,t) \in \mathbb{R}^3 : \ y = \varphi(x,t), \ x \in \Omega_f, \ t \in (0,T)\} \tag{5.34}$$

and the parameters are the same as in (5.12). For the least-squares setting we have the spaces

$$X = Y = \{v \in L^2(Q) : \ \nabla_y v \in [L^2(Q)]^2, \ v = 0 \text{ on } \Sigma\},$$

together with the norm

$$\|v\|_X := \sqrt{\langle \nu_l \nabla_y v, \nabla_y v \rangle_{L^2(Q)}}.$$

The operators $A : Y \to Y^*, B : X \to Y^*$ are defined in the variational sense satisfying

$$\langle Ap, q \rangle_Q := \langle \nu_l \nabla_y p, \nabla_y q \rangle_{L^2(Q)}, \quad \langle Bu, q \rangle_Q := \langle \nu_l \nabla_y u, \nabla_y q \rangle_{L^2(Q)}$$

for all $u \in X$, $p, q \in Y$. It is clear that the operators A, B fulfill the assumptions of Section 3.1 with the constants $c_1^A = c_1^B = c_2^A = c_2^B = 1$. The abstract variational formulation (3.12) then reads to find $(u,p) \in X \times Y$ such that

$$\langle \nu_l \nabla_y p, \nabla_y q \rangle_{L^2(Q)} + \langle \nu_l \nabla_y u, \nabla_y q \rangle_{L^2(Q)} = \langle f, q \rangle_Q, \quad \langle \nu_l \nabla_y v, \nabla_y p \rangle_{L^2(Q)} = 0 \tag{5.35}$$

is satisfied for all $(v,q) \in X \times Y$. For the discretization we choose $X_H = S_H^1(\mathcal{T}_H) \cap X$ and $Y_h = Y_H = S_H^2(\mathcal{T}_H) \cap Y$, which are defined with respect to some admissible and locally quasi-uniform decomposition $\mathcal{T}_H$ of Q into shape regular simplicial finite elements. We have $X_H \subset Y_h$ and the discrete inf-sup condition

$$\|u_H\|_X = \frac{\langle \nu_l \nabla_y u_H, \nabla_y u_H \rangle_{L^2(Q)}}{\|u_H\|_Y} \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle Bu_H, q_h \rangle_Q}{\|q_h\|_Y},$$

i.e., (3.15). This ensures unique solvability of the mixed discrete system which reads to find $(u_H, p_h) \in X_H \times Y_h$ such that

$$\begin{aligned} \langle \nu_l \nabla_y p_h, \nabla_y q_h \rangle_{L^2(Q)} + \langle \nu_l \nabla_y u_H, \nabla_y q_h \rangle_{L^2(Q)} &= \langle f, q_h \rangle_Q, \\ \langle \nu_l \nabla_y v_H, \nabla_y p_h \rangle_{L^2(Q)} &= 0 \end{aligned} \tag{5.36}$$

holds for all $(v_H, q_h) \in X_H \times Y_h$. In our numerical simulations we use the global error estimator

$$\eta_H^2 = \|p_h\|_Y^2 = \langle \nu_l \nabla_y p_h, \nabla_y p_h \rangle_{L^2(Q)} = \sum_{\tau \in \mathcal{T}_H} \langle \nu_l \nabla_y p_h, \nabla_y p_h \rangle_{L^2(\tau)} = \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2, \quad (5.37)$$

with the local indicators

$$\eta_\tau^2 = \langle \nu_l \nabla_y p_h, \nabla_y p_h \rangle_{L^2(\tau)} \quad \text{for } \tau \in \mathcal{T}_H,$$

to drive an adaptive refinement scheme. We use the Dörfler criterion [58] as a marking strategy and refine the marked elements via newest vertex bisection. As mentioned the right-hand side f is given in terms of a three-phase alternating current with the winding scheme indicated in Fig. 5.10, i.e.,

$$\begin{aligned} f(y,t) = {} & j_U(t)\chi_{Q_{U_+}}(y,t) + j_V(t)\chi_{Q_{V_+}}(y,t) + j_W(t)\chi_{Q_{W_+}}(y,t) \\ & - j_U(t)\chi_{Q_{U_-}}(y,t) - j_V(t)\chi_{Q_{V_-}}(y,t) - j_W(t)\chi_{Q_{W_-}}(y,t), \end{aligned} \quad (5.38)$$

for $(y,t) \in Q$, where χ denotes the indicator function or characteristic function. The current densities are given by means of the functions

$$\begin{aligned} j_U(t) &= \frac{I \cdot N_w}{A_c} \sin\left(\alpha(t) + \frac{\pi}{4}\right), \\ j_V(t) &= \frac{I \cdot N_w}{A_c} \sin\left(\alpha(t) + \frac{\pi}{4} + \frac{4\pi}{3}\right), \\ j_W(t) &= \frac{I \cdot N_w}{A_c} \sin\left(\alpha(t) + \frac{\pi}{4} + \frac{2\pi}{3}\right), \end{aligned}$$

where $I = 12\ A$ denotes the amplitude of the impressed current, $N_w = 64$ is the number of turns, and $A_c = 3.2962 \cdot 10^{-5}\ m^2$ is the area of one coil. The implementation was done in `Netgen/NGSolve` [152], where we used the sparse direct solver `Pardiso` [149] to solve the resulting linear systems.

We started with an initial mesh consisting of 44459 dofs, where we inserted 12 time slices in order to have a sufficient resolution in the time direction at the beginning of the adaptive refinement process. The mesh between these time slices is completely unstructured, see Fig. 5.11. The adaptive refinement process generated a space-time mesh which obtains stronger refinements around the coils and in the airgap between stator and rotor, see Fig. 5.11 and Fig. 5.12, where we present a more detailed view on the mesh. The convergence behaviour of the error estimator until refinement level $L = 2$ is depicted in Fig. 5.13. We see that the rate attemps to approach linear convergence. We mention that the evaluation of η_H on level $L = 2$ is done on a mesh with $\dim(S_H^1(Q)) = 277754$ dofs and $\dim(S_H^1(Q) + S_H^2(Q)) = 2395988$ dofs for the corresponding saddle point system (5.36). A further refinement was not possible as

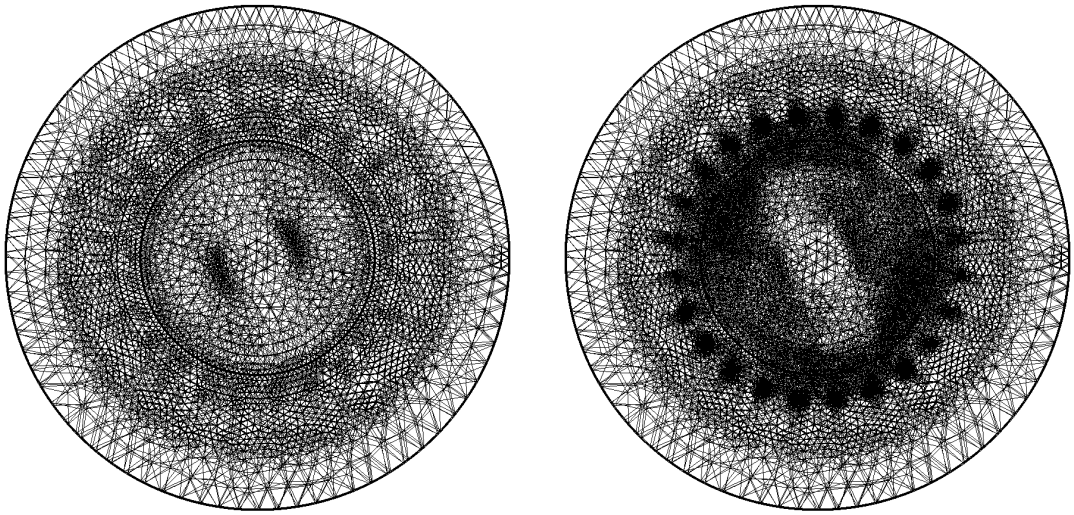


Figure 5.11: Left: Initial mesh with 44459 dofs. Right: Adaptive refined mesh on level $L = 2$ with 277754 dofs obtained for a linear material behaviour. Both meshes are cut at $t = 0.003$ s.

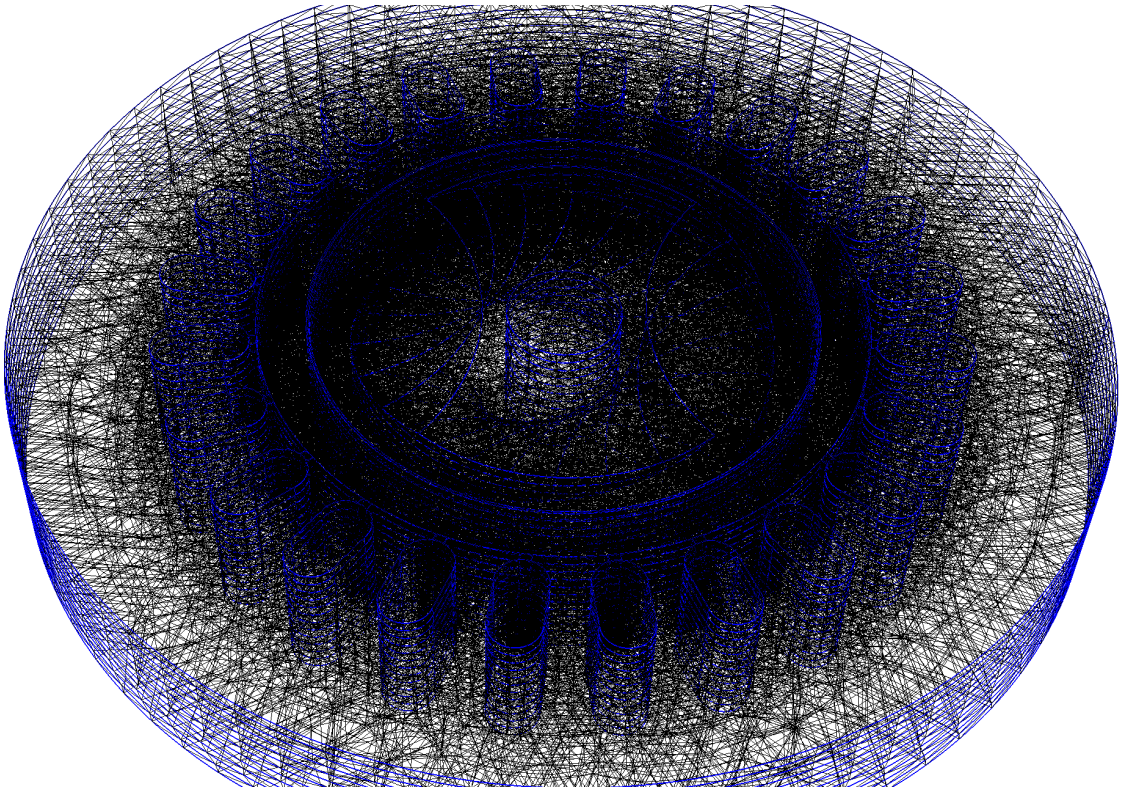


Figure 5.12: Detailed view of the generated adaptive space-time mesh. The blue lines indicate the layout of the motor and the timeslices.

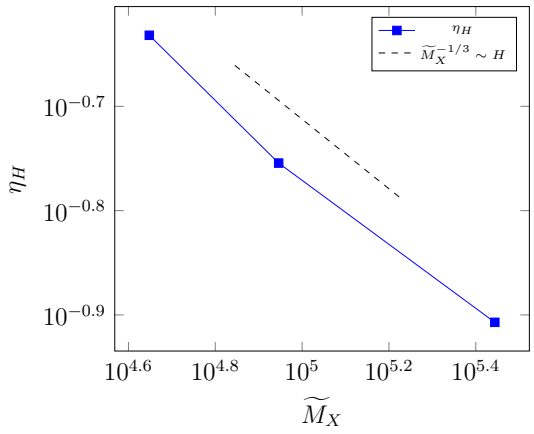


Figure 5.13: Convergence behaviour of the error estimator η_H during the adaptive refinement process in case of a linear material behaviour.

we reached the limits of the direct solver. In Fig. 5.14 we present the numerical solution u_H at different times in the space-time domain including the fieldlines. We see a stronger concentration of the fieldlines in the stator, where the poles join and also in the rotor around the shaft.

5.3.3 Eddy current approximation

In this section we consider the eddy current approximation of Maxwell's equations, i.e. (5.9), for the simulation of the synchronous reluctance machine given in Fig. 5.10. For this we use similar as in the previous section a space-time approach where we consider the movement of the rotor within the mesh. The space-time cylinder Q and its lateral boundary Σ can be described by means of the deformation mapping (5.29) and are given by (5.31), (5.32), respectively. In the eddy current model we also have to consider the velocity field induced by the deformation mapping (5.29). A straight forward calculation gives

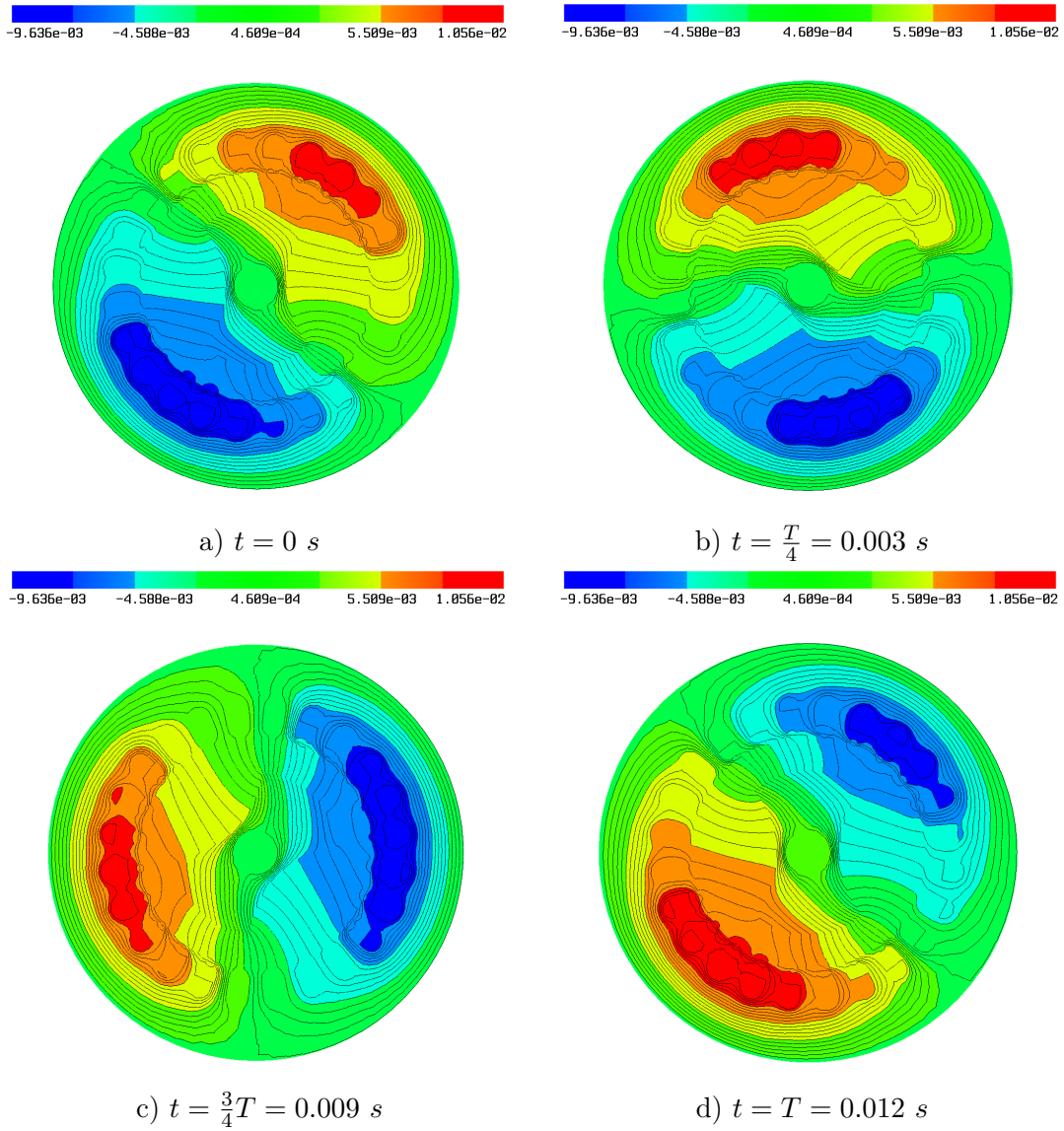
$$v(y,t) = y'(t) = \begin{cases} \alpha'(t)(-y_2, y_1)^\top & \text{for } (y,t) \in Q_r \\ 0 & \text{else} \end{cases},$$

where Q_r is the space-time cylinder of the rotor with its cross-section Ω_r for $t = 0$ indicated in Fig. 5.10. The derivative of the rotation angle $\alpha(t)$ can be computed from (5.30) and reads

$$\alpha'(t) = \frac{\pi}{T}.$$

Note that this velocity field is divergence free, i.e.,

$$\operatorname{div}_y v(y,t) = 0 \quad \text{for } (y,t) \in Q. \tag{5.39}$$



a) $t = 0\ s$ b) $t = \frac{T}{4} = 0.003\ s$

c) $t = \frac{3}{4}T = 0.009\ s$ d) $t = T = 0.012\ s$

Figure 5.14: Left: Numerical solution u_H obtained on refinement level $L = 2$ with 277754 dofs for different times t in the space-time cylinder.

The application of the least-squares framework to (5.9) is done for a linear material behaviour, where the linear reluctivity ν_l is given by (5.33). The values of the electric conductivity in the iron sheets and in the coils are usually negligible due to the laminated structure of the sheets and the insulation of the wires. However, we assume in our simulations a small value of σ in ferromagnetic material and in the coils of the machine and solely neglect the value of σ in air. Hence, we choose

$$\sigma(y,t) = \begin{cases} 0.001 & \text{for } (y,t) \in Q_f \cup Q_c \\ 0 & \text{else} \end{cases},$$

where Q_c denotes the space-time cylinder of the coils which are marked in yellow in Fig. 5.10 and Q_f is the space-time cylinder of the ferromagnetic material given by (5.34). In view of the abstract minimal residual framework we have the spaces

$$\begin{aligned} Y &:= \left\{ q \in L^2(Q) : \ \nabla_y q \in [L^2(Q)]^2, \ q = 0 \text{ on } \Sigma \right\}, \\ X &:= \left\{ u \in Y : \ \sigma[\partial_t u + v \cdot \nabla_y u] \in Y^*, \ u(x,0) = 0 \text{ for } x \in \Omega_f \cup \Omega_c \right\} \end{aligned}$$

together with the norms

$$\begin{aligned} \|q\|_Y^2 &:= \langle \nu_l \nabla_y q, \nabla_y q \rangle_{L^2(Q)}, \\ \|u\|_X^2 &:= \|u\|_Y^2 + \|\sigma[\partial_t u + v \cdot \nabla_y u]\|_{Y^*}^2 = \|u\|_Y^2 + \|w_u\|_Y^2, \end{aligned}$$

where $w_u \in Y$ is the unique solution to the variational problem

$$\langle \nu_l \nabla_y w_u, \nabla_y q \rangle_{L^2(Q)} = \langle \sigma(\partial_t u + v \cdot \nabla_y u), q \rangle_Q$$

for all $q \in Y$. The operator $A : Y \to Y^*$ is defined as

$$\langle Ap, q \rangle_Q := \langle \nu_l \nabla_y p, \nabla_y q \rangle_{L^2(Q)}$$

for $p, q \in Y$, which is bounded, elliptic and self-adjoint. The operator $B : X \to Y^*$ reads

$$\langle Bu, q \rangle_Q := \langle \sigma(\partial_t u + v \cdot \nabla_y u), q \rangle_Q + \langle \nu_l \nabla_y u, \nabla_y q \rangle_{L^2(Q)}$$

for $u \in X$, $q \in Y$. The necessary conditions on the operator B, i.e., boundedness, injectivity and surjectivity are shown in [78, 83]. We briefly comment on these. Similar as in the case of a fixed domain Ω, see e.g. Lemma 4.9 we conclude that the operator B is bounded satisfying

$$\langle Bu, q \rangle_Q \leq \sqrt{2} \|u\|_X \|q\|_Y$$

for all $u \in X$, $q \in Y$. Furthermore, a careful look at the inf-sup proof of B in Lemma 4.9 reveals that it also carries over to the case of a moving domain $\Omega(t)$. The crucial part is to show the nonnegativity of the expression

$$\langle \sigma(\partial_t u + v \cdot \nabla_x u), u \rangle_Q = \int_0^T \int_{\Omega(t)} \sigma(\partial_t u + v \cdot \nabla_y u) u \, \mathrm{d}y \mathrm{d}t$$

Using the total time derivative we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}u(y,t) = \partial_t u(y,t) + v(y,t)\cdot\nabla_y u(y,t). \tag{5.40}$$

Moreover, an application of Reynold's transport theorem [61] and (5.39) gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)} u(y,t)\mathrm{d}y &= \int_{\Omega(t)} \left[\partial_t u(y,t) + \mathrm{div}_y(u(y,t)v(y,t))\right]\mathrm{d}y \\ &= \int_{\Omega(t)} \left[\partial_t u(y,t) + \nabla_y u(y,t)\cdot v(y,t)\right]\mathrm{d}y \\ &= \int_{\Omega(t)} \frac{\mathrm{d}}{\mathrm{d}t}u(y,t)\,\mathrm{d}y\end{aligned} \tag{5.41}$$

With (5.40) and (5.41) we conclude

$$\begin{aligned}\langle\sigma(\partial_t u + v\cdot\nabla_y u), u\rangle_Q &= \int_0^T\int_{\Omega(t)} \sigma\frac{\mathrm{d}}{\mathrm{d}t}uu\,\mathrm{d}y\mathrm{d}t = \frac{1}{2}\int_0^T\int_{\Omega(t)}\frac{\mathrm{d}}{\mathrm{d}t}\sigma[u]^2\,\mathrm{d}y\mathrm{d}t \\ &= \frac{1}{2}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)}\sigma u^2\,\mathrm{d}y\mathrm{d}t \\ &= \frac{1}{2}\left(\int_{\Omega(T)}\sigma u^2(y,T)\,\mathrm{d}y - \int_\Omega \sigma u^2(x,0)\,\mathrm{d}x\right) \\ &= \frac{1}{2}\int_{\Omega(T)}\sigma u^2(y,T)\,\mathrm{d}y \geq 0,\end{aligned} \tag{5.42}$$

where we use that the integral over the domain Ω vanishes since $u(x,0) = 0$ for $x \in \Omega_f \cup \Omega_c$ and $\sigma = 0$ for $x \in \Omega_a$. The property (5.42) ensures that the operator B is inf-sup stable satisfying

$$\|u\|_X \leq \sup_{0\neq q\in Y}\frac{\langle Bu, q\rangle_Q}{\|q\|_Y} \tag{5.43}$$

for all $u \in X$. The proof of the surjectivity is more involved. We refer to [78, 83]. The abstract variational formulation (3.12) in this case reads to find $(u,p) \in X \times Y$ such that

$$\begin{aligned}\langle\nu_l\nabla_y p, \nabla_y q\rangle_{L^2(Q)} + \langle\sigma(\partial_t u + v\cdot\nabla_y u), q\rangle_Q + \langle\nu_l\nabla_y u, \nabla_y q\rangle_{L^2(Q)} &= \langle f, q\rangle_Q, \\ \langle\sigma(\partial_t z + v\cdot\nabla_y z), p\rangle_Q + \langle\nu_l\nabla_y z, \nabla_y p\rangle_{L^2(Q)} &= 0\end{aligned} \tag{5.44}$$

is satisfied for all $(z,q) \in X\times Y$. For the discretization we consider $X_H = S^1_H(\mathcal{T}_H)\cap X$ and $Y_h = Y_H = S^2_H(\mathcal{T}_H)\cap Y$, which are defined with respect to some admissible and locally quasi-uniform decomposition $\mathcal{T}_H$ of Q into shape regular simplicial finite

elements. The discrete variational formulation of (5.44) is to find $(u_H, p_h) \in X_H \times Y_h$ such that

$$\begin{aligned} \langle \nu_l \nabla_y p_h, \nabla_y q_h \rangle_{L^2(Q)} + \langle \sigma(\partial_t u_H + v \cdot \nabla_y u_H), q_h \rangle_Q + \langle \nu_l \nabla_y u_H, \nabla_y q_h \rangle_{L^2(Q)} &= \langle f, q_h \rangle_Q, \\ \langle \sigma(\partial_t z_H + v \cdot \nabla_y z_H), p_h \rangle_Q + \langle \nu_l \nabla_y z_H, \nabla_y p_h \rangle_{L^2(Q)} &= 0 \end{aligned} \tag{5.45}$$

holds for all $(z_H, q_h) \in X_H \times Y_h$. To ensure unique solvability of (5.45) we can proceed as in the case of a fixed domain. This means we introduce the mesh dependent norm

$$\|u\|_{X,h} := \sqrt{\|u\|_Y^2 + \|w_{uh}\|_Y^2} \leq \|u\|_X,$$

where $w_{uh} \in Y$ is the unique solution to the variational problem

$$\langle \nu_l \nabla_y w_{uh}, \nabla_y q_h \rangle_{L^2(Q)} = \langle \partial_t u + v \cdot \nabla_y u, q_h \rangle_Q \quad \text{for all } q_h \in Y_h. \tag{5.46}$$

Now, the proof of a discrete inf-sup condition can be shown similar as in the continuous case for a moving domain. In particular, we define the test function $\bar{q}_h := u_H + w_{u_H h} \in Y_h$, where $w_{u_H h} \in Y_h$ solves (5.46). Then it follows

$$\langle B u_H, \bar{q}_h \rangle_Q = \|\bar{q}_h\|_Y^2, \quad \|\bar{q}_h\|_Y^2 \geq \|u_H\|_{X,h}^2,$$

and we conclude the discrete inf-sup stability condition

$$\|u_H\|_{X,h} \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle B u_H, q_h \rangle_Q}{\|q_h\|_Y} \quad \text{for all } u_H \in X_H, \tag{5.47}$$

which gives unique solvability of the mixed problem (5.45).

In our numerical simulations we use the error estimator (5.37) together with the Dörfler criterion [58] with parameter $\theta = 0.5$ to drive an adaptive refinement scheme. The marked elements are then refined using newest vertex bisection. For the excitation f of the electric machine we choose the same three-phase alternating current as described in (5.38). The implementation was done in `Netgen/NGSolve` [152], where we used the sparse direct solver `Pardiso` [149].

We started the adaptive refinement process with the same initial mesh consisting of 44459 dofs and 12 timeslices as in the quasistatic case, see Fig. 5.15. After three iterations the mesh obtains stronger refinements around the coils and in the airgap between stator and rotor, see Fig. 5.15. A detailed view of the generated mesh is provided in 5.16. Here we indicated the layout of the motor as well as the time slices in blue. We see a completely unstructured mesh with respect to space and time between the time slices. In Fig. 5.17 we present the convergence behaviour of the error estimator η_H. We observe a drop in the rate from level $L = 1$ to $L = 2$,

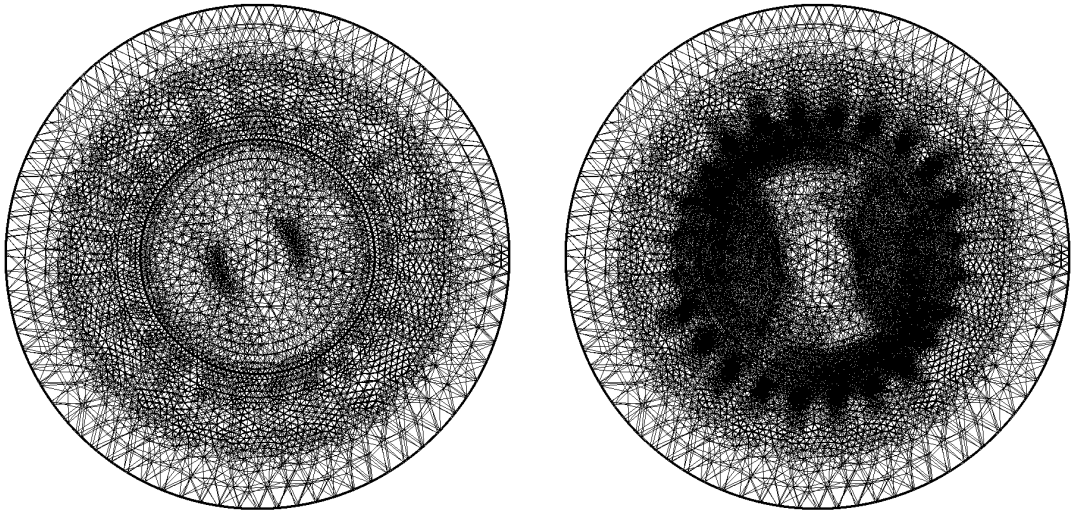


Figure 5.15: Cross-section of the spac-time mesh for $t = 0.003\ s$. Left: Initial mesh with 44459 dofs. Right: Adaptive refined mesh on level $L = 2$ with 221495 dofs obtained for a linear material behaviour.

but still have an error reduction. To make a clear statement about the rate one needs to compute some further iterations, which was not possible as we ran into the limits of the direct solver. Finally, in Fig. 5.18 we depict the numerical solution of the third component of the magnetic vector potential u_H for homogeneous initial conditions together with the corresponding fieldlines. Here we plot the solution on the cross-section of the electric motor at specific time points t. For the time points $t \neq 0$ we observe a stronger concentration of the magnetic fieldlines in the stator of the machine as well as in the rotor around the shaft.

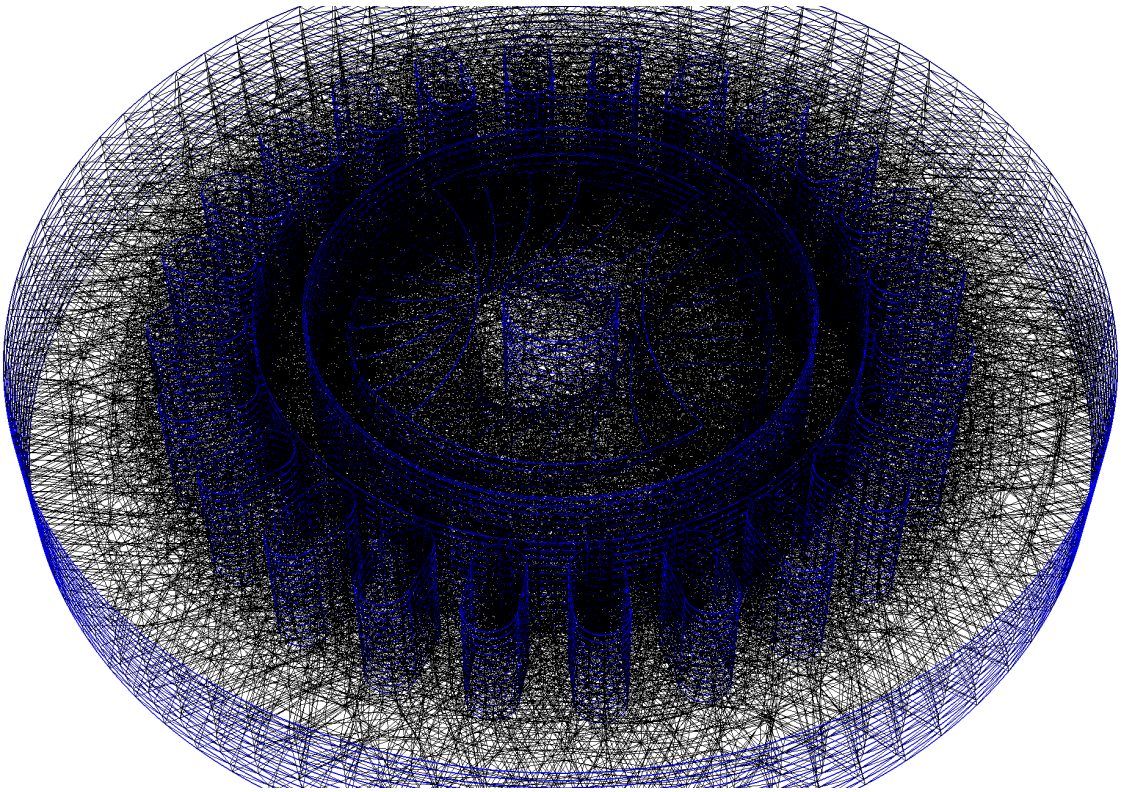


Figure 5.16: Detailed view of the generated adaptive space-time mesh for the eddy current problem. The blue lines indicate the layout of the motor and the timeslices.

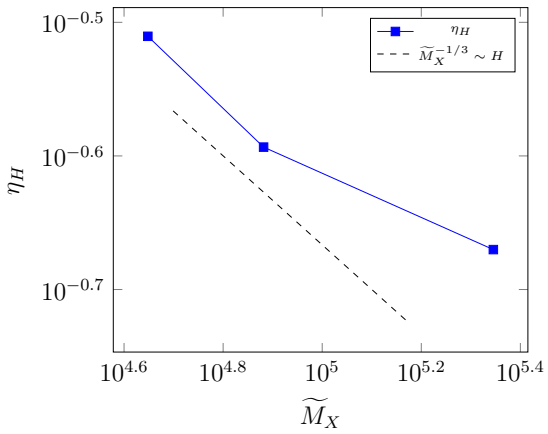
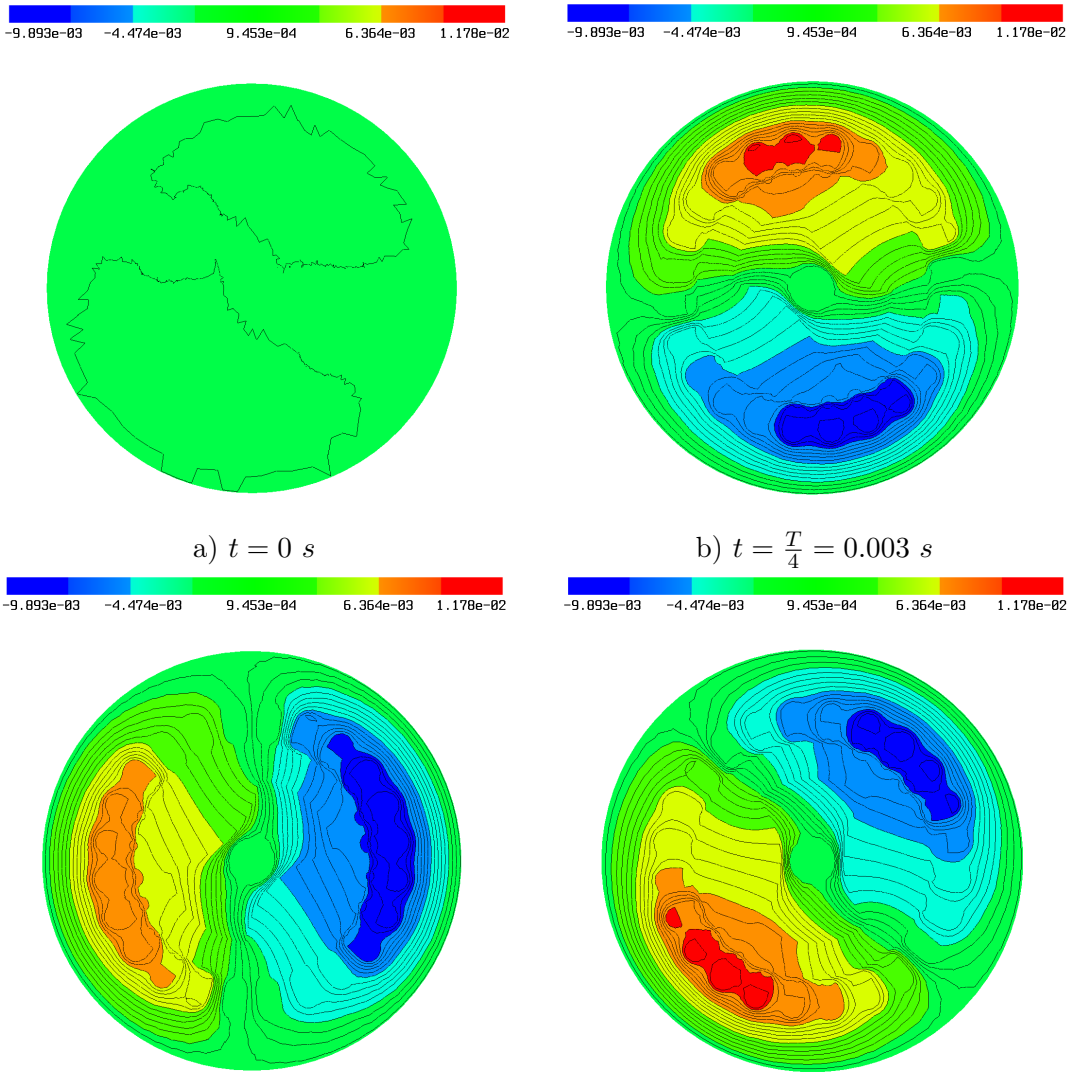


Figure 5.17: Convergence behaviour of the error estimator η_H during the adaptive refinement process in case of a linear material behaviour.



a) $t = 0\ s$ b) $t = \frac{T}{4} = 0.003\ s$

c) $t = \frac{3}{4}T = 0.009\ s$ d) $t = T = 0.012\ s$

Figure 5.18: Left: Numerical solution u_H obtained on refinement level $L = 2$ with 221495 dofs for different times t in the space-time cylinder.

6 Conclusions & Outlook

In this thesis we solved an abstract operator equation $Bu = f$ by means of a least-squares/minimal residual approach. In particular, we put a stong emphasis on the combination of a least-squares method with a space-time discretization scheme in the sense of [157] and on the simulation of electric machines. To take into account the different types of PDEs, i.e., elliptic, parabolic, as well as linear and nonlinear we presented an abstract least-squares framework for the solution of $Bu = f$. Assuming that $B : X \to Y^*$ is an isomorphism and using the Riesz operator $A : Y \to Y^*$ the first-order optimaliy system was given by means of a saddle point system

$$\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} \begin{bmatrix} p \\ u \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}. \tag{6.1}$$

We showed a complete stability and a priori error analysis of the mixed system (6.1). Furthermore, we proved that the inbuilt error estimator, i.e., p_h is efficient and reliable modulo some data oscillation term. Additionally, assuming a saturation assumption an alternative proof of the reliability of the error indicator was provided. At several points we also highlighted the connection to the DPG framework presented in [50, 52, 55]. In addition to that we also discussed an extension to the nonlinear case with possible solution strategies involving Newton's and Gauß-Newton's method.

Next, we demonstrated the application of the abstract least-squares framework to various parabolic evolution equations, involving the heat equation, the convection-diffusion equation and a heat equation with nonlinear reaction term. For the heat equation we were able to show a discrete inf-sup stability condition with respect to a mesh dependent norm and provided numerical examples which confirmed our theoretical findings as well as a comparison to the FOSLS method by Führer and Karkulik [72]. In a next step we applied the least-squares method to the solution of a time dependent convection-diffusion equation. Stability of the discrete system could be shown with similar techniques as in the case of the heat equation. Numerical examples including spatial and temporal boundary layers were given. Last but not least a semilinear problem was considered and a comparison between the Newton and the Gauß-Newton method was given. Both methods lead to the same rates for erros and estimators. However, the Newton method needed fewer iterations than the Gauß-Newton method.

Finally, we applied the least-squares method to the simulation of the electromagnetic quantities in an electric machine. Since electric machines are low frequency appli-

cations the eddy current approximation of Maxwell's equations served as a physical model. In a first step we considered the 2d magnetostatic approximation, i.e., all field quantities are independent of time. This is one of the most widely used model to simulate electric machines. We ended up with a quasilinear elliptic PDE. The analysis was carried out using the theory of monotone operators. This enabled us to prove the well-posedness of the resulting linear systems in Gauß-Newton's method as well as to show that the computed search direction is a descent direction. Numerical results computed on a synchronous reluctance machine lead to an adaptive mesh which obtained stronger refinements at the interfaces between ferromagnetic and nonferromagnetic materials. In a second step and third step we considered the quasistatic and the eddy current approximation. In both models the movement of the rotor was considered within the mesh. Numerical results were presented which demonstrated the applicability of the least-squares framework.

Open questions and future work

The work presented in this thesis can be extended into several directions:

The development of preconditioners which can be used in robust iterative solution strategies to solve the discrete system resulting from (6.1) is an open task. In particular, this involves preconditioners for A and the Schur complement $S = B^*A^{-1}B$. For time dependent PDEs these preconditioners have to take care of the time related anisotropy which is built-in in the considered space-time formulation. As mentionend in [172] a possibility would be to explore specialized directional smoothers and appropriate semi-coarsening in a geometric multigrid preconditioner. The resulting finite element matrices A_h and S_h are symmetric and positive definite. Thus, possible iterative solution strategies are a conjugate gradient (CG) method [92] for the solution of the discrete Schur complement system or the minimal residual (MINRES) method [135] for the solution of the Galerkin discretization of the mixed system (6.1), which is symmetric but indefinite. In a further step one could exploit parallelization and domain decomposition methods.

A further direction would be to apply the least-squares approach in a 3d-1d space-time setting, i.e., to consider a four-dimensional space-time cylinder Q. This involves the use of adaptive mesh refinement routines for the 4d case, see e.g. [147, 160, 131].

We only considered a magnetic field computation for the simulation of an electric machine. However, electric machines are multiphysical objects. Thus, in a further step one could also consider a more enriched physical model of the machine which involves the coupling to other physical fields. In particular, the couplings to physical models resulting from thermodynamics, structural mechanics or acoustics are of interest.

References

[1] R. Adams and J. Fournier, *Sobolev Spaces*, Pure and Applied Mathematics, Academic Press, 2003.

[2] J. H. Adler, T. A. Manteuffel, S. F. McCormick, J. W. Nolting, J. W. Ruge, and L. Tang, *Efficiency Based Adaptive Local Refinement for First-Order System Least-squares Formulations*, SIAM J. Sci. Comput., 33 (2011), pp. 1–24.

[3] J. H. Adler and P. S. Vassilevski, *Error Analysis for Constrained First-Order System Least-Squares Finite-Element Methods*, SIAM J. Sci. Comput., 36 (2014), pp. A1071–A1088.

[4] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure and Applied Mathematics, John Wiley & Sons, New York, 2000.

[5] P. Alotto, A. Bertoni, I. Perugia, and D. Schötzau, *Discontinuous finite element methods for the simulation of rotating electrical machines*, COMPEL - The international journal for computation and mathematics in electrical and electronic engineering, 20 (2001), pp. 448–462.

[6] R. Andreev, *Stability of sparse space-time finite element discretizations of linear parabolic evolution equations*, IMA J. Numer. Anal., 33 (2013), pp. 242–260.

[7] R. Araya and P. Venegas, *An a posteriori error estimator for an unsteady advection–diffusion–reaction problem*, Computers and Mathematics with Applications, 66 (2014), pp. 2456–2476.

[8] W. Arendt and M. Kreuter, *Mapping theorems for Sobolev spaces of vector-valued functions*, Studia Math., 240 (2018), pp. 275–299.

[9] D. N. Arnold, A. Mukherjee, and L. Pouly, *Locally Adapted Tetrahedral Meshes using Bisection*, SIAM J. Sci. Comput., 22 (2000), pp. 431–448.

[10] M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke, and R. Umla, *An assessment of discretizations for convection-dominated convection-diffusion equations*, Comput. Methods Appl. Mech. Engrg., 200 (2011), pp. 3395–3409.

[11] I. Babuška and A. Miller, *A feedback finite element method with a posteriori error estimation: Part I. The finite element method and some basic properties of the a posteriori error estimator*, Computer Methods in Applied Mechanics and Engineering, 61 (1987), pp. 1–40.

[12] C. Bacuta, D. Hayes, and T. O'Grady, *Saddle point least squares discretization for convection-diffusion*, Appl. Anal., 103 (2024), pp. 2241–2268.

[13] W. Bangerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2003.

[14] R. E. Bank and R. K. Smith, *A Posteriori Error Estimates Based on Hierarchical Bases*, SIAM Journal on Numerical Analysis, 30 (1993), pp. 921–935.

[15] S. Bartels, *Numerical Approximation of Partial Differential Equations*, vol. 64 of Texts in Applied Mathematics, Springer, 2016.

[16] N. Beranek, M. A. Reinhold, and K. Urban, *A space-time variational method for optimal control problems: well-posedness, stability and numerical solution*, Comput. Optim. Appl., 86 (2023), pp. 767–794.

[17] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Local error estimates and adaptive refinement for first-order system least squares (FOSLS)*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43.

[18] G. Bertotti, *Hysteresis in Magnetism: For Physicists, Materials Scientists, and Engineers*, Academic Press Series in Electromagnetism, Elsevier Science, 1998.

[19] A. Binder, *Elektrische Maschinen und Antriebe: Grundlagen, Betriebsverhalten*, Springer Berlin Heidelberg, 2017.

[20] P. Binev, W. Dahmen, and R. DeVore, *Adaptive Finite Element Methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.

[21] P. B. Bochev and M. D. Gunzburger, *Least-squares Finite Element Methods*, vol. 166 of Applied Mathematical Sciences, Springer, New York, 2009.

[22] P. B. Bochev, M. D. Gunzburger, and J. N. Shadid, *Stability of the SUPG finite element method for transient advection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2301–2323.

[23] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, vol. 44 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2013.

[24] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, vol. 15 of Texts in Applied Mathematics, Springer-Verlag, New York, 1994.

[25] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer, New York, 2011.

[26] M. O. Bristeau, O. Pironneau, R. Glowinski, J. Périaux, and P. Perrier, *On the numerical solution of nonlinear problems in fluid dynamics by least squares and finite element methods. I: Least square formulations and conjugate gradient solution of the continuous problems*, Comput. Methods Appl. Mech. Engrg., 17-18 (1979), pp. 619–657.

[27] M. O. Bristeau, O. Pironneau, R. Glowinski, J. Périaux, P. Perrier, and G. Poirier, *On the numerical solution of nonlinear problems in fluid dynamics by least squares and finite element methods. II: Application to transonic flow simulations*, Comput. Methods Appl. Mech. Engrg., 51 (1985), pp. 363–394.

[28] D. Broersen and R. Stevenson, *A robust Petrov-Galerkin discretisation of convection-diffusion equations*, Comput. Math. Appl., 68 (2014), pp. 1605–1618.

[29] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[30] A. Buffa, Y. Maday, and F. Rapetti, *A Sliding Mesh-Mortar Method for a two Dimensional Currents Model of Electric Engines*, ESAIM: M2AN, 35 (2001), pp. 191–228.

[31] T. Bui-Thanh and O. Ghattas, *A PDE-constrained optimization approach to the discontinuous Petrov-Galerkin method with a trust region inexact Newton-CG solver*, Comput. Methods Appl. Mech. Engrg., 278 (2014), pp. 20–40.

[32] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-Order System Least Squares for Second-Order Partial Differential Equations: Part I*, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[33] A. Cangiani, E. H. Georgoulis, and S. Metcalfe, *Adaptive discontinuous Galerkin methods for nonstationary convection-diffusion problems*, IMA Journal of Numerical Analysis, 34 (2014), pp. 1578–1597.

[34] C. Carstensen, *An adaptive mesh-refining algorithm allowing for an H^1 stable L^2 projection onto Courant finite element spaces*, Constr. Approx., 20 (2004), pp. 549–564.

[35] C. Carstensen, P. Bringmann, F. Hellwig, and P. Wriggers, *Nonlinear discontinuous Petrov-Galerkin methods*, Numer. Math., 139 (2018), pp. 529–561.

[36] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan, *A posteriori error control for DPG methods*, SIAM J. Numer. Anal., 52 (2014), pp. 1335–1353.

[37] C. Carstensen, M. Feischl, M. Page, and D. Praetorius, *Axioms of adaptivity*, Comput. Math. Appl., 67 (2014), pp. 1195–1253.

[38] C. Carstensen, D. Gallistl, and J. Gedicke, *Justification of the saturation assumption*, Numer. Math., 134 (2016), pp. 1–25.

[39] E. Casas, C. Ryll, and F. Tröltzsch, *Sparse Optimal Control of the Schlögl and FitzHugh-Nagumo Systems*, Comput. Methods Appl. Math., 13 (2013), pp. 415–442.

[40] A. Cesarano, C. Dapogny, and P. Gangl, *Space-time shape optimization of rotating electric machines*, Math. Models Methods Appl. Sci., 34 (2024), pp. 2647–2708.

[41] J. Chan, L. Demkowicz, and R. Moser, *A DPG method for steady viscous compressible flow*, Comput. & Fluids, 98 (2014), pp. 69–90.

[42] J. Chan and J. A. Evans, *A Minimum-Residual Finite Element Method for the Convection-Diffusion Equation*, Tech. Report 13-12, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, 2013.

[43] J. Chan, N. Heuer, T. Bui-Thanh, and L. Demkowicz, *A robust DPG method for convection-dominated diffusion problems II: Adjoint boundary conditions and mesh-dependent test norms*, Computers and Mathematics with Applications, 67 (2014), pp. 771–795.

[44] H. Chen, G. Fu, J. Li, and W. Qiu, *First order least squares method with weakly imposed boundary condition for convection dominated diffusion problems*, Comput. Math. Appl., 68 (2014), pp. 1635–1652.

[45] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol. 4 of Studies in Mathematics and its Applications, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978.

[46] A. Cohen, W. Dahmen, and G. Welper, *Adaptivity and variational stabilization for convection-diffusion equations*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 1247–1273.

[47] W. Dahmen, H. Monsuur, and R. Stevenson, *Least squares solvers for ill-posed PDEs that are conditionally stable*, ESAIM: Math. Model. Numer. Anal., 57 (2023), pp. 2227–2255.

[48] J. Daube, P. Nägele, and S. F., *Bochner-Räume.* `https://aam.uni-freiburg.de/agru/lehre/ws15/bochner/bochner.pdf`, 2016.

[49] R. Dautray, A. Craig, M. Artola, M. Cessenat, J. Lions, and H. Lanchon, *Mathematical Analysis and Numerical Methods for Science and Technology: Vol. 5: Evolution Problems I*, Mathematical Analysis and Numerical Methods for Science and Technology, Springer Berlin Heidelberg, 1999.

[50] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation*, Comput. Methods Appl. Mech. Engrg., 199 (2010), pp. 1558–1572.

[51] L. Demkowicz and J. Gopalakrishnan, *Analysis of the DPG Method for the Poisson Equation*, SIAM J. Numer. Anal., 49 (2011), pp. 1788–1809.

[52] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions*, Numer. Methods Partial Differential Equations, 27 (2011), pp. 70–105.

[53] L. Demkowicz and J. Gopalakrishnan, *A primal DPG method without a first-order reformulation*, Comput. Math. Appl., 66 (2013), pp. 1058–1064.

[54] L. Demkowicz and J. Gopalakrishnan, *The discontinuous Petrov-Galerkin method*, Acta Numerica, 34 (2025), pp. 293–384.

[55] L. Demkowicz, J. Gopalakrishnan, and A. H. Niemi, *A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity*, Appl. Numer. Math., 62 (2012), pp. 396–427.

[56] L. Demkowicz and N. Heuer, *Robust DPG Method for Convection-Dominated Diffusion Problems*, SIAM J. Numer. Anal., 51 (2013), pp. 2514–2537.

[57] P. Deuflhard, *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, vol. 35 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2011.

[58] W. Dörfler, *A Convergent Adaptive Algorithm for Poisson's Equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[59] W. Dörfler and R. H. Nochetto, *Small data oscillation implies the saturation assumption*, Numer. Math., 91 (2002), pp. 1–12.

[60] G. Dziuk, *Theorie und Numerik partieller Differentialgleichungen*, Walter de Gruyter GmbH & Co. KG, Berlin, 2010.

[61] C. Eck, H. Garcke, and P. Knabner, *Mathematische Modellierung*, Springer Berlin Heidelberg, 2017.

[62] T. Ellis, J. Chan, and L. Demkowicz, *Robust DPG Methods for Transient Convection-Diffusion*, in Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations, vol. 114 of Lect. Notes Comput. Sci. Eng., Springer, 2016, pp. 179–203.

[63] C. Emson, C. Riley, D. Walsh, K. Ueda, and T. Kumano, *Modelling eddy currents induced by rotating systems*, IEEE Transactions on Magnetics, 34 (1998), pp. 2593–2596.

[64] A. Ern and J. Guermond, *Theory and Practice of Finite Elements*, Applied Mathematical Sciences, Springer New York, 2004.

[65] A. Ern, I. Smears, and M. Vohralík, *Guaranteed, Locally Space-Time Efficient, and Polynomial-Degree Robust a Posteriori Error Estimates for High-Order Discretizations of Parabolic Problems*, SIAM J. Numer. Anal., 55 (2017), pp. 2811–2834.

[66] L. C. Evans, *Partial Differential Equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, 2010.

[67] S. Ferraz-Leite, C. Ortner, and D. Praetorius, *Convergence of simple adaptive Galerkin schemes based on h-h/2 error estimators*, Numer. Math., 116 (2010), pp. 291–316.

[68] M. Fortin, *An analysis of the convergence of mixed finite element methods*, RAIRO Anal. Numér., 11 (1977), pp. 341–354.

[69] F. N. Fritsch and R. E. Carlson, *Monotone Piecewise Cubic Interpolation*, SIAM J. Numer. Anal., 17 (1980), pp. 238–246.

[70] T. Führer, *On a mixed FEM and a FOSLS with H^{-1} loads*, Comput. Methods Appl. Math., 24 (2024), pp. 355–370.

[71] T. Führer, N. Heuer, and M. Karkulik, *MINRES for Second-Order PDEs with Singular Data*, SIAM J. Numer. Anal., 60 (2022), pp. 1111–1135.

[72] T. Führer and M. Karkulik, *Space-time least-squares finite elements for parabolic equations*, Comput. Math. Appl., 92 (2021), pp. 27–36.

[73] S. Funken, D. Praetorius, and P. Wissgott, *Efficient implementation of adaptive P1-FEM in Matlab*, Comput. Methods Appl. Math., 11 (2011), pp. 460–490.

[74] M. J. Gander, *50 Years of Time Parallel Time Integration*, in Multiple Shooting and Time Domain Decomposition Methods, T. Carraro et al, ed., vol. 9 of Contrib. Math. Comput. Sci., Springer, 2015, pp. 69–113.

[75] M. J. Gander and M. Neumüller, *Analysis of a New Space-Time Parallel Multigrid Algorithm for Parabolic Problems*, SIAM J. Sci. Comput., 38 (2016), pp. A2173–A2208.

[76] P. Gangl, *Sensitivity-Based Topology and Shape Optimization with Application to Electrical Machines*, PhD thesis, Johannes Kepler University Linz, 2016.

[77] P. Gangl, M. Gobrial, and O. Steinbach, *A Parallel Space-Time Finite Element Method for the Simulation of an Electric Motor*, in Domain Decomposition Methods in Science and Engineering XXVII, Z. Dostál et al, ed., Cham, 2024, Springer Nature Switzerland, pp. 255–262.

[78] P. Gangl, M. Gobrial, and O. Steinbach, *A Space-Time Finite Element Method for the Eddy Current Approximation of Rotating Electric Machines*, Computational Methods in Applied Mathematics, 25 (2025), pp. 441–457.

[79] P. Gangl, S. Köthe, C. Mellak, A. Cesarano, and A. Mütze, *Multi-objective free-form shape optimization of a synchronous reluctance machine*, COMPEL-The international journal for computation and mathematics in electrical and electronic engineering, 41 (2022), pp. 1849–1864.

[80] P. Gangl and N. Krenn, *Topology optimization of a rotating electric machine by the topological derivative*, PAMM, 23 (2023), p. e202200052.

[81] G. Gantner and R. Stevenson, *Further results on a space-time FOSLS formulation of parabolic PDEs*, ESAIM Math. Model. Numer. Anal., 55 (2021), pp. 283–299.

[82] G. Gantner and R. Stevenson, *Improved rates for a space-time FOSLS of parabolic PDEs*, Numerische Mathematik, (2023), pp. 1–25.

[83] M. Gobrial, *Space-time Finite Element Methods for the Eddy Current Problem and Applications*, PhD thesis, Graz University of Technology, 2025.

[84] J. Gopalakrishnan, *Five lectures on DPG methods*, 2014, `https://arxiv.org/abs/1306.0557`.

[85] J. Gopalakrishnan and W. Qiu, *An analysis of the practical DPG method*, Math. Comp., 83 (2014), pp. 537–552.

[86] C. Grossmann, H. Roos, and M. Stynes, *Numerical Treatment of Partial Differential Equations*, Springer Berlin Heidelberg, 2007.

[87] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, vol. 8 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1993.

[88] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2010.

[89] B. Heise, *Mehrgitter-Newton-Verfahren zur Berechnung nichtlinearer magnetischer Felder*, PhD thesis, Technical University of Chemnitz, 1991.

[90] B. Heise, *Analysis of a Fully Discrete Finite Element Method for a Nonlinear Magnetic Field Problem*, SIAM J. Numer. Anal., 31 (1994), pp. 745–759.

[91] P. W. Hemker, *A singularly perturbed model problem for numerical computation*, J. Comput. Appl. Math., 76 (1996), pp. 277–285.

[92] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[93] D. Hoonhout, R. Löscher, O. Steinbach, and C. Urzúa-Torres, *Stable least-squares space-time boundary element methods for the wave equation*, 2023, `https://arxiv.org/abs/2312.12547`.

[94] T. J. R. Hughes, L. P. Franca, and M. Mallet, *A new finite element formulation for computational fluid dynamics. VI. Convergence analysis of the generalized SUPG formulation for linear time-dependent multidimensional advective-diffusive systems*, Comput. Methods Appl. Mech. Engrg., 63 (1987), pp. 97–112.

[95] N. Ida and J. Bastos, *Electromagnetics and Calculation of Fields*, Springer New York, 1997.

[96] J. Jackson, *Classical Electrodynamics*, Wiley, 2021.

[97] A. Jha, V. John, and P. Knobloch, *Adaptive Grids in the Context of Algebraic Stabilizations for Convection-Diffusion-Reaction Equations*, SIAM J. Sci. Comput., 45 (2023), pp. B564–B589.

[98] V. John, P. Knobloch, and O. Pártl, *A Numerical Assessment of Finite Element Discretizations for Convection-Diffusion-Reaction Equations Satisfying Discrete Maximum Principles*, Comput. Methods Appl. Math., 23 (2023), pp. 969–988.

[99] JSOL-Corporation, *JMAG: Simulation technology for electromechanical design.* `https://www.jmag-international.com/`.

[100] M. Jung and U. Langer, *Methode der finiten Elemente für Ingenieure: Eine Einführung in die numerischen Grundlagen und Computersimulation*, Springer Fachmedien Wiesbaden, 2012.

[101] B. Kaltenbacher, M. Kaltenbacher, and S. Reitzinger, *Identification of nonlinear b–h curves based on magnetic field computations and multigrid methods for ill-posed problems*, European Journal of Applied Mathematics, 14 (2003), p. 15–38.

[102] M. Kaltenbacher, *Numerical Simulation of Mechatronic Sensors and Actuators: Finite Elements for Computational Multiphysics*, Springer Berlin Heidelberg, 2015.

[103] A. Kost, *Numerische Methoden in der Berechnung elektromagnetischer Felder*, Springer Berlin Heidelberg, 2013.

[104] C. Köthe, *PDE-constrained shape optimization for coupled problems using space-time finite elements*, Master's thesis, Graz University of Technology, 2020.

[105] S. Köthe, *Deterministic and stochastic shape and topology optimization of an electric machine*, Master's thesis, Graz University of Technology, 2020.

[106] C. Köthe, R. Löscher, and O. Steinbach, *Adaptive least-squares space-time finite element methods*, 2023, `https://arxiv.org/abs/2309.14300`.

[107] E. Lange, F. Henrotte, and K. Hameyer, *A Variational Formulation for Nonconforming Sliding Interfaces in Finite Element Analysis of Electric Machines*, IEEE Transactions on Magnetics, 46 (2010), pp. 2755–2758.

[108] U. Langer, S. E. Moore, and M. Neumüller, *Space-time isogeometric analysis of parabolic evolution problems*, Comput. Methods Appl. Mech. Engrg., 306 (2016), pp. 342–363.

[109] U. Langer, D. Pauly, and S. Repin (Eds.), *Maxwell's Equations. Analysis and Numerics*, vol. 24 of Radon Series on Computational and Applied Mathematics, De Gruyter, Berlin, 2019.

[110] U. Langer and A. Schafelner, *Adaptive Space-Time Finite Element Methods for Non-autonomous Parabolic Problems with Distributional Sources*, Comput. Methods Appl. Math., 20 (2020), pp. 677–693.

[111] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang, *Space-Time Finite Element Discretization of Parabolic Optimal Control Problems with Energy Regularization*, SIAM J. Numer. Anal., 59 (2021), pp. 675–695.

[112] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang, *Unstructured Space-Time Finite Element Methods for Optimal Control of Parabolic Equations*, SIAM J. Sci. Comput., 43 (2021), pp. A744–A771.

[113] U. Langer and O. Steinbach (Eds.), *Space-Time Methods: Applications to Partial Differential Equations*, vol. 25 of Radon Ser. Comput. Appl. Math., De Gruyter, Berlin, Boston, 2019.

[114] U. Langer and M. Zank, *Efficient Direct Space-Time Finite Element Solvers for Parabolic Initial-Boundary Value Problems in Anisotropic Sobolev Spaces*, SIAM J. Sci. Comput., 43 (2021), pp. A2714–A2736.

[115] G. Lei, J. Zhu, Y. Guo, C. Liu, and B. Ma, *A Review of Design Optimization Methods for Electrical Machines*, Energies, 10 (2017).

[116] J. Li, *A Nonlinear Mixed Problem Framework for Discontinuous Petrov-Galerkin (DPG) Methods*, PhD thesis, The University of Texas at Austin, 2024.

[117] M. Łoś, P. Sepúlveda, M. Sikora, and M. Paszyński, *Solver algorithm for stabilized space-time formulation of advection-dominated diffusion problem*, Computers and Mathematics with Applications, 152 (2023), pp. 67–80.

[118] R. Löscher, *On unified frameworks for space-time optimal control and least squares problems*, PhD thesis, Technische Universität Graz, 2023.

[119] M. Majidi and G. Starke, *Least-Squares Galerkin Methods for Parabolic Problems. I. Semidiscretization in Time*, SIAM J. Numer. Anal., 39 (2001), pp. 1302–1323.

[120] M. Majidi and G. Starke, *Least-Squares Galerkin Methods for Parabolic Problems. II. The Fully Discrete Case and Adaptive Algorithms*, SIAM J. Numer. Anal., 39 (2001/02), pp. 1648–1666.

[121] J. M. Maubach, *Local Bisection Refinement for N-Simplicial Grids Generated by Reflection*, SIAM J. Sci. Comput., 16 (1995), pp. 210–227.

[122] J. C. Maxwell, *A Treatise on Electricity and Magnetism. Vol. 1*, Oxford Classic Texts in the Physical Sciences, The Clarendon Press, Oxford University Press, New York, 1998. With prefaces by W. D. Niven and J. J. Thomson, Reprint of the third (1891) edition.

[123] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, 2000.

[124] C. Mellak, A. Cesarano, P. Gangl, J. Deuringer, and A. Muetze, *Free-Form Rotor Optimization for Synchronous Reluctance Machines used in X-ray Tubes*, in 2023 IEEE International Electric Machines & Drives Conference (IEMDC), 2023, pp. 1–7.

[125] C. Mellak, K. Krischan, and A. Muetze, *Synchronous Reluctance Machines as Drives for Rotary Anode X-Ray Tubes-A Feasibility Study*, in 2018 XIII International Conference on Electrical Machines (ICEM), 2018, pp. 2613–2618.

[126] N. G. Meyers and J. Serrin, $H = W$, Proc. Nat. Acad. Sci. U.S.A., 51 (1964), pp. 1055–1056.

[127] H. Monsuur, R. Stevenson, and J. Storn, *Minimal residual methods in negative or fractional Sobolev norms*, Math. Comp., 93 (2024), pp. 1027–1052.

[128] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data Oscillation and Convergence of Adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[129] P. Morin, K. G. Siebert, and A. Veeser, *A basic convergence result for conforming adaptive finite elements*, Math. Models Methods Appl. Sci., 18 (2008), pp. 707–737.

[130] M. Neumüller, *Space-Time Methods: Fast Solvers and Applications*, PhD thesis, Graz University of Technology, 2013.

[131] M. Neumüller and O. Steinbach, *Refinement of flexible space–time finite element meshes and discontinuous Galerkin methods*, Computing and Visualization in Science, 14 (2011), pp. 189–205.

[132] M. Neumüller, P. S. Vassilevski, and U. E. Villa, *Space-Time CFOSLS Methods with AMGe Upscaling*, in Domain Decomposition Methods in Science and Engineering XXIII, C.-O. Lee et al, ed., Cham, 2017, Springer International Publishing, pp. 253–260.

[133] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2006.

[134] J. T. Oden and L. F. Demkowicz, *Applied Functional Analysis*, Textbooks in Mathematics, CRC Press, 2018.

[135] C. C. Paige and M. A. Saunders, *Solutions of Sparse Indefinite Systems of Linear Equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[136] C. Pechstein, *Multigrid-Newton-Methods for Nonlinear Magnetostatic Problems*, Master's thesis, Johannes Kepler University Linz, 2004.

[137] C. Pechstein and B. Jüttler, *Monotonicity-preserving interproximation of B-H-curves*, J. Comput. Appl. Math., 196 (2006), pp. 45–57.

[138] T. Preston, A. Reece, and P. Sangha, *Induction motor analysis by time-stepping techniques*, IEEE Transactions on Magnetics, 24 (1988), pp. 471–474.

[139] S. Repin, *A Posteriori Estimates for Partial Differential Equations*, vol. 4 of Radon Series on Computational and Applied Mathematics, Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[140] D. Rodger, H. Lai, and P. Leonard, *Coupled elements for problems involving movement (switched reluctance motor)*, IEEE Transactions on Magnetics, 26 (1990), pp. 548–550.

[141] H. Roos, M. Stynes, and L. Tobiska, *Robust Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion-Reaction and Flow Problems*, Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 2008.

[142] E. Rothe, *Über die Wärmeleitungsgleichung mit nichtkonstanten Koeffizienten im räumlichen Falle. I, II.*, Math. Ann., 104 (1931), pp. 340–354,355–362.

[143] M. Růžička, *Nichtlineare Funktionalanalysis: Eine Einführung*, Masterclass, Springer Berlin Heidelberg, 2020.

[144] N. Sadowski, Y. Lefevre, M. Lajoie-Mazenc, and J. Cros, *Finite element torque calculation in electrical machines while considering the movement*, IEEE Transactions on Magnetics, 28 (1992), pp. 1410–1413.

[145] R. Saidur, *A review on electrical motors energy use and energy savings*, Renewable and Sustainable Energy Reviews, 14 (2010), pp. 877–898.

[146] A. Schafelner, *Space-time Finite Element Methods*, PhD thesis, Johannes Kepler University Linz, 2021.

[147] A. Schafelner and P. S. Vassilevski, *Numerical results for adaptive (negative norm) constrained first order system least squares formulations*, Comput. Math. Appl., 95 (2021), pp. 256–270.

[148] O. Schenk, M. Bollhöfer, and R. A. Römer, *On Large-Scale Diagonalization Techniques for the Anderson model of Localization*, SIAM Rev., 50 (2008), pp. 91–112.

[149] O. Schenk and K. Gärtner, *PARDISO*, in Encyclopedia of Parallel Computing, D. Padua, ed., Boston, MA, 2011, Springer US, pp. 1458–1464.

[150] O. Schenk, A. Wächter, and M. Hagemann, *Matching-based preprocessing algorithms to the solution of saddle-point problems in large-scale nonconvex interior-point optimization*, Comput. Optim. Appl., 36 (2007), pp. 321–341.

[151] F. Schlögl, *A characteristic critical quantity in nonequilibrium phase transitions*, Zeitschrift für Physik B Condensed Matter, 52 (1983), pp. 51–60.

[152] J. Schöberl, *C++ Implementation of Finite Elements in NGSolve*, Tech. Report 30, Institute for Analysis and Scientific Computing, TU Wien, 2014.

[153] C. Schwab and R. Stevenson, *Space-time adaptive wavelet methods for parabolic evolution problems*, Math. Comp., 78 (2009), pp. 1293–1318.

[154] B. Schweizer, *Partielle Differentialgleichungen. Eine anwendungsorientierte Einführung*, Springer-Verlag, 2023.

[155] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.

[156] O. Steinbach, *Numerical Approximation Methods for Elliptic Boundary Value Problems*, Springer, New York, 2008.

[157] O. Steinbach, *Space-time finite element methods for parabolic problems*, Comput. Meth. Appl. Math., 15 (2015), pp. 551–566.

[158] O. Steinbach, *An adaptive least squares boundary element method for elliptic boundary value problems*, 2023. Berichte aus dem Institut für Angewandte Mathematik, Bericht 2023/1, TU Graz.

[159] O. Steinbach and H. Yang, *An Algebraic Multigrid Method for an Adaptive Space-Time Finite Element Discretization*, in Large-Scale Scientific Computing, I. Lirkov and S. Margenov, eds., Springer, 2018, pp. 66–73.

[160] O. Steinbach and H. Yang, *Comparison of algebraic multigrid methods for an adaptive space-time finite-element discretization of the heat equation in 3D and 4D*, Numer. Linear Algebra Appl., 25 (2018), p. e2143.

[161] O. Steinbach and H. Yang, *Space-time finite element methods for parabolic evolution equations: discretization, a posteriori error estimation, adaptivity and solution*, in Space-Time methods: Applications to Partial Differential Equations, vol. 25 of Radon Ser. Comput. Appl. Math., De Gruyter, Berlin, 2019, pp. 207–248.

[162] R. Stevenson, *Optimality of a Standard Adaptive Finite Element Method*, Found. Comput. Math., 7 (2007), pp. 245–269.

[163] R. Stevenson, *The completion of locally refined simplicial partitions created by bisection*, Math. Comp., 77 (2008), pp. 227–241.

[164] R. Stevenson and J. Westerdiep, *Minimal residual space-time discretizations of parabolic equations: Asymmetric spatial operators*, Comput. Math. Appl., 101 (2021), pp. 107–118.

[165] R. Stevenson and J. Westerdiep, *Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations*, IMA J. Numer. Anal., 41 (2021), pp. 28–47.

[166] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, vol. 25 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[167] C. T. Traxler, *An Algorithm for Adaptive Mesh Refinement in n Dimensions*, Computing, 59 (1997), pp. 115–137.

[168] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*, Vieweg Studium, Vieweg+Teubner Verlag, 2010.

[169] I. Tsukerman, *Accurate computation of 'ripple solutions' on moving finite element meshes*, IEEE Transactions on Magnetics, 31 (1995), pp. 1472–1475.

[170] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, Journal of Computational and Applied Mathematics, 50 (1994), pp. 67–83.

[171] R. Verfürth, *A Posteriori Error Estimation Techniques for Finite Element Methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[172] K. Voronin, C. S. Lee, M. Neumüller, P. Sepulveda, and P. S. Vassilevski, *Space-time discretizations using constrained first-order system least squares (CFOSLS)*, J. Comput. Phys., 373 (2018), pp. 863–876.

[173] J. Xu and L. Zikatanov, *Some observations on Babuška and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.

[174] S. Zaglmayer, *High Order Finite Elements for Electromagnetic Field Computation*, PhD thesis, Johannes Kepler University Linz, 2006.

[175] M. Zank, *Inf-Sup Stable Space-Time Methods for Time-Dependent Partial Differential Equations*, PhD thesis, Graz University of Technology, 2019.

[176] E. Zeidler, *Nonlinear Functional Analysis and Its Applications II/A: Linear Monotone Operators*, Springer-Verlag New York, 1990.

[177] E. Zeidler, *Nonlinear Functional Analysis and Its Applications. II/B: Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

[178] E. Zeidler, *Applied Functional Analysis: Applications to Mathematical Physics*, vol. 108 of Applied Mathematical Sciences, Springer-Verlag, New York, 1995.

[179] E. Zeidler, *Applied Functional Analysis: Main Principles and Their Applications*, vol. 109 of Applied Mathematical Sciences, Springer, New York, 1995.

[180] W. P. Ziemer, *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, vol. 120 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1989.

[181] O. C. Zienkiewicz and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, International Journal for Numerical Methods in Engineering, 24 (1987), pp. 337–357.

[182] W. Zulehner, *Numerische Mathematik: Eine Einführung anhand von Differentialgleichungsproblemen. Band 1. Stationäre Probleme*, Mathematik Kompakt, Birkhäuser Verlag, Basel, 2008.

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 1 Steffen Alvermann
Effective Viscoelastic Behavior of Cellular Auxetic Materials
2008
ISBN 978-3-902465-92-4

Vol. 2 Sendy Fransiscus Tantono
The Mechanical Behaviour of a Soilbag
under Vertical Compression
2008
ISBN 978-3-902465-97-9

Vol. 3 Thomas Rüberg
Non-conforming FEM/BEM Coupling in Time Domain
2008
ISBN 978-3-902465-98-6

Vol. 4 Dimitrios E. Kiousis
Biomechanical and Computational Modeling
of Atherosclerotic Arteries
2008
ISBN 978-3-85125-023-7

Vol. 5 Lars Kielhorn
A Time-Domain Symmetric Galerkin BEM
for Viscoelastodynamics
2009
ISBN 978-3-85125-042-8

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 6 Gerhard Unger
Analysis of Boundary Element Methods for Laplacian Eigenvalue Problems
2009
ISBN 978-3-85125-081-7

Vol. 7 Gerhard Sommer
Mechanical Properties of Healthy and Diseased Human Arteries
2010
ISBN 978-3-85125-111-1

Vol. 8 Mathias Nenning
Infinite Elements for Elasto- and Poroelastodynamics
2010
ISBN 978-3-85125-130-2

Vol. 9 Thanh Xuan Phan
Boundary Element Methods for Boundary Control Problems
2011
ISBN 978-3-85125-149-4

Vol. 10 Loris Nagler
Simulation of Sound Transmission through Poroelastic Plate-like Structures
2011
ISBN 978-3-85125-153-1

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 11 Markus Windisch
Boundary Element Tearing and Interconnecting Methods for Acoustic and Electromagnetic Scattering
2011
ISBN 978-3-85125-152-4

Vol. 12 Christian Walchshofer
Analysis of the Dynamics at the Base of a Lifted Strongly Buoyant Jet Flame Using Direct Numerical Simulation
2011
ISBN 978-3-85125-185-2

Vol. 13 Matthias Messner
Fast Boundary Element Methods in Acoustics
2012
ISBN 978-3-85125-202-6

Vol. 14 Peter Urthaler
Analysis of Boundary Element Methods for Wave Propagation in Porous Media
2012
ISBN 978-3-85125-216-3

Vol. 15 Peng Li
Boundary Element Method for Wave Propagation in Partially Saturated Poroelastic Continua
2012
ISBN 978-3-85125-236-1

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 16 Andreas Jörg Schriefl
Quantification of Collagen Fiber Morphologies in Human Arterial Walls
2013
ISBN 978-3-85125-238-5

Vol. 17 Thomas S. E. Eriksson
Cardiovascular Mechanics
2013
ISBN 978-3-85125-277-4

Vol. 18 Jianhua Tong
Biomechanics of Abdominal Aortic Aneurysms
2013
ISBN 978-3-85125-279-8

Vol. 19 Jonathan Rohleder
Titchmarsh–Weyl Theory and Inverse Problems for Elliptic Differential Operators
2013
ISBN 978-3-85125-283-5

Vol. 20 Martin Neumüller
Space-Time Methods
2013
ISBN 978-3-85125-290-3

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 21 Michael J. Unterberger
Microstructurally-Motivated Constitutive Modeling of Cross-Linked Filamentous Actin Networks
2013
ISBN 978-3-85125-303-0

Vol. 22 Vladimir Lotoreichik
Singular Values and Trace Formulae for Resolvent Power Differences of Self-Adjoint Elliptic Operators
2013
ISBN 978-3-85125-304-7

Vol. 23 Michael Meßner
A Fast Multipole Galerkin Boundary Element Method for the Transient Heat Equation
2014
ISBN 978-3-85125-350-4

Vol. 24 Lorenz Johannes John
Optimal Boundary Control in Energy Spaces
2014
ISBN 978-3-85125-373-3

Vol. 25 Hannah Weisbecker
Softening and Damage Behavior of Human Arteries
2014
ISBN 978-3-85125-370-2

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 26 Bernhard Kager
Efficient Convolution Quadrature based Boundary Element Formulation for Time-Domain Elastodynamics
2015
ISBN 978-3-85125-382-5

Vol. 27 Christoph M. Augustin
Classical and All-floating FETI Methods with Applications to Biomechanical Models
2015
ISBN 978-3-85125-418-1

Vol. 28 Elias Karabelas
Space-Time Discontinuous Galerkin Methods for Cardiac Electromechanics
2016
ISBN 978-3-85125-461-7

Vol. 29 Thomas Traub
A Kernel Interpolation Based Fast Multipole Method for Elastodynamic Problems
2016
ISBN 978-3-85125-465-5

Vol. 30 Matthias Gsell
Mortar Domain Decomposition Methods for Quasilinear Problems and Applications
2017
ISBN 978-3-85125-522-5

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 31 Christian Kühn
Schrödinger operators and singular infinite rank perturbations
2017
ISBN 978-3-85125-551-5

Vol. 32 Michael H. Gfrerer
Vibro-Acoustic Simulation of Poroelastic Shell Structures
2018
ISBN 978-3-85125-573-7

Vol. 33 Markus Holzmann
Spectral Analysis of Transmission and Boundary Value Problems for Dirac Operators
2018
ISBN 978-3-85125-642-0

Vol. 34 Osman Gültekin
Computational Inelasticity of Fibrous Biological Tissues with a Focus on Viscoelasticity, Damage and Rupture
2019
ISBN 978-3-85125-655-0

Vol. 35 Justyna Anna Niestrawska
Experimental and Computational Analyses of Pathological Soft Tissues – Towards a Better Understanding of the Pathogenesis of AAA
2019
ISBN 978-3-85125-678-9

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 36 Marco Zank
Inf-Sup Stable Space-Time Methods for Time-Dependent Partial Differential Equations
2020
ISBN 978-3-85125-721-2

Vol. 37 Christoph Irrenfried
Convective turbulent near wall heat transfer at high Prandtl numbers
2020
ISBN 978-3-85125-724-3

Vol. 38 Christopher Albert
Hamiltonian Theory of Resonant Transport Regimes in Tokamaks with Perturbed Axisymmetry
2020
ISBN 978-3-85125-746-5

Vol. 39 Daniel Christopher Haspinger
Material Modeling and Simulation of Phenomena at the Nano, Micro and Macro Levels in Fibrous Soft Tissues of the Cardiovascular System
2021
ISBN 978-3-85125-802-8

Vol. 40 Markus Alfons Geith
Percutaneous Coronary Intervention
2021
ISBN 978-3-85125-801-1

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 41 Dominik Pölz
Space-Time Boundary Elements for Retarded Potential Integral Equations
2021
ISBN 978-3-85125-811-0

Vol. 42 Douglas Ramalho Queiroz Pacheco
Stable and stabilised finite element methods for incompressible flows of generalised Newtonian fluids
2021
ISBN 978-3-85125-856-1

Vol. 43 Peter Schlosser
Superoscillations and their Schrödinger time evolution
2022
ISBN 978-3-85125-930-8

Vol. 44 Raphael Watschinger
Fast space-time boundary element methods for the heat equation
2023
ISBN 978-3-85125-949-0

Vol. 45 Ishan Gupta
Modelling Growth and Formation of Thrombus: A Multiphasic Approach
2023
ISBN 978-3-85125-964-3

Monographic Series TU Graz
Computation in Engineering and Science

Vol. 46 Mario Gobrial
Space-time Finite Element Methods
for the Eddy Current Problem and Applications
2025
ISBN 978-3-99161-048-9

Vol. 47 Christian Stelzer-Landauer
Approximation of Dirac Operators with
Delta-Shell Potentials in the Norm Resolvent Sense
2025
ISBN 978-3-99161-052-6

Vol. 48 Köthe Christian
Adaptive least-squares
space-time finite element methods
2025
ISBN 978-3-99161-060-1

Adaptive least-squares space-time finite element methods

In this thesis we consider a minimal residual/least-squares approach for the solution of an abstract operator equation. In particular, the first objective is to apply a least-squares method in combination with a space-time discretization scheme for the solution of parabolic evolution equations. Moreover, a second objective is to consider the approach for the simulation of electric machines.

Firstly, we present an abstract minimal residual framework together with a complete stability and a priori error analysis. Furthermore, it is shown that under a saturation assumption the method obtains an efficient and reliable error indicator.

Secondly, we apply the least-squares framework together with a space-time discretization scheme to several parabolic PDEs, including the heat equation, the convection-diffusion equation and a heat equation with nonlinear reaction term. The inbuilt error estimator is used to drive an adaptive refinement scheme and numerical experiments confirm our theoretical findings.

Finally, we apply the method to the simulation of different electric machine models, including the magnetostatic approximation, a quasistatic approximation and the eddy current problem. Numerical examples demonstrate the correctness of the proposed method.

MONOGRAPHIC SERIES TU GRAZ
COMPUTATION IN ENGINEERING AND SCIENCE

Verlag der Technischen Universität Graz
www.tugraz-verlag.at

ISBN 978-3-99161-060-1
ISSN 1990-357X

ISBN 978-3-99161-060-1
9 783991 610601