GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

On a Discontinuous Galerkin discretization for a degenerate diffusion equation

Bachelor's thesis

submitted by

Tim van Beeck

Supervisor:

Prof. Dr. Christoph Lehrenfeld

Second Assesor:

Prof. Dr. Gert Lube

Institute for Numerical and Applied Mathematics
University of Göttingen

December 3, 2021

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe.Wörtlich oder sinngemäß aus anderen Werken entnommene Stellen habe ich unter Angabe der Quellen kenntlich gemacht.

Göttingen, den 03.12.2021.

Acknowledgements

First and foremost, I would like to express my gratitude to Prof. Dr. Christoph Lehrenfeld for supervising this thesis and proposing such an interesting topic. Further, I am thankful for his guidance in the process of creating the thesis. His willingness to discuss the topic at hand, any arising problems, and questions of mine has greatly benefited me.

Secondly, I am grateful to Prof. Dr. Gert Lube for serving as a Second Assessor on this thesis.

Finally, I would like to thank Paul Stocker for his suggestions for improvement.

Abstract

In this thesis, we develop a discretization for a degenerate diffusion equation in the sense that the diffusion only acts along a velocity field, which is a problem arising in the context of a numerical model for solar and stellar oscillations. The main contribution of this thesis is the derivation and analysis of a suitable symmetric interior penalty Discontinuous Galerkin method. Proving coercivity and continuity, the well-posedness of the discrete problem is established. Afterwards, a Ceá-type a priori error estimate is shown, and standard interpolation results yield an optimal convergence result in an energy-like norm. We test the method numerically for some example problems and consider especially constant and non-constant densities.

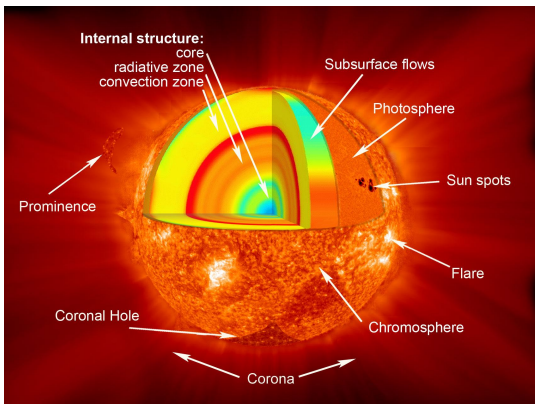
Contents

Introduction

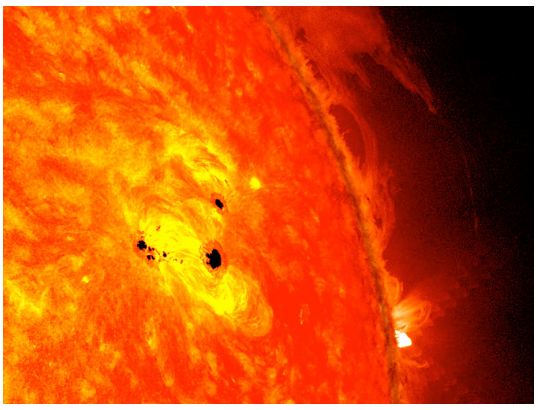
Motivation

The sun has been studied intensively, yet many questions, like the origin of its magnetic field, still remain open. One scientific field that might provide some answers is helioseismology which studies the solar interior through solar oscillations.
In general, helioseismology methods can be classified into two classes: global helioseismology and local helioseismology. The former studies the structure and physical conditions in the interior of the sun by observing its modes of oscillation and building models to match the data, whereas the latter uses surface flows in the convection zone to model three dimensional subsurface flows.
As such, local helioseismology is important for studying phenomena like sunspots, darker regions on the surface of the sun. Sunspots appear darker because they have a lower temperature than their surrounding, which correlates with a stronger magnetic field in this region. Figure 1 shows the different layers in the sun and a sunspot.
A more extensive overview over local helioseismology is provided by Gizon et al. in [GBS10].

Interpreting the solar seismic waves requires numerical simulations. Hence, a task of a project of the CRC1456[1] aims to develop a numerical model for the equations of solar and stellar oscillation. This thesis deals with a sub-problem of this task under simplifying assumptions.



(a) Sunlayers, image credit: NASA/Goddard[2]



(b) Sunspot, image credit: NASA/SDO/AIA[3]

Figure 1: An overview of the different layers of the sun (left) and a sunspot (right).

[1] specifically project C04, which can be found at https://www.uni-goettingen.de/en/630954.html
[2] available at https://www.nasa.gov/images/content/462977main_sun_layers_full.jpg, accessed 11.10.2021.
[3] available at https://spaceplace.nasa.gov/solar-activity/en/solar-activity2.en.jpg, accessed 11.10.2021.

Let us describe the time-harmonic equations of solar and stellar oscillations without magnetic fields as derived by Lynden-Bell and Ostriker [LO67]:
Consider the background velocity u, the density ρ, the pressure p and the gravitational potential ϕ of a stationary equilibrium solution of the conservation equations of mass and momentum, given by the Euler equations and a potential equation for the gravitational potential. Then, the displacement perturbations ξ of Lagrangian particles and the Eulerian perturbation φ of the gravitational background potential ϕ satisfy

$$\begin{aligned}\rho(-i\omega + u\cdot\nabla + \Omega\times)^2\xi - \nabla(\rho c_s^2\nabla\cdot\xi) + (\nabla\cdot\xi)\nabla p - \nabla(\nabla p\cdot\xi)\\ +(\mathrm{Hess}(p)\xi - \rho\mathrm{Hess}(\phi))\xi - i\gamma\rho\omega\xi + \rho\nabla\varphi = s \text{ in } D,\\ -\frac{1}{4\pi G}\Delta\varphi + \nabla\cdot(\rho\xi) = 0 \text{ in } \mathbb{R}^3,\end{aligned} \tag{0.1}$$

where $D\subset\mathbb{R}^3$ is a bounded Lipschitz domain. Further G denotes the gravitational constant, c_s the speed of sound, s the source terms caused by turbulent convection, Ω the uniform angular velocity of the frame of reference and ω the angular frequency of the waves.

Recently, Halla and Hohage [HH21] proved the well-posedness of these equations, that is the existence, uniqueness, and stability of solutions. Apart from this paper, there are few results on the theoretical properties of these type of equations. In the following, we will briefly describe the theoretical framework used in the paper. In particular, the following is only relevant for the larger context and is not treated in the remainder of this thesis.
Before we introduce a weak formulation of (0.1), we will pose some assumptions on the background flow u. First of all, we require u to be $H(\mathrm{div})$-type regular. Furthermore, u is assumed to be subsonic, which means that the velocity of the background flow is lower that the speed of sound. Mathematically, we can describe the latter in the following way: For sufficiently smooth ρ, c_s, D and homogeneous pressure p and gravity ϕ, u satisfies

$$\|c_s^{-1}u\|_\infty < 1.$$

Additionally, the Cowling approximation, cf. to [Chr03], sets $\varphi = 0$ reducing the problem to a problem for ξ only.

Now, we define the following function space:

$$\mathbf{X} := \{\xi\in L^2(D;\mathbb{C}^3) \mid \mathrm{div}\,\xi\in L^2(D;\mathbb{C}),\ u\cdot\nabla\xi\in L^2(D;\mathbb{C}^3),\ \xi\cdot\nu = 0 \text{ on } \partial D\}.$$

We equip this space with the canonical inner product

$$\langle\xi,\xi'\rangle_X = \langle\nabla\cdot\xi,\nabla\cdot\xi'\rangle + \langle u\cdot\nabla\xi, u\cdot\nabla\xi'\rangle + \langle\xi,\xi'\rangle,\qquad \xi,\xi'\in\mathbf{X},$$

where $\langle\cdot,\cdot\rangle$ denotes the standard $L^2(\mathbb{C})$ inner product. With the assumptions on u from above, this space becomes a Hilbert space [HH21, Lemma 2.1].
Then, the weak formulation of (0.1) reads as:
Find $\xi\in\mathbf{X}$ such that

$$a(\xi,\xi') = \sum_{i=1}^{3} a^i(\xi,\xi') = \langle s,\xi'\rangle \text{ for all } \xi'\in\mathbf{X}, \tag{0.2}$$

where we define the sesquilinear forms $a^i(\cdot,\cdot)$, $1\le i\le 3$, through

$$\begin{aligned}a^1(\xi,\xi') &:= \langle\rho c_s^2(\nabla+\mathbf{q})\cdot\xi, (\nabla+\mathbf{q})\cdot\xi'\rangle - \langle\rho c_s^2\mathbf{q}\cdot\xi,\mathbf{q}\cdot\xi'\rangle,\\ a^2(\xi,\xi') &:= -\langle\rho\mathcal{D}\xi,\mathcal{D}\xi'\rangle + \langle(\mathrm{Hess}(p)-\rho\mathrm{Hess}(\phi)\xi,\xi'\rangle,\\ a^3(\xi,\xi') &:= -i\rho\gamma\omega\langle\xi,\xi'\rangle,\ \text{for } \xi,\xi'\in\mathbf{X}.\end{aligned}$$

Here, for ease of notation, the differential operator $\mathcal{D} := (-i\omega + u \cdot \nabla + \Omega\times)$ relating to transport and the function $\mathbf{q} := \rho^{-1}c^{-2}\nabla p$ were introduced.

The following generalized Helmholtz decomposition is crucial for proving the well-posedness of the problem.

$$\mathbf{X} = \mathbf{V} \oplus \mathbf{W} \oplus \mathbf{Z}, \tag{0.3}$$

where

- $\mathbf{V} \subset \{\nabla v \mid v \in H^2(D),\ \nabla v \cdot \nu = 0 \text{ on } \partial D\}$ is compactly embedded in L^2;
- $\mathbf{W} = \{\xi \in \mathbf{X} \mid (\nabla + q) \cdot \xi = 0\}$;
- $\mathbf{Z}$ is finite dimensional.

Note that for $\mathbf{q} = 0$ and $\mathbf{Z} = \{0\}$ this decomposition is the classical Helmholtz decomposition where $\mathbf{u} \in \mathbf{X}$ is decomposed into a divergence free and a curl free function. For the proof of the decomposition we refer to [HH21, Lemma 3.5].

This result can be used to show that the operator A induced by the bilinear form $a(\cdot,\cdot)$ is weakly T-coercive [HH21, Theorem 3.11]. The concept of weak T-coercivity was introduced by Halla [Hal19] and implies that the operator is Fredholm with index zero. This in turn implies the well-posedness of the problem[4].

As mentioned above, there are few similar theoretical results on these type of equations. Consequently, there are even fewer numerical discretization schemes. We want to develop a finite element based discretization with provable error bounds that is tailored to the structure of this problem. Especially, the decomposition (0.3) of the continuous problem should be transferred to the discrete case.
Assuming $\mathbf{Z} = \{0\}$ and $\mathrm{Hess}(p) - \rho\mathrm{Hess}(\phi) = 0$, we need to consider a decomposition of finite element spaces

$$\mathbf{X}_h = \mathbf{V}_h \oplus \mathbf{W}_h. \tag{0.4}$$

To simplify things further, we assume constant pressure such that $\mathbf{q} = 0$.
From the decomposition (0.3) we have that

$$\mathbf{W} = \{\xi \in \mathbf{X} \mid \nabla \cdot \xi = 0\}, \tag{0.5}$$

which means that $\mathbf{W}$ is the divergence free subspace of $\mathbf{X}$. Hence, we consider a discrete version of this space:

$$\mathbf{W}_h := \{\xi_h \in \mathbf{X}_h \mid \nabla \cdot \xi_h = 0\}. \tag{0.6}$$

Further, we choose $\mathbf{V}_h = \mathbf{W}_h^{\perp}$ to be the orthogonal complement of $\mathbf{W}_h$ in $\mathbf{X}_h$.

For the treatment of divergence-free subspaces, we are considering $H(\mathrm{div}, D)$-conforming, that is normal continuous, finite element spaces $\mathbf{X}_h$, which are in general not tangential-continuous. That means, that functions $\xi_h \in \mathbf{X}_h$ can be discontinuous across element interfaces. Hence, we are aiming for a discontinuous Galerkin discretization.

[4] for an overview of the "Fredholm alternative" we refer to [Eva10, Chapter 6.2.3].

This thesis deals with a sub-problem of this task under further simplifying assumptions and thus can be viewed as a starting point for further work. We restrict to function spaces over $\mathbb{R}^d$ and the standard L^2-scalar product and remove all zero order terms. This eliminates the damping term and yields $\mathcal{D} = \partial_u := u \cdot \nabla$. The bilinear form in the weak formulation (0.2) becomes

$$a(\xi, \xi') = \sum_{i=1}^{2} a^i(\xi, \xi') = \langle \rho c^2 \nabla \cdot \xi, \nabla \cdot \xi \rangle - \langle \rho \partial_u \xi, \partial_u \xi' \rangle, \quad \xi, \xi' \in \mathbf{X}. \tag{0.7}$$

Now, for a discrete version of this bilinear form we consider

$$a_h(\xi_h, \xi_h') = \sum_{i=1}^{2} a_h^i(\xi_h, \xi_h'), \quad \xi_h, \xi_h' \in \mathbf{X_h}, \tag{0.8}$$

where $a_h^1(\xi_h, \xi_h') = a^1(\xi_h, \xi_h)$ and $a_h^2(\cdot, \cdot)$ is a discrete version of $a^2(\cdot, \cdot)$ that is yet to be defined. Using the decomposition (0.3), we can split $\xi_h = v_h + w_h$ for $v_h \in \mathbf{V}_h$, $w_h \in \mathbf{W}_h$. Choosing $\xi_h' = v_h - w_h$ yields

$$\begin{aligned} a_h(\xi_h, \xi_h') &= \langle \rho c^2 \nabla \cdot v_h, \nabla \cdot v_h \rangle - \langle \rho \partial_u v_h, \partial_u v_h \rangle + \langle \rho \partial_u w_h, \partial_u w_h \rangle \\ &= \underbrace{\langle \rho c^2 \nabla \cdot v_h, \nabla \cdot v_h \rangle + a_h^2(v_h, v_h)}_{=: \widetilde{a}_h(v_h, v_h)} \underbrace{- a_h^2(w_h, w_h)}_{=: b_h(w_h, w_h)} \end{aligned} \tag{0.9}$$

The bilinear form $b_h(\cdot, \cdot)$ will be defined and analysed in this thesis. It corresponds to a diffusion equation of the following form

$$-\nabla \cdot (\rho(u \otimes u) \nabla w) = f \text{ in } D. \tag{0.10}$$

To make this problem uniquely solvable, in particular to avoid a non-trivial kernel, we add a volume term w. Further, we multiply the volume term with ρ so that both terms scale equally in dependence of the density. This helps to avoid conditioning issues in the numerical simulations.

Ultimately, this thesis aims to develop and analyse a Discontinuous Galerkin discretization for the following problem:
For a given smooth velocity field $u \in L^\infty(D, \mathbb{R}^d)$,a sufficiently smooth density $\rho : D \to \mathbb{R}_{>0}$, and a source term $f \in L^2(D)$, we search for $w : D \to \mathbb{R}$ such that

$$\rho w - \nabla \cdot (\rho(u \otimes u) \nabla w) = f \text{ in } D. \tag{0.11}$$

Structure of this thesis

This thesis is structured as followed:

Chapter 1 introduces the model problem and poses the continuous problem. Further, we derive a suitable Discontinuous Galerkin discretization (DG) of (0.11). The main focus will lie on a discrete bilinear form describing the second order operator in (0.11).

Chapter 2 focuses on a theoretical analysis of the discrete problem. We will prove coercivity and continuity, which imply the well-posedness of the discrete problem. Then, a Ceá-type a priori error estimate and corresponding interpolation results will be developed. Afterwards, we will dive deeper into some specific issues, namely we will prove a non-standard inverse inequality and investigate the stabilization term used in the DG formulation in more detail.

Chapter 3 conducts numerical experiments to test the developed method. The discretization was implemented with the software package `Netgen/NGSolve` which can be found at

`https://ngsolve.org`

In particular, we will study convergence rates of the errors in the L^2- and an energy-like norm and the influences of the penalization parameter and a non-constant density.

Chapter 4 summarizes the results from this thesis and gives an overview of open problems.

The appendix consisting of **chapters A**, **B** and **C** contains supplementary material. The first chapter of the appendix gives a brief overview over triangulations and useful inequalities. In the second chapter, we will present the code used to implement the method. Finally, the last chapter contains additional plots and tables that complement the material in chapter 3.

Requirements and Notation

In order to understand this work, a basic knowledge of Finite Element Methods and the associated concepts of functional analysis is required.
For ease of presentation, the L^2-norm respectively the L^2-scalar product on a domain Ω are denoted as

$$\|\cdot\|_\Omega, \quad \text{resp.} \quad \langle\cdot,\cdot\rangle_\Omega.$$

We will also use the notation $a \lesssim b$ to indicate that a is less or equal than a constant times b, i.e.

$$a \lesssim b :\Leftrightarrow \exists c > 0 \text{ independent of } a \text{ and } b \text{ s.t. } a \leq c \cdot b. \tag{0.12}$$

We will use $\gtrsim$ analogously and $\simeq$ if $a \lesssim b$ and $a \gtrsim b$. On occasion, we will use Sobolev-spaces $H^p(\Omega)$ and denote the corresponding Sobolev-norm as $\|\cdot\|_{H^p(\Omega)}$.

Chapter 1

Derivation of a suitable DG discretization

In this chapter, we will introduce the model problem considered in this thesis. Then, we will pose the continuous problem and show that it is well-posed. Afterwards, we will derive a discrete Discontinuous Galerkin formulation for the model problem.

1.1 Model problem

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain and $u \in L^\infty(D, \mathbb{R}^d)$ a smooth velocity field. Further, let the density $\rho : D \to \mathbb{R}_{>0}$ be sufficiently smooth. Then, we consider the problem: Find $w : D \to \mathbb{R}$ such that

$$\begin{aligned} \rho w - \nabla \cdot (\rho(u \otimes u)\nabla w) &= f \text{ in } D, \\ w &= 0 \text{ on } \partial D^{\text{in}}. \end{aligned} \tag{1.1}$$

Here, with ν denoting the normal vector, we define

$$\partial D^{\text{in}} = \partial D \cap \{x \in D \mid u \cdot \nu \neq 0\}. \tag{1.2}$$

Note further, that for $a, b \in \mathbb{R}^d$, we define the dyadic product as

$$a \otimes b = ab^T. \tag{1.3}$$

Additionally, we assume that the density can be bounded from below and above:

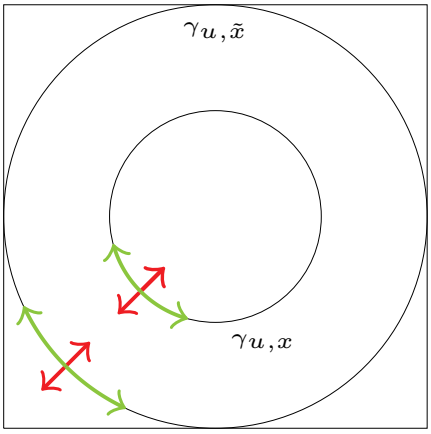
Assumption 1.1. *There exist constants $\underline{\rho}, \overline{\rho} \in \mathbb{R}_{>0}$ such that*

$$\underline{\rho} \leq \rho(x) \leq \overline{\rho} \quad \forall x \in D.$$

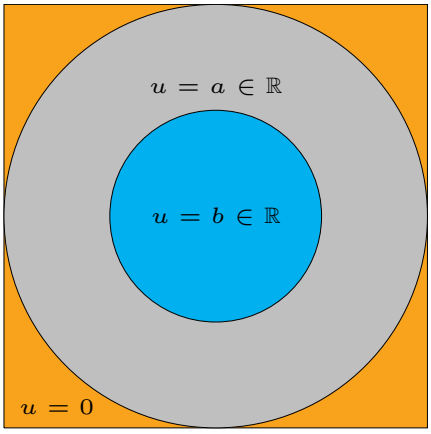
This problem is a reaction-diffusion equation. However, the diffusion operator is degenerative in the sense that it only acts along the velocity field u. Especially, that means that the diffusion operator vanishes whenever $u = 0$. Figure 1.1a visualizes the direction of the diffusion for a rotational velocity field. In particular, we see that there is no communication between the two trajectories.
Furthermore, the kernel of the diffusion operator is not trivial:

$$\ker(-\nabla \cdot (\rho(u \otimes u)\nabla w)) = \{w \in L^2(D) \mid w \text{ is constant along } \gamma_{x,u}, x \in D\}, \tag{1.4}$$

where $\gamma_{x,u}$ describes the trajectory through x along u in D. Figure 1.1b displays a non-trivial function in the kernel of the diffusion operator for our example from above. As a consequence of the non-trivial kernel, we need the volume term ρw to ensure that the problem is well-posed, cf. to the more detailed discussion in section 3.5.



(a) Direction of the diffusion



(b) Non-trivial function in the kernel

Figure 1.1: The direction of the diffusion along the velocity field u (left), where a green arrow indicates that there is diffusion and a red arrow indicates that there is no diffusion, and a function in the kernel of the diffusion operator (right).

1.2 The continuous setting

Before we derive the discrete problem, we will examine a continuous weak formulation of (1.1). To this end, we define the following function space:

$$W := \left\{ v \in L^2(D) \mid \rho \partial_u v \in L^2(D), v = 0 \text{ on } \partial D^{\text{in}} \right\}, \tag{1.5}$$

and equip it with the following norm:

$$|||w|||_W := \|\rho^{\frac{1}{2}} w\|_D + \|\rho^{\frac{1}{2}} \partial_u w\|_D. \tag{1.6}$$

Remark 1.1 (On the kernel of $|||\cdot|||_W$)**.** *Obviously, $|||\cdot|||_W$ satisfies absolute homogeneity and the triangle inequality. The kernel of $|||\cdot|||_W$, that is $\{w \in W \mid |||w|||_W = 0\}$, is trivial as well: While the second term would be zero for e.g. a constant term, the first term ensures that $|||w|||_W = 0 \Leftrightarrow w = 0$. Hence, $|||\cdot|||_W$ defines a proper norm.*

A common challenge when working with partial differential equations (PDEs) in general, is that physical meaningful solutions might not satisfy the classical definition of a derivative. For instance, functions that solve wave equations can have discontinuities and are not differentiable in the classical sense. However, as we are working with variational formulations, we only have to make sense of derivatives in integral equations and want to be able to apply partial integration.

Following Halla and Hohage [HH21, Equation (16)], we define the differential operator $\rho\partial_u$ in a weak sense:

$$\langle \rho \partial_u w, w' \rangle_D := -\langle \rho w, \partial_u w' \rangle_D - \langle \operatorname{div}(\rho u) w, w' \rangle_D \qquad \forall w' \in C_0^\infty(D). \tag{1.7}$$

Here, we assume that $\rho \partial_u w \in L^2(D)$.
To derive a weak formulation, we multiply (1.1) with a test function $w' \in W$ and integrate over the domain D. This yields:

$$\langle \rho w, w' \rangle_D - \int_D \nabla \cdot (\rho (u \otimes u) \nabla w) \cdot w' \, \mathrm{d}x. \tag{1.8}$$

Now, we apply partial integration on the second term and get

$$\langle \rho w, w' \rangle_D + \int_D (\rho(u \otimes u)\nabla w) \cdot \nabla w' \ \mathrm{d}x - \underbrace{\int_{\partial D} (\rho(u \otimes u)\nabla w) \cdot w' \nu \ \mathrm{d}s}_{=0}. \tag{1.9}$$

The boundary term vanishes as $w' = 0$ on ∂D^{in} and $u \cdot \nu = 0$ otherwise.

Further, we note that for the dyadic product holds

$$a(b \cdot c) = (a \otimes b) \cdot c \text{ for } a, b, c \in \mathbb{R}^d. \tag{1.10}$$

Hence, we can write the second term as:

$$\int_D \rho(u \cdot \nabla)w \cdot (u \cdot \nabla)w' \ \mathrm{d}x. \tag{1.11}$$

Recalling the definition $\partial_u \cdot = (u \cdot \nabla)\cdot$, (1.9) becomes

$$\langle \rho w, w' \rangle_D + \langle \rho \partial_u w, \partial_u w' \rangle_D =: \mathcal{B}(w, w'). \tag{1.12}$$

Altogether, we obtain the following weak formulation:
Find $w \in W$ such that

$$\mathcal{B}(w, w') = \langle f, w' \rangle_D \quad \forall w' \in W. \tag{1.13}$$

Ultimately, the weak formulation motivates the choice of the space W, as we need w and $\rho \partial_u v$ to be square integrable.

Remark 1.2 (Symmetry of the continuous problem)**.** *At the first glance, the continuous problem (1.13) does not appear to be symmetric. However, when one considers either the scalar product $\langle \rho \cdot, \cdot \rangle_D$ or splits the density such that the scalar products are given by $\langle \rho^{\frac{1}{2}} \cdot, \rho^{\frac{1}{2}} \cdot \rangle_D$, one quickly asserts that the problem is indeed symmetric.*

We want to consider the well-posedness of the weak formulation (1.13). To this end, we want to apply the Riesz theorem [Eva10, Appendix D, Thm. 2].

Theorem 1.1 (Riesz representation theorem)**.** *Let V be a Hilbert space and $\ell : V \to \mathbb{R}$ be a continuous linear functional. Then, there exits a unique $u_\ell \in V$ such that*

$$\ell(v) = \langle u_\ell, v \rangle_V \tag{1.14}$$

and

$$\|\ell\|_{V*} = \|u_\ell\|_V. \tag{1.15}$$

The right-hand side of the weak formulation (1.13) is linear as a scalar product. Furthermore, it is continuous due to the Cauchy-Schwarz inequality. If we show that $(W, \mathcal{B}(\cdot, \cdot))$ is a Hilbert space, theorem 1.1 yields the existence of a $w \in W$ such that

$$\mathcal{B}(w, w') = \langle f, w' \rangle_D. \tag{1.16}$$

The second results from Riesz' theorem gives us the stability of the solution w, which means that the continuous weak formulation is well-posed. The following lemma shows that W is indeed a Hilbert space.

Lemma 1.2. *The space $(W, \mathcal{B}(\cdot,\cdot))$ is a Hilbert space.*

Proof. The bilinear form $\mathcal{B}(\cdot,\cdot)$ is an inner product, because by definition there is

$$|||w|||_W = \sqrt{\mathcal{B}(w,w)}. \tag{1.17}$$

To prove the claim, we have to show completeness of W with respect to the $|||\cdot|||_W$-norm. Let $\{w_n\} \subset W$ be a Cauchy sequence.There holds that $W \subset L^2(D)$ and $\|w\|_D \leq |||w|||_W$ for all $\in W$, which means that $\{w_n\}$ is Cauchy in $L^2(D)$ as well. As $L^2(D)$ itself is a complete space, there exits a $\widetilde{w} \in L^2(D)$ such that $w_n \overset{L^2(D)}{\to} w$ as $n \to \infty$. To conclude the proof, we are left with showing that $\widetilde{w} \in W$ and that $|||w_n - \widetilde{w}|||_W \to 0$.
For the former, consider (1.7) with w_n: For all $w' \in C_0^\infty(D)$ there holds

$$\begin{aligned}\langle \rho\partial_u w_n, w'\rangle_D &\overset{(1.7)}{=} -\langle \rho w_n, \partial_u w'\rangle_D - \langle \mathrm{div}(\rho u) w_n, w'\rangle_D \\ &\overset{n\to\infty}{\to} -\langle \rho\widetilde{w}, \partial_u w'\rangle_D - \langle \mathrm{div}(\rho u)\widetilde{w}, w'\rangle_D \\ &= \langle \rho\partial_u \widetilde{w}, w'\rangle_D.\end{aligned} \tag{1.18}$$

Consequently, $\rho\partial_u\widetilde{w} \in L^2(D)$ and hence, $\widetilde{w} \in W$.
For the latter, let $n, m \geq N_\epsilon$ where N_ϵ stems from the Cauchy sequence property:

$$|||w_n - w_m|||_W \leq \epsilon \quad \forall n, m \geq N_\epsilon. \tag{1.19}$$

Now, we insert $w_n - w_m$ into (1.7): For all $w' \in C_0^\infty(D)$ there holds

$$\begin{aligned}\langle \rho\partial_u(w_n - w_m), w'\rangle_D &\overset{(1.7)}{=} -\langle \rho(w_n - w_m), \partial_u w'\rangle_D - \langle \mathrm{div}(\rho u)(w_n - w_m), w'\rangle_D \\ &\leq \sup_{\substack{w'\in[C_0^\infty(D)]^d,\\ \|w'\|_D=1}} -\langle \rho(w_n - w_m), \partial_u w'\rangle_D - \langle \mathrm{div}(\rho u)(w_n - w_m), w'\rangle_D \\ &\overset{(A.7)}{\leq} \sup_{\substack{w'\in[C_0^\infty(D)]^d,\\ \|w'\|_D=1}} -\|\rho(w_n - w_m)\|_D\|\partial_u w'\|_D - \|\mathrm{div}(\rho u)(w_n - w_m)\|_D \\ &\leq \sup_{\substack{w'\in[C_0^\infty(D)]^d,\\ \|w'\|_D=1}} \epsilon(-\|\rho\|_D\|\partial_u w'\|_D - \|\mathrm{div}(\rho u)\|_D)\end{aligned} \tag{1.20}$$

For $\epsilon \to 0$ this expression approaches zero as all the other terms are bounded. Letting $m \to \infty$ yields that $|||w_n - \widetilde{w}|||_W \to 0$. □

As described above, this result automatically yields the well-posedness of the continuous problem:

Theorem 1.3 (Well-posedness of the continuous problem)**.** *There exists a unique and stable solution $w \in W$ to the continuous problem (1.13).*

1.3 The discrete setting

We have posed the continuous weak formulation and shown its well-posedness. Now, we transfer the problem to the discrete setting. Therefore, let $\mathcal{T}_h$ be an admissible triangulation of D that is shape-regular and quasi-uniform[1].
The standard Galerkin formulation of (1.1) reads as:
Find $w_h \in W_h \subset W$ such that

$$\mathcal{B}(w_h, w_h') = \langle f, w_h' \rangle_D \quad w_h' \in W_h.$$

However, as explained in the introduction, we aim for a DG formulation of the problem. DG methods are non-conforming methods in the sense that the discrete functions do not have to be continuous across element interfaces. This means, that we consider the space of polynomials of degree k on every element $T \in \mathcal{T}_h$:

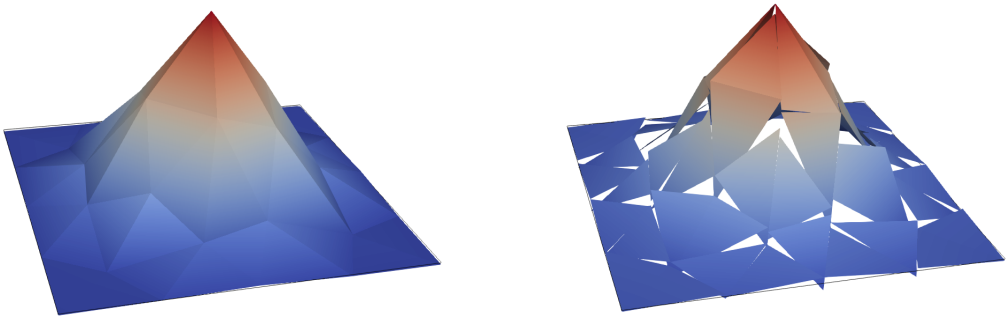
$$W_h := V_h^{k,d} = \{v \in L^2(D) \mid v|_T \in \mathcal{P}^k(T) \quad \forall T \in \mathcal{T}_h\}. \tag{1.21}$$

In contrast to the standard Galerkin method, we have that $W_h \not\subset W$ as the piecewise polynomials can be discontinuous. Figure 1.2 visualizes the difference between approximations with continuous and discontinuous piecewise polynomials.

Furthermore, we have to introduce a modified discrete bilinear form $\mathcal{B}_h(\cdot,\cdot)$ as $\mathcal{B}(\cdot,\cdot)$ is in general not well-defined on W_h. On the one hand, we want this bilinear form to be consistent, which means that for the true solution w there should hold

$$\mathcal{B}_h(w, w_h') = \langle f, w_h' \rangle_D \qquad \forall w_h' \in W_h.$$

On the other hand, we also want it to provide stability, which in this case means that is should fulfil discrete coercivity. To achieve this, we will introduce a stabilization term that penalizes the element interface jumps of the discrete functions.



(a) Continuous Polynomials (b) Discontinuous Polynomials

Figure 1.2: L^2-best-approximation of a Gaussian bump $w = e^{-2(x^2+y^2)}$ with continuous and discontinuous linear polynomials.

It is clear at the first glance, that the discrete formulation takes the form

$$\mathcal{B}_h(w_h, w_h') = \langle \rho w_h, w_h' \rangle_D + b_h(w_h, w_h'), \tag{1.22}$$

[1] for details on admissible and quasi-uniform triangulations we refer to chapter A.

where $b_h(\cdot,\cdot)$ is a bilinear form corresponding to the degenerated diffusion operator

$$-\nabla\cdot(\rho(u\otimes u)\nabla w_h). \tag{1.23}$$

In the following, we will focus on deriving this bilinear form. To do this, we first have to introduce some notations. Let $\mathcal{F}_h$ denote the space of all facets of the triangulation $\mathcal{T}_h$, that is

$$\mathcal{F}_h = \{\overline{\partial T_1}\cap\overline{\partial T_2} \mid T_1\neq T_2, T_1, T_2\in\mathcal{T}_h\}\cup\{\overline{\partial T}\cap\overline{\partial D} \mid T\in\mathcal{T}_h\}.$$

Further, we introduce the following jump- and average operators:
For two elements T_1 and T_2 with a common facet F and a function $w_h\in W_h$ we define:

- the jump of w_h as $[\![w_h]\!] := w_h|_{T_1} - w_h|_{T_2}$,
- the average of w_h as $\{\!\{w_h\}\!\} := \frac{1}{2}(w_h|_{T_1} + w_h|_{T_2})$.

On the boundary we define

$$[\![w]\!] = w,$$
$$\{\!\{w\}\!\} = w.$$

Later on, we will use the following result:

Lemma 1.4. *There holds*

$$[\![uv]\!] = \{\!\{u\}\!\}[\![v]\!] + \{\!\{v\}\!\}[\![u]\!]. \tag{1.24}$$

Proof. Direct calculation, see [PE12, p. 123]. □

Now, we can derive the discrete formulation of (1.23). As in the derivation of the continuous problem, we multiply (1.23) with a test function $w_h'\in W_h$ and integrate over the domain D. However, as we are aiming for a DG method, we decompose the domain into the mesh elements and write the integral as a sum over integrals over all mesh elements:

$$\sum_{T\in\mathcal{T}_h}\int_T -\nabla\cdot(\rho(u\otimes u)\nabla w_h)\cdot w_h' \,\mathrm{d}x. \tag{1.25}$$

Applying partial integration on each element T and using formula (1.10) yields

$$\sum_{T\in\mathcal{T}_h}\int_T \rho\partial_u w_h\partial_u w_h' \,\mathrm{d}x + \int_{\partial T}(-\rho\partial_u w_h u)\cdot w_h'\nu \,\mathrm{d}s. \tag{1.26}$$

Now, we take a closer look at the boundary integral. First of all, we switch from the sum over all element boundaries to the sum over all facets in the mesh. Each facet appears twice, as two elements are always connected by one facet. Hence, using the notation $u_\nu = u\cdot\nu$, we can rewrite

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T}(-\rho\partial_u w_h u)\cdot w_h'\nu \,\mathrm{d}s = \sum_{F\in\mathcal{F}_h}\int_F u_\nu[\![-\rho\partial_u w_h w_h']\!] \,\mathrm{d}s. \tag{1.27}$$

Now, applying formula (1.24) yields

$$\sum_{F\in\mathcal{F}_h}\int_F u_\nu(\{\!\{-\rho\partial_u w_h\}\!\}[\![w_h']\!] + \{\!\{w_h'\}\!\}[\![-\rho\partial_u w_h]\!]) \,\mathrm{d}s. \tag{1.28}$$

The latter term vanishes for the true solution w. Hence, it can be dropped without making the formulation inconsistent.

To solve the linear system that arises in the discretization, it is generally preferred to preserve the symmetry of the continuous problem. Hence, we add the term

$$\sum_{F\in\mathcal{F}_h} \langle u_\nu \{\!\{ -\rho\partial_u w_h' \}\!\}, [\![w_h]\!] \rangle_F. \tag{1.29}$$

Note that this term is consistent as $[\![w]\!] = 0$ for the exact solution and on the boundary. By the choice of our method, we allow functions to be discontinuous over element interfaces. These jumps have to be controlled in order to ensure coercivity of the formulation. To achieve this, we introduce another term, called stabilization term. We choose a penalization parameter $\lambda > 0$ and add:

$$\sum_{F\in\mathcal{F}_h} \langle \frac{\rho\lambda}{h} |u_\nu|^2 [\![w_h]\!], [\![w_h']\!] \rangle_F. \tag{1.30}$$

In section 2.2.2 we will see that this term and in particular the choice of the penalization parameter λ is crucial for the coercivity and thus, the well-posedness of the problem. Altogether, we propose the following discrete DG bilinear form

$$\begin{aligned} b_h(w_h, w_h') = \sum_{T\in\mathcal{T}_h} \langle \rho\partial_u w_h, \partial_u w_h' \rangle_T + \sum_{F\in\mathcal{F}_h} \Big\{ & \langle u_\nu \{\!\{ -\rho\partial_u w_h \}\!\}, [\![w_h']\!] \rangle_F \\ & + \langle u_\nu \{\!\{ -\rho\partial_u w_h' \}\!\}, [\![w_h]\!] \rangle_F \\ & + \langle \frac{\rho\lambda}{h} |u_\nu|^2 [\![w_h]\!], [\![w_h']\!] \rangle_F \Big\}. \end{aligned}$$

Then, the discrete DG formulation of problem (1.1) is:
Find $w_h \in W_h = V_h^{k,d}$ such that

$$\mathcal{B}_h(w_h, w_h') = \langle f, w_h' \rangle_D \qquad \forall w_h' \in W_h, \tag{1.31}$$

where

$$\begin{aligned} \mathcal{B}_h(w_h, w_h') = & \sum_{T\in\mathcal{T}_h} \langle \rho w_h, w_h' \rangle_T + \langle \rho\partial_u w_h, \partial_u w_h' \rangle_T \\ & + \sum_{F\in\mathcal{F}_h} \Big\{ \langle u_\nu \{\!\{ -\rho\partial_u w_h \}\!\}, [\![w_h']\!] \rangle_F \\ & \qquad + \langle u_\nu \{\!\{ -\rho\partial_u w_h' \}\!\}, [\![w_h]\!] \rangle_F \\ & \qquad + \langle \frac{\rho\lambda}{h} |u_\nu|^2 [\![w_h]\!], [\![w_h']\!] \rangle_F \Big\}. \end{aligned} \tag{1.32}$$

Remark 1.3 (On non-homogeneous Dirichlet boundary conditions)**.** *In this section we have only considered homogeneous Dirichlet boundary conditions. If we consider boundary conditions of the form*

$$w = g \text{ on } \partial D^{in}, \tag{1.33}$$

the right-hand side of (1.31) has to be modified. We will do this for the example problem for the numerical experiments in (3.2).

Chapter 2

Analysis of the discrete problem

In this chapter, we will present a theoretical analysis of the method derived in the previous chapter. We will follow a classical structure by proving consistency, coercivity and continuity, which yields the well-posedness of the discrete problem. With the help of these results, we will present an a priori error bound and corresponding interpolation results. Finally, we will investigate one crucial tool in the analysis, the inverse inequality and the stabilization mechanism, in more detail.

2.1 Preliminaries

The solution of the discrete problem does not need to be continuous, so we cannot use the $H^1(D)$-norm for the analysis. Instead, we introduce two discrete norms, which allow us to show the well-posedness of the discrete problem.
To show coercivity, we introduce the $\|\cdot\|_\rho$-norm defined through

$$\|w_h\|_\rho^2 := \sum_{T\in\mathcal{T}_h} \left\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 + h^{-1}|u_\nu|^2\|\rho^{\frac{1}{2}}[\![w_h]\!]\|_{\partial T}^2 \right\}, \qquad \forall w_h \in W_h. \quad (2.1)$$

For continuity however, we need a stronger norm:

$$\|w_h\|_{\rho,*}^2 := \|w_h\|_\rho^2 + \sum_{T\in\mathcal{T}_h} h\|\rho^{\frac{1}{2}}\partial_u w_h\|_{\partial T}^2, \qquad \forall w_h \in W_h. \quad (2.2)$$

Both are indeed norms: the triangle inequality and absolute homogeneity are inherited from the L^2-norm and the kernel of both norms is trivial, since the first term ensures that the norm is only zero when $\|w_h\|_T = 0$, which is only the case if $w_h = 0$.

Further, we need a non-standard inverse inequality for the proof of coercivity. As of now, we will assume that this inequality holds. A more detailed discussion including a proof can be found in section 2.5.

Assumption 2.1. *There exists a constant $c_{tr} > 0$ such that for all $w_h \in \mathcal{P}^k(T)$ holds:*

$$h\|\rho^{\frac{1}{2}}\partial_u w_h\|_{\partial T}^2 \le c_{tr}^2\|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2. \quad (2.3)$$

As a consequence of this inequality, the norms defined above are equivalent for $w_h \in W_h$. There holds

$$0 \le \|w_h\|_{\rho,*}^2 - \|w_h\|_\rho^2 = \sum_{T\in\mathcal{T}_h} h\|\rho^{\frac{1}{2}}\partial_u w_h\|_{\partial T}^2 \le \sum_{T\in\mathcal{T}_h} c_{\mathrm{tr}}^2\|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 \le c_{\mathrm{tr}}^2\|w_h\|_\rho^2,$$

which implies that

$$\frac{1}{\sqrt{1+c_{\text{tr}}^2}}\|w_h\|_{\rho,*} \leq \|w_h\|_\rho \leq \|w_h\|_{\rho,*}.$$

This equivalence will allow us to show coercivity in the $\|\cdot\|_\rho$- and continuity in the $\|\cdot\|_{\rho,*}$-norm.

Recall (1.22), i.e. that we denote

$$\mathcal{B}_h(w_h, w_h') = \langle \rho w_h, w_h' \rangle_D + b_h(w_h, w_h'). \tag{2.4}$$

2.2 Well-posedness of the discrete problem

2.2.1 Consistency

A direct consequence of the derivation in chapter 1 is consistency, mainly due to the fact that $[\![w]\!] = 0$ for the exact solution $w \in W$:

Corollary 2.1 (Consistency). *The bilinearform $\mathcal{B}_h(\cdot,\cdot)$ is consistent, i.e. for the exact solution $w \in W$ of (0.11) holds that*

$$\mathcal{B}_h(w, w_h') = \langle f, w_h' \rangle_D \quad \forall w_h' \in W_h \tag{2.5}$$

Let us quickly note that this condition is equivalent to Galerkin orthogonality as

$$\mathcal{B}_h(w - w_h, w_h') = 0 \Leftrightarrow \mathcal{B}_h(w, w_h') = \langle f, w_h' \rangle_D = \mathcal{B}_h(w_h, w_h'), \tag{2.6}$$

for the exact solution w, the discrete solution w_h and for all $w_h' \in W_h$. This result is useful for the proof of a Ceá-type error estimate that can be found in section 2.3.

2.2.2 Coercivity

While consistency is a direct consequence of the derivation of the method, proving the coercivity of the bilinear form $\mathcal{B}_h(\cdot,\cdot)$ requires a bit more effort. In particular, we will apply the Cauchy-Schwarz- and Young's inequalities that can be found in chapter A.

Proposition 2.2 (Coercivity). *For all $w_h \in W_h$ and λ sufficiently large, there holds that*

$$\mathcal{B}_h(w_h, w_h) \geq \frac{1}{2}\|w_h\|_\rho^2 \geq \alpha_{\mathcal{B}_h}\|w_h\|_{\rho,*}^2, \tag{2.7}$$

with $\alpha_{\mathcal{B}_h} \in \mathbb{R}$ independent of the mesh size h.

Proof. The proof relies on the Cauchy-Schwarz and Young's inequalities, cf. (A.7) and (A.9). Furthermore, we require the inverse inequality (2.3). Let $w_h \in W_h$ be arbitrary.

Then, there holds

$$
\begin{aligned}
\mathcal{B}_h(w_h, w_h) &= \sum_{T\in\mathcal{T}_h} \left\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 \right\} \\
&\quad + \sum_{F\in\mathcal{F}_h} \left\{ 2\langle u_\nu \{\!\{ -\rho \partial_u w_h \}\!\}, [\![w_h]\!] \rangle_F + \frac{\lambda}{h} |u_\nu|^2 \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_F^2 \right\} \\
&\geq \sum_{T\in\mathcal{T}_h} \left\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 \right. \\
&\quad \left. - 2\langle u_\nu |\rho \partial_u w_h|, |[\![w_h]\!]| \rangle_{\partial T} + \frac{\lambda}{2h} |u_\nu|^2 \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T}^2 \right\} \\
&\overset{(A.7)}{\geq} \sum_{T\in\mathcal{T}_h} \left\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 \right. \\
&\quad \left. - 2|u_\nu| \|\rho^{\frac{1}{2}} \partial_u w_h\|_{\partial T} \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T} + \frac{\lambda}{2h} |u_\nu|^2 \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T}^2 \right\} \\
&\overset{(2.3)}{\geq} \sum_{T\in\mathcal{T}_h} \left\{ \|w_h\|_T^2 + \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 \right. \\
&\quad \left. - 2|u_\nu| c_{\mathrm{tr}} h^{-\frac{1}{2}} \|\rho^{\frac{1}{2}} \partial_u w_h\|_T \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T} + \frac{\lambda}{2h} |u_\nu|^2 \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T}^2 \right\} \\
&\overset{(A.9)}{\geq} \sum_{T\in\mathcal{T}_h} \left\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 \right. \\
&\quad \left. - \frac{1}{2} \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 - \frac{2c_{\mathrm{tr}}^2}{h} |u_\nu|^2 \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T}^2 + \frac{\lambda}{2h} |u_\nu|^2 \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T}^2 \right\} \\
&\geq \sum_{T\in\mathcal{T}_h} \left\{ \frac{1}{2} \|\rho^{\frac{1}{2}} w_h\|_T^2 + \frac{1}{2} \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 + h^{-1} |u_\nu|^2 (\frac{\lambda}{2} - 2c_{\mathrm{tr}}^2) \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T}^2 \right\} \\
&\geq \min(\frac{1}{2}, \frac{\lambda}{2} - 2c_{\mathrm{tr}}^2) \|w_h\|_\rho^2 \\
&\geq \frac{1}{2} \|w_h\|_\rho^2
\end{aligned}
\tag{2.8}
$$

for $\lambda \geq 1 + 4c_{\mathrm{tr}}^2$. The first inequality stems from the fact that each facet appears twice in the sum over all element boundaries and that we can bound the average, that is half a contribution from one element and its neighbour, by the full contribution from both elements. As the $\|\cdot\|_\rho$- and the $\|\cdot\|_{\rho,*}$-norms are equivalent, the second inequality holds with constant $\alpha_{\mathcal{B}_h} = \frac{1}{2\sqrt{1+c_{\mathrm{tr}}^2}}$. □

Remark 2.1. *While the assumption on the stabilization parameter λ is crucial for coercivity, it can be problematic. We will explore this issue in more detail in section 2.6.*

2.2.3 Continuity

Having shown coercivity, we will now prove that the bilinear form $\mathcal{B}_h(\cdot,\cdot)$ is continuous with respect to the $\|\cdot\|_{\rho,*}$-norm. To this end, we will again make use of the Cauchy-Schwarz inequality.

Proposition 2.3. *For $w, w' \in W_* := W + W_h$ there holds that*

$$
\mathcal{B}_h(w, w') \leq \beta_{\mathcal{B}_h} \|w\|_{\rho,*} \|w'\|_{\rho,*} \tag{2.9}
$$

for $\beta_{\mathcal{B}_h} > 0$ independent of h.

Proof. Similarly to the proof of coercivity, we will first switch from the sum over all facets to the sum over all elements. Then, we will apply both variants of the Cauchy-Schwarz inequality in lemma A.3. For $w, w' \in W + W_h$, there holds:

$$\begin{aligned}
\mathcal{B}_h(w,w') &= \sum_{T\in\mathcal{T}_h} \Big\{ \langle \rho w, w'\rangle_T + \langle \rho^{\frac12}\partial_u w, \rho^{\frac12}\partial_u w'\rangle_T \Big\} + \sum_{F\in\mathcal{F}_h} \Big\{ \langle u_\nu \{\!\{ -\rho\partial_u w \}\!\}, [\![w']\!]\rangle_F \\
&\qquad + \langle u_\nu \{\!\{ -\rho\partial_u w' \}\!\}, [\![w]\!]\rangle_F + \frac{\lambda}{h}|u_\nu|^2 \langle \rho^{\frac12}[\![w]\!], \rho^{\frac12}[\![w']\!]\rangle_F \Big\} \\
&\le \sum_{T\in\mathcal{T}_h} \Big\{ \langle \rho w, w'\rangle_T + \langle \rho^{\frac12}\partial_u w, \rho^{\frac12}\partial_u w'\rangle_T + \langle u_\nu|\rho\partial_u w|, |[\![w']\!]|\rangle_{\partial T} \\
&\qquad + \langle u_\nu|\rho\partial_u w'|, |[\![w]\!]|\rangle_{\partial T} + \frac{\lambda}{2h}|u_\nu|^2 \langle \rho^{\frac12}[\![w]\!], \rho^{\frac12}[\![w']\!]\rangle_{\partial T} \Big\} \\
&\overset{(A.7)}{\le} \sum_{T\in\mathcal{T}_h} \Big\{ \|\rho^{\frac12} w\|_T \|\rho^{\frac12} w'\|_T + \|\rho^{\frac12}\partial_u w\|_T \|\rho^{\frac12}\partial_u w'\|_T \\
&\qquad + |u_\nu| \|\rho^{\frac12}\partial_u w\|_{\partial T} \|\rho^{\frac12}[\![w']\!]\|_{\partial T} + |u_\nu| \|\rho^{\frac12}\partial_u w'\|_{\partial T} \|\rho^{\frac12}[\![w]\!]\|_{\partial T} \\
&\qquad + \frac{\lambda}{2h}|u_\nu|^2 \|\rho^{\frac12}[\![w]\!]\|_{\partial T} \|\rho^{\frac12}[\![w']\!]\|_{\partial T} \Big\} \\
&\overset{(A.8)}{\le} \sum_{T\in\mathcal{T}_h} \Big\{ \Big(\|\rho^{\frac12} w\|_T^2 + \|\rho^{\frac12}\partial_u w\|_T^2 + h\|\rho^{\frac12}\partial_u w\|_{\partial T}^2 \\
&\qquad + h^{-1}|u_\nu|^2 \|\rho^{\frac12}[\![w]\!]\|_{\partial T}^2 + \frac{\lambda}{2h}|u_\nu|^2 \|\rho^{\frac12}[\![w]\!]\|_{\partial T}^2 \Big)^{\frac12} \\
&\qquad \cdot \Big(\|\rho^{\frac12} w'\|_T^2 + \|\rho^{\frac12}\partial_u w'\|_T^2 + h\|\rho^{\frac12}\partial_u w'\|_{\partial T}^2 \\
&\qquad + h^{-1}|u_\nu|^2 \|\rho^{\frac12}[\![w']\!]\|_{\partial T}^2 + \frac{\lambda}{2h}|u_\nu|^2 \|\rho^{\frac12}[\![w']\!]\|_{\partial T}^2 \Big)^{\frac12} \Big\} \\
&\overset{(A.8)}{\le} \Big(\sum_{T\in\mathcal{T}_h} \Big\{ \|\rho^{\frac12} w\|_T^2 + \|\rho^{\frac12}\partial_u w\|_T^2 + h\|\rho^{\frac12}\partial_u w\|_{\partial T}^2 \\
&\qquad + h^{-1}|u_\nu|^2 \|\rho^{\frac12}[\![w]\!]\|_{\partial T}^2 + \frac{\lambda}{2h}|u_\nu|^2 \|\rho^{\frac12}[\![w]\!]\|_{\partial T}^2 \Big\} \Big)^{\frac12} \\
&\qquad \cdot \Big(\sum_{T\in\mathcal{T}_h} \Big\{ \|\rho^{\frac12} w'\|_T^2 + \|\rho^{\frac12}\partial_u w'\|_T^2 + h\|\rho^{\frac12}\partial_u\|_{\partial T}^2 \\
&\qquad + h^{-1}|u_\nu|^2 \|\rho^{\frac12}[\![w']\!]\|_{\partial T}^2 + \frac{\lambda}{2h}|u_\nu|^2 \|\rho^{\frac12}[\![w']\!]\|_{\partial T}^2 \Big\} \Big)^{\frac12} \\
&\le \beta_{\mathcal{B}_h} \|w\|_{\rho,*} \|w'\|_{\rho,*}
\end{aligned} \tag{2.10}$$

for $\beta_{\mathcal{B}_h} = 1 + \frac{\lambda}{2}$. □

With these results, we can show the well-posedness of the discrete problem:

Theorem 2.4 (Well-posedness of the discrete problem)**.** *There exists a unique solution to the discrete problem: Find $w_h \in W_h$ such that*

$$\mathcal{B}_h(w_h, w_h') = \langle f, w_h'\rangle_D \qquad \forall w_h' \in W_h. \tag{2.11}$$

Further, this solution admits the following stability estimate

$$\|w_h\|_\rho \le \frac{1}{\alpha_{\mathcal{B}_h}\underline{\rho}} \|f\|_D. \tag{2.12}$$

Proof. We have shown that the discrete bilinear form $\mathcal{B}_h(\cdot,\cdot)$ is coercive and continuous. Furthermore, the right-hand side of the discrete problem (1.31) is not only linear, but also bounded due to the Cauchy-Schwarz inequality. As such, we can apply the Lax-Milgram lemma [Eva10, Chapter 6 Theorem 1] to conclude the existence of a unique solution $w_h \in W_h$ to the discrete problem.
The stability of the solution follows directly:

$$\|w_h\|_\rho \leq \alpha_{\mathcal{B}_h} \underbrace{\mathcal{B}_h(w_h, w_h)}_{\langle f, w_h\rangle_D} \leq \alpha_{\mathcal{B}_h}\|f\|_D \underbrace{\|w_h\|_D}_{\leq \underline{\rho}^{-1}\|w_h\|_\rho} \leq \frac{1}{\alpha_{\mathcal{B}_h}\underline{\rho}}\|f\|_D. \tag{2.13}$$

□

2.3 A priori error estimates

In theorem 2.4 we argued that coercivity and continuity imply the existence of a unique solution $w_h \in W_h$. Now, we want to derive a Ceá-type estimate on the approximation error $\|w - w_h\|_\rho$, where $w \in W$ denotes the exact solution to the problem.

Proposition 2.5. *Let $w \in W$ be the exact solution to (0.11) and $w_h \in W_h$ be the discrete solution to the problem: Find $w_h \in W_h$ such that*

$$\mathcal{B}_h(w_h, w_h') = \langle f, w_h'\rangle_D \quad \forall w_h' \in W_h.$$

Then, there holds

$$\|w - w_h\|_\rho \leq \|w - w_h\|_{\rho,*} \leq C \inf_{v_h \in W_h} \|w - v_h\|_{\rho,*}, \tag{2.14}$$

with a constant C independent of h.

Proof. The first inequality simply holds true by construction of the $\|\cdot\|_\rho$- and the $\|\cdot\|_{\rho,*}$-norm. Further, for any $v_h \in W_h$ the triangle inequality yields

$$\|w - w_h\|_{\rho,*} \leq \|w - v_h\|_{\rho,*} + \|v_h - w_h\|_{\rho,*}. \tag{2.15}$$

In the following, we will bound the discrete error $\|v_h - w_h\|_{\rho,*}$ by the approximation error $\|w - v_h\|_{\rho,*}$ using the coercivity, the Galerkin orthogonality and the continuity of the bilinear form $\mathcal{B}_h(\cdot,\cdot)$. We have

$$\begin{aligned}
\|v_h - w_h\|_{\rho,*}^2 &\overset{(2.7)}{\leq} \frac{1}{\alpha_{\mathcal{B}_h}}\mathcal{B}_h(v_h - w_h, v_h - w_h) \\
&\overset{(2.6)}{=} \frac{1}{\alpha_{\mathcal{B}_h}}\Big(\mathcal{B}_h(v_h - w, v_h - w_h) + \underbrace{\mathcal{B}_h(w - w_h, v_h - w_h)}_{=0, \text{ as } v_h - w_h \in W_h}\Big) \\
&\overset{(2.9)}{\leq} \frac{\beta_{\mathcal{B}_h}}{\alpha_{\mathcal{B}_h}}\|v_h - w\|_{\rho,*}\|v_h - w_h\|_{\rho,*}.
\end{aligned} \tag{2.16}$$

Dividing by $\|v_h - w_h\|_{\rho,*}$ yields

$$\|v_h - w_h\|_{\rho,*} \leq \frac{\beta_{\mathcal{B}_h}}{\alpha_{\mathcal{B}_h}}\|w - v_h\|_{\rho,*}. \tag{2.17}$$

Inserting this estimate into (2.15) gives us

$$\|w - w_h\|_{\rho,*} \leq \Big(1 + \frac{\beta_{\mathcal{B}_h}}{\alpha_{\mathcal{B}_h}}\Big)\|w - v_h\|_{\rho,*}. \tag{2.18}$$

As $v_h \in W_h$ was chosen arbitrary, we can take the infimum over all $v_h \in W_h$ and conclude

$$\|w - w_h\|_{\rho,*} \leq C \inf_{v_h \in W_h} \|w - v_h\|_{\rho,*}, \tag{2.19}$$

where $C = 1 + \frac{\beta_{\mathcal{B}_h}}{\alpha_{\mathcal{B}_h}}$. We note that the constants $\alpha_{\mathcal{B}_h}$ and $\beta_{\mathcal{B}_h}$ only depend on the stabilization parameter λ and the constant c_{tr} from the inverse inequality. □

2.4 Interpolation estimates

The following section is dedicated to a typical bound on the approximation error in the $H^l(D)$-norm, $l \geq 2$. To this end, we will assume that the exact solution is sufficiently regular, i.e. $w \in W \cap H^l(D)$ and apply standard interpolation results.
Let us introduce the L^2-projection $\Pi_h : L^2(D) \to W_h$ defined through

$$\langle \Pi_h w, v_h \rangle_D = \langle w, v_h \rangle_D \qquad \forall v_h \in W_h. \tag{2.20}$$

When restricting to a mesh element $T \in \mathcal{T}_h$,we obtain

$$\langle \Pi_h w|_T, v_h \rangle_T = \langle w, v_h \rangle_T \qquad \forall v_h \in W_h. \tag{2.21}$$

Hence, we can make sense of the L^2-projection element-wise.
For the proof of an H^l-norm error estimate, we will make use of the following standard interpolation results:

Lemma 2.6. *Let $v \in H^s(D)$ for $s \in \{2, \dots k+1\}$ where k is the polynomial degree. Further, let $m \leq s$. Then, there holds*

$$|v - \Pi_h v|_{H^m(T)} \lesssim h^{s-m}|v|_{H^s(T)}; \tag{2.22}$$

$$\|v - \Pi_h v\|_F \lesssim h^{s-\frac{1}{2}}|v|_{H^s(T)}; \tag{2.23}$$

$$\|\nabla(v - \Pi_h v)\|_F \lesssim h^{s-\frac{3}{2}}|v|_{H^s(T)}. \tag{2.24}$$

Proof. The proof follows standard arguments and can be found in [PE12, Lemmata 1.58 and 1.59]. □

Now, we can show the following result, which will give us a standard convergence rate.

Proposition 2.7 (H^l-norm error estimate)**.** *Let $w \in H^l(D)$ for $2 \leq l \leq k+1$. Then, there holds*

$$\inf_{v_h \in W_h} \|w - v_h\|_{\rho,*} \lesssim \bar{\rho}^{\frac{1}{2}} h^{l-1} \|w\|_{H^l(D)}, \tag{2.25}$$

where the implied constants do not depend on ρ.

Proof. There holds for any arbitrary $v_h \in W_h$ that

$$\inf_{v_h \in W_h} \|w - v_h\|_{\rho,*} \leq \|w - v_h\|_{\rho,*}.$$

Denoting $\tilde{w} = w - v_h$, there is

$$\|\tilde{w}\|^2_{\rho,*} = \sum_{T\in\mathcal{T}_h} \left\{ \|\rho^{\frac{1}{2}}\tilde{w}\|^2_T + \|\rho^{\frac{1}{2}}\partial_u\tilde{w}\|^2_T + h^{-1}|u_\nu|^2\|\rho^{\frac{1}{2}}[\![\tilde{w}]\!]\|^2_{\partial T} + h\|\rho^{\frac{1}{2}}\partial_u\tilde{w}\|^2_{\partial T} \right\} \tag{2.26}$$

In order to get the desired result, we will choose $v_h = \Pi_h w$ and use the interpolation results from lemma 2.6 to bound each term in the $\|\cdot\|_{\rho,*}$-norm.
For the first one, (2.22) with $m = 0$ directly yields

$$\|\rho^{\frac{1}{2}}\tilde{w}\|^2_T \lesssim \bar{\rho}h^{2l}\|w\|^2_{H^l(T)}. \tag{2.27}$$

The same result can be applied with $m = 1$ to bound the second term in the following way

$$\|\rho^{\frac{1}{2}}\partial_u\tilde{w}\|^2_T \leq \bar{\rho}\|u\|^2_\infty\|\nabla\tilde{w}\|^2_T \lesssim \bar{\rho}\|u\|^2_\infty h^{2l-2}\|w\|^2_{H^l(T)}. \tag{2.28}$$

Here, we use that by assumption (1.1) there exists a constant $\overline{\rho}$ such that $\rho(x) \leq \overline{\rho}$ for all $x \in D$. For the third term, we simply replace the jump of $\tilde{w}$ by just $\tilde{w}$ on the boundary and use (2.23) to get

$$\|\rho^{\frac{1}{2}}[\![\tilde{w}]\!]\|^2_{\partial T} \lesssim \bar{\rho}\|\tilde{w}\|^2_{\partial T} \lesssim \bar{\rho}h^{2l-1}\|w\|^2_{H^l(T)}. \tag{2.29}$$

Finally, using (2.24) the last term can be bounded in the following way:

$$\|\rho^{\frac{1}{2}}\partial_u\tilde{w}\|^2_{\partial T} \lesssim \bar{\rho}\|u\|^2_\infty\|\nabla\tilde{w}\|^2_{\partial T} \lesssim \bar{\rho}\|u\|^2_\infty h^{2l-3}\|w\|^2_{H^l(T)}. \tag{2.30}$$

Putting this all together, we obtain

$$\|\tilde{w}\|^2_{\rho,*} \lesssim \sum_{T\in\mathcal{T}_h} \left\{ \bar{\rho}h^{2l}\|w\|^2_{H^l(T)} + 3\bar{\rho}\|u\|^2_\infty h^{2l-2}\|w\|^2_{H^l(T)} \right\} \tag{2.31}$$

Taking the square root, and using the fact that $h^l \lesssim h^{l-1}$, we arrive at

$$\|\tilde{w}\|_{\rho,*} \lesssim \bar{\rho}^{\frac{1}{2}}h^{l-1}\|w\|_{H^l(D)}. \tag{2.32}$$

□

By combining this result with (2.14,) we get that

$$\|w - w_h\|_{\rho,*} \lesssim \bar{\rho}^{\frac{1}{2}}h^{l-1}\|w\|_{H^l(D)} \tag{2.33}$$

for a sufficiently smooth solution $w \in H^l(D)$, $l \geq 2$. This result implies that for $w \in H^{k+1}(D)$, the discrete solution converges to the exact solution with order k.

Remark 2.2 (On the regularity assumptions on w)**.** *Functions in the space W do not have to fulfil the regularity assumptions. The differential operator ∂_u only acts along the velocity field, so that functions in W might not have higher regularity than L^2 in directions orthogonal to the velocity field. As such, even as functions $w \in W$ fulfil $\rho\partial_u w \in L^2(D)$, they might not have enough regularity to apply proposition 2.7.*

Remark 2.3 (On L^2-error estimates)**.** *Using a broken version of the Poincaré-inequality, we can show that for $w \in W \cap H^{k+1}(D)$*

$$\|w - w_h\|_D \leq \underline{\rho}^{-\frac{1}{2}}\|w - w_h\|_{\rho,*} \lesssim \underline{\rho}^{-\frac{1}{2}}\bar{\rho}^{\frac{1}{2}}h^k\|w\|_{H^{k+1}(D)}. \tag{2.34}$$

However, this estimate is suboptimal, and usually it can be improved. With the help of a duality argument, one can show that

$$\|w - w_h\|_D \lesssim h\|w - w_h\|_{\rho,*}. \tag{2.35}$$

To do this, one requires the problem to be L^2-H^2-regular. This means that for the solution of the dual problem, which is in our case the same as the primal problem due to symmetry, there should hold that $\|v\|_{H^2(D)} \lesssim \|\tilde{f}\|_D$. Usually, second order derivatives have a smoothing effect that allows to apply such an argumentation. However, similar as in remark 2.2, where we argued that functions in W do not have to fulfil higher regularity assumption, this is not the case here. As the diffusion operator only acts along the velocity field, there is no smoothing effect in directions orthogonal to the velocity field. Hence, we cannot apply a duality argument yielding improved rates of convergence in the L^2-norm.

2.5 On the inverse inequality

Having concluded the standard error analysis, we now want to dive deeper into some specific issues, starting with the inverse inequality. Our non-standard inverse inequality (2.3) is a crucial ingredient in the proof of coercivity, as we need to choose the stabilization parameter λ sufficiently large in relation to the constant c_{tr} arising there.
Consequently, the constant in the a-priori error estimate depends on this constant as well:

$$C = 1 + \frac{\beta_{\mathcal{B}_h}}{\alpha_{\mathcal{B}_h}} = 1 + \frac{1 + \frac{\lambda}{2}}{\frac{1}{2\sqrt{1+c_{\text{tr}}^2}}} = 1 + (2+\lambda)\sqrt{1+c_{\text{tr}}^2}. \tag{2.36}$$

To ensure stability, we need to choose $\lambda \geq 1 + 4c_{\text{tr}}^2$ and hence

$$C \geq 1 + (3 + 4c_{\text{tr}}^2)\sqrt{1+c_{\text{tr}}^2}. \tag{2.37}$$

This expression gets large, if c_{tr} is large, which means that the constant in the a-priori error estimate increases with the constant in the inverse inequality. Hence, it seems to be reasonable to investigate how the constant c_{tr} behaves in dependence of the density, the polynomial degree and the shape regularity of the mesh $\mathcal{T}_h$.
To this end, we will prove assumption 2.1.

Proposition 2.8. *For any $w_h \in \mathcal{P}^k(T)$ there holds*

$$h\|\rho^{\frac{1}{2}}\partial_u w_h\|_{\partial T}^2 \leq c_{tr}^2\|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2,$$

where c_{tr} depends on the shape regularity, the polynomial degree k and on the density ρ, specifically on $\frac{\max \rho|_T}{\min \rho|_T}$.

Proof. To prove the inverse inequality, we will make use of the results given in section A.1. Let $T \in \mathcal{T}_h$ be an arbitrary mesh element and $F \in \partial T$ one of its facets. According to Lemma A.1, there exists a bijective affine mapping $\Phi : \widehat{T} \to T$ such that $\Phi(\widehat{F}) = F$ for a facet $\widehat{F}$ of the reference simplex $\widehat{T}$ in $\mathbb{R}^d$. For a polynomial $w_T \in \mathcal{P}^k(T)$ we denote $\widehat{w_T} = w_T \circ \Phi$. There holds the following transformation rule

$$\int_F w_T \ \mathrm{d}s = \int_{\widehat{F}} J\widehat{w_T} \ \mathrm{d}s, \tag{2.38}$$

where $J = \sqrt{\det(D\Phi D\Phi^T)}$. Especially, we have that

$$|J| \lesssim h_F^{d-1} \lesssim h_T^{d-1}, \tag{2.39}$$

where h_F is the length of the facet F, h_T the size of an element T and d the dimension. Due to the assumption of quasi-uniformity, we have that $h_T \simeq h$. In the following, we will denote $\widehat{\rho} = \rho \circ \Phi$, $\widehat{u} = D\Phi^{-1}(u \circ \Phi)$ and

$$\widehat{\partial}_{\widehat{u}} \cdot = (\widehat{u} \cdot \widehat{\nabla}) \cdot = (u \circ \Phi)^T (D\Phi^{-T})(D\Phi)^T (\nabla \circ \Phi) \cdot .$$

Then, there holds

$$\begin{aligned} \|\rho^{\frac{1}{2}} \partial_u w_T\|_F^2 &= \int_F (\rho^{\frac{1}{2}} \partial_u w_T)^2 \ \mathrm{d}s \overset{(2.38)}{=} \int_{\widehat{F}} J \cdot (\rho^{\frac{1}{2}} \partial_u w_T)^2 \circ \Phi \ \mathrm{d}s \\ &= \int_{\widehat{F}} J \cdot (\widehat{\rho}^{\frac{1}{2}} \widehat{\partial}_{\widehat{u}} \widehat{w_T})^2 \ \mathrm{d}s \lesssim h_T^{d-1} \|\widehat{\rho}^{\frac{1}{2}} \widehat{\partial}_{\widehat{u}} \widehat{w_T}\|_{\widehat{F}}^2 \end{aligned} \tag{2.40}$$

Now, we want to apply Lemma A.2. In particular, we use equation (A.5), which states that

$$|\widehat{u}|_{H^m(\widehat{T})} \lesssim \left(\frac{h_T}{\rho_{\widehat{T}}}\right)^m \rho_T^{-\frac{d}{2}} |u|_{H^m(T)}.$$

For ease of presentation, we will denote $\widehat{f} := \widehat{\rho}^{\frac{1}{2}} \widehat{\partial}_{\widehat{u}} \widehat{w_T}$ and $f := \rho^{\frac{1}{2}} \partial_u w_T$. There holds:

$$\begin{aligned} h_T^{d-1} \|\widehat{f}\|_{\widehat{F}}^2 \lesssim h_T^{d-1} \|\widehat{f}\|_{\partial\widehat{T}}^2 &\lesssim h_T^{d-1} \|\widehat{f}\|_{H^1(\widehat{T})}^2 \\ &\simeq h_T^{d-1} \left(\|\widehat{f}\|_{\widehat{T}}^2 + |\widehat{f}|_{H^1(\widehat{T})}^2 \right) \\ &\overset{(A.5)}{\lesssim} h_T^{d-1} \left(\rho_T^{-d} \|f\|_T^2 + \left(\frac{h_T}{\rho_{\widehat{T}}}\right)^2 \rho_T^{-d} |f|_{H^1(T)}^2 \right) \end{aligned} \tag{2.41}$$

The second step uses the equivalence of norms on finite dimensional spaces. We note that the constants implied in the $\lesssim$ depend on ρ, in particular on $\left(\frac{\max(rho|_T}{\min \rho|_T}\right)$.
In the last step, we apply equations (A.5) with $m = 0$ for the first term and with $m = 1$ for the second term. For the next steps, we note that by definition $\frac{1}{\rho_T} = \frac{\sigma_T}{h_T}$. Due to shape regularity, we have that $\sigma_T \leq \sigma$ for some constant σ. Hence, we can pull the σ into the constant implied by $\lesssim$. As we are on the reference element, we can do the same for $\frac{1}{\rho_{\widehat{T}}}$. Hence, we get

$$\begin{aligned} & h_T^{d-1} \left(\rho_T^{-d} \|f\|_T^2 + \left(\frac{h_T}{\rho_{\widehat{T}}}\right)^2 \rho_T^{-d} |f|_{H^1(T)}^2 \right) \\ & \lesssim h_T^{d-1} \left(\left(\frac{\sigma_T}{h_T}\right)^d \|f\|_T^2 + \left(\frac{h_T}{\rho_{\widehat{T}}}\right)^2 \left(\frac{\sigma_T}{h_T}\right)^d |f|_{H^1(T)}^2 \right) \\ & \lesssim h_T^{-1} \left(\|f\|_T^2 + h_T^2 |f|_{H^1(T)}^2 \right). \end{aligned} \tag{2.42}$$

To conclude, we have to show that

$$h_T^{-1} \left(\|f\|_T^2 + h_T^2 |f|_{H^1(T)}^2 \right) \lesssim h_T^{-1} \|f\|_T^2.$$

To this end, we will use the second result from lemma A.2, equation (A.6), which states

$$|u|_{H^m(T)} \lesssim \left(\frac{h_{\widehat{T}}}{\rho_T}\right)^m h_T^{\frac{d}{2}} |\widehat{u}|_{H^m(\widehat{T})}.$$

Again, we use equivalence of norms and that we can bound $\sigma_T \leq \sigma$. Furthermore, note that $h_{\widehat{T}}$ can be bounded as well. There holds

$$\begin{aligned}\|f\|_{H^1(T)} &\overset{(A.6)}{\lesssim} \Big(\frac{h_{\widehat{T}}}{\rho_T}\Big) h_T^{\frac{d}{2}} \|f\|_{H^1(T)} \lesssim \Big(\frac{h_{\widehat{T}}}{\rho_T}\Big) h_T^{\frac{d}{2}} \|\widehat{f}\|_{\widehat{T}} \\ &\lesssim h_{\widehat{T}} \Big(\frac{\sigma_T}{h_T}\Big) h_T^{\frac{d}{2}} \|\widehat{f}\|_{\widehat{T}} \overset{(A.5)}{\lesssim} h_{\widehat{T}} \Big(\frac{\sigma_T}{h_T}\Big) h_T^{\frac{d}{2}} \underbrace{\rho_T^{\frac{-d}{2}}}_{=\frac{h_T}{\sigma_T}} \|f\|_T \\ &\lesssim h_T^{-1} \underbrace{h_T^{\frac{d}{2}} h_T^{\frac{-d}{2}}}_{=1} \|f\|_T \lesssim h_T^{-1} \|f\|_T.\end{aligned} \tag{2.43}$$

Since we assume quasi uniformity, we have that $h_T \simeq h$ and thus

$$h\|\rho^{\frac{1}{2}} \partial_u w_T\|_F^2 \leq c_{\text{tr}}^2 \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2. \tag{2.44}$$

Summing over all $F \in \partial T$ yields (2.3).
Note that the constant c_{tr} usually depends on the shape regularity and the polynomial degree k [WH03]. In particular, we have that $c_{\text{tr}} \sim k^2$. Furthermore, as ρ can be non-constant, we have a dependence on ρ stemming from the equivalence of norms as well. □

2.6 On refined penalization

Until now, we have noted mainly two facts about the stabilization term and the corresponding choice of a penalization parameter λ:

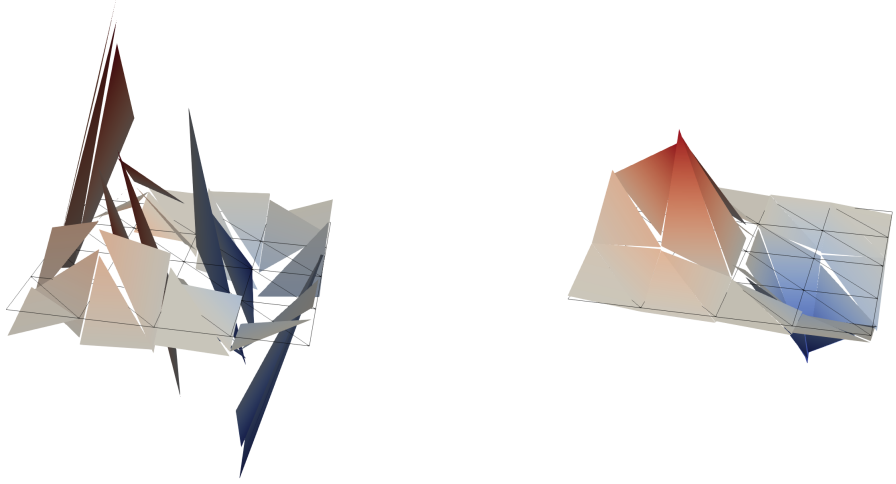
- the term exerts control over the jumps across element interfaces,
- λ has to be chosen large enough, so that the problem is well-posed.

Figure 2.1, which shows a discrete solution of the example problem from chapter 3 for two different penalization parameters, visualizes the first effect. Regarding the second point, we know, from the proof of proposition 2.2, that choosing $\lambda \geq 1 + 4c_{\text{tr}}^2$ is sufficient to ensure the coercivity of the problem. Note, however, that this analysis might not be sharp.

One major drawback of symmetric interior penalty methods in general is caused by the stabilization term: the condition number of the matrices arising in the linear systems depend on the penalization parameter [Cas02]. As such, choosing a suitable penalization parameter might not be a trivial task. On the one hand, λ has to be large enough to ensure coercivity, but on the other hand, choosing λ too large will have a negative impact on the condition number.
In practice, one often tries out different order of magnitudes, that is $\lambda = 1, 10, 100, \ldots$, and compares how well the method performs. If one encounters issues with the choice of the penalization parameter, several remedies can be applied. In the following, we will present two possibilities to fine-tune the usage of the penalization: a generalized eigenvalue problem and a Bassi-Rebay type stabilization.

2.6.1 Generalized eigenvalue problem

One straightforward approach is to make sure that we choose the stabilization parameter as close to the necessary minimum as possible, cf. [Leh21, Section 14.1.6]. Formally, this



(a) Discrete solution with $\lambda = 3$ (b) Discrete solution with $\lambda = 20$

Figure 2.1: Solution of the example problem (3.3) for $k = 1$ with $\lambda = 3$ (left) and $\lambda = 20$ (right).

means that we have to solve the following generalized element-local eigenvalue problem: Find $w_T \in \mathcal{P}^k(T)\backslash\mathbb{R}$ and $\mu \in \mathbb{R}$ such that

$$B(w_T, v_T) = \mu C(w_T, v_T) \qquad \forall v_T \in \mathcal{P}^k(T), \tag{2.45}$$

where $B(w, v) := h\langle \rho^{\frac{1}{2}} \partial_u w, \rho^{\frac{1}{2}} \partial_u v\rangle_{\partial T}$ and $C(w, v) = \langle \rho^{\frac{1}{2}} \partial_u w, \rho^{\frac{1}{2}} \partial_u v\rangle_T$.
Then, with defining

$$\lambda_T^* = 8 \cdot \mu_{\max} = 8 \cdot \max_{w_T \in \mathcal{P}^k(T)\backslash\mathbb{R}} \frac{B(w_T, w_T)}{C(w_T, w_T)} = 8 \cdot \max_{w_T \in \mathcal{P}^k(T)\backslash\mathbb{R}} \frac{\|\rho^{\frac{1}{2}} \partial_u w_T\|_{\partial T}^2 h}{\|\rho^{\frac{1}{2}} \partial_u w_T\|_T^2} \tag{2.46}$$

and choosing $\lambda_T \geq \lambda_T^*$, we automatically get that

$$\begin{aligned} \langle u_\nu | \rho \partial_u w_h |, |[\![w_h]\!]| \rangle_{\partial T} &\overset{(A.7)}{\leq} |u_\nu| \|\rho^{\frac{1}{2}} \partial_u w_h \cdot \nu\|_{\partial T} \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T} \\ &\leq |u_\nu| \|\rho^{\frac{1}{2}} \partial_u w_h\|_T \sqrt{\frac{\rho \frac{\lambda_T}{8}}{h}} \|[\![w_h]\!]\|_{\partial T}. \end{aligned} \tag{2.47}$$

With this bound, we can circumvent the use of the inverse inequality in the proof of coercivity and thus get coercivity independent of the constant from the inverse inequality.

We define the following discrete norm depending on λ_T:

$$\|w_h\|_{\rho,\lambda_T}^2 = \sum_{T \in \mathcal{T}_h} \left\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}} \partial_u w_h\|_T^2 + h^{-1} |u_\nu|^2 \frac{\lambda_T}{2} \|\rho^{\frac{1}{2}} [\![w_h]\!]\|_{\partial T}^2 \right\}. \tag{2.48}$$

Now, we can show coercivity of $\mathcal{B}_h(\cdot, \cdot)$ with respect to this modified norm for all $\lambda_T \geq \lambda_T^*$.

Lemma 2.9. *There holds for all $\lambda_T \geq \lambda_T^*$ that*

$$\mathcal{B}_h(w_h, w_h) \geq \frac{1}{2} \|w_h\|_{\rho,\lambda_T}^2. \tag{2.49}$$

Proof. Following a similar argumentation as in the proof of proposition 2.2 with the additional use of (2.47) we get that

$$\begin{aligned}
\mathcal{B}_h(w_h, w_h) &\geq \sum_{T\in\mathcal{T}_h} \Big\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 \\
&\qquad - 2\langle u_\nu|\rho\partial_u w_h|, |[\![w_h]\!]|\rangle_{\partial T} + \frac{\lambda_T}{2h}|u_\nu|^2\|\rho^{\frac{1}{2}}[\![w_h]\!]\|_{\partial T}^2\Big\} \\
&\overset{(2.47)}{\geq} \sum_{T\in\mathcal{T}_h} \Big\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 \\
&\qquad - 2|u_\nu|\|\rho^{\frac{1}{2}}\partial_u w_h\|_T\|\sqrt{\frac{\rho\frac{\lambda_T}{8}}{h}}[\![w_h]\!]\|_{\partial T} + \frac{\lambda_T}{2h}|u_\nu|^2\|\rho^{\frac{1}{2}}[\![w_h]\!]\|_{\partial T}^2\Big\} \\
&\overset{(A.9)}{\geq} \sum_{T\in\mathcal{T}_h} \Big\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 \\
&\qquad - \frac{1}{2}\|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 - 2|u_\nu|^2\|\sqrt{\frac{\rho\frac{\lambda_T}{8}}{h}}[\![w_h]\!]\|_{\partial T}^2 + \frac{\lambda_T}{2h}|u_\nu|^2\|\rho^{\frac{1}{2}}[\![w_h]\!]\|_{\partial T}^2\Big\} \\
&\geq \sum_{T\in\mathcal{T}_h} \Big\{ \frac{1}{2}\|\rho^{\frac{1}{2}} w_h\|_T^2 + \frac{1}{2}\|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 \\
&\qquad + h^{-1}|u_\nu|^2 \underbrace{\Big(-\frac{2\lambda_T}{8} + \frac{\lambda_T}{2}\Big)}_{=\frac{1}{4}\lambda_T} \|\rho^{\frac{1}{2}}[\![w_h]\!]\|_{\partial T}^2\Big\} \\
&\geq \frac{1}{2}\|w_h\|_{\rho,\lambda_T}^2.
\end{aligned} \tag{2.50}$$

□

Let us stress, again, that this result does not depend directly on the constant from the inverse inequality. Further, we note that we can modify the norm for continuity in a similar way:

$$\|w_h\|_{\rho,\lambda_T,*}^2 = \|w_h\|_{\rho,\lambda_T}^2 + \sum_{T\in\mathcal{T}_h} h\|\rho^{\frac{1}{2}}\partial_u w_h\|_{\partial T}^2, \tag{2.51}$$

and proceed analogously with an a-priori error analysis as before.

Remark 2.4. *The prefactor in the definition of λ_T^* can be chosen differently. Let $m > 4$ and $\lambda_T^* = m \cdot \mu_{\max}$. Then the estimate (2.47) reads as follows:*

$$\langle u_\nu|\rho\partial_u w_h|, |[\![w_h]\!]|\rangle_{\partial T} \leq |u_\nu|\|\rho^{\frac{1}{2}}\partial_u w_h\|_T\|\sqrt{\frac{\rho\frac{\lambda_T}{m}}{h}}[\![w_h]\!]\|_{\partial T}.$$

We can than show coercivity as above simply by modifying the norm:

$$\|w_h\|_{\rho,\lambda_T;m}^2 = \sum_{T\in\mathcal{T}_h} \Big\{ \|\rho^{\frac{1}{2}} w_h\|_T^2 + \|\rho^{\frac{1}{2}}\partial_u w_h\|_T^2 + h^{-1}|u_\nu|^2\frac{(m-4)\lambda_T}{m}\|\rho^{\frac{1}{2}}[\![w_h]\!]\|_{\partial T}^2\Big\}.$$

However, the coercivity constant may become smaller than $\frac{1}{2}$.

2.6.2 Bassi-Rebay type stabilization

While the main idea of the generalized eigenvalue problem is choosing λ as small as possible, but still large enough to ensure coercivity, there are more sophisticated approaches. One of them is the Bassi-Rebay (BR) stabilization introduced by Bassi and Rebay [BR97], which we will briefly derive for our problem in this section.
To avoid using the inverse inequality directly, the BR-stabilization introduces a lifting operator, allowing us to write the facets integrals as volume integrals. For $w_h' \in L^2(\mathcal{F}_h)$ we define such a lifting operator $r : L^2(\mathcal{F}_h) \to \left[V_h^{k,d}\right]^d$ through:

$$\sum_{T\in\mathcal{T}_h} \langle \rho u \cdot \tau_h, u \cdot r(w_h')\rangle_T = \sum_{F\in\mathcal{F}_h} \langle u_\nu \{\!\{ \rho u \cdot \tau_h \}\!\}, w_h'\rangle_F \qquad \forall \tau_h \in \left[V_h^{k,d}\right]^d. \tag{2.52}$$

Without the stabilization term, the bilinear form $b_h(\cdot,\cdot)$ reads as

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} \langle \rho \partial_u w_h, \partial_u w_h'\rangle_T + \sum_{F\in\mathcal{F}_h} \Big\{ &\langle u_\nu \{\!\{ -\rho \partial_u w_h \}\!\}, [\![w_h']\!]\rangle_F \\ &+ \langle u_\nu \{\!\{ -\rho \partial_u w_h' \}\!\}, [\![w_h]\!]\rangle_F \Big\}.\end{aligned} \tag{2.53}$$

The lifting operator allows us to rewrite the facet terms as integrals over the elements

$$-\sum_{F\in\mathcal{F}_h} \langle u_\nu \{\!\{ \rho \partial_u w_h \}\!\}, [\![w_h']\!]\rangle_F \overset{\tau_h = \nabla w_h}{=} -\sum_{T\in\mathcal{T}_h} \langle \rho \underbrace{u \cdot \nabla w_h}_{=\partial_u w_h}, u \cdot r([\![w_h']\!])\rangle_T. \tag{2.54}$$

Hence, (2.53) becomes

$$\begin{aligned}&\sum_{T\in\mathcal{T}_h} \langle \rho \partial_u w_h, \partial_u w_h'\rangle_T - \langle \rho \partial_u w_h, u \cdot r([\![w_h']\!])\rangle_T - \langle \rho \partial_u w_h', u \cdot r([\![w_h]\!])\rangle_T \\ &= \sum_{T\in\mathcal{T}_h} \langle \rho(\partial_u w_h - u \cdot r([\![w_h]\!])), \partial_u w_h' - u \cdot r([\![w_h']\!])\rangle_T - \langle \rho u \cdot r([\![w_h]\!]), u \cdot r([\![w_h']\!])\rangle_T\end{aligned} \tag{2.55}$$

To avoid stability issues, we compensate for the last term. This yields

$$\begin{aligned}b_h^{BR}(w_h, w_h') = \sum_{T\in\mathcal{T}_h} \Big\{ &\langle \rho(\partial_u w_h - u \cdot r([\![w_h]\!])), \partial_u w_h' - u \cdot r([\![w_h']\!])\rangle_T \\ &+ \eta \langle \rho u \cdot r([\![w_h]\!]), u \cdot r([\![w_h']\!])\rangle_T \Big\},\end{aligned} \tag{2.56}$$

for a user dependent parameter $\eta > 1$ that assures non-negativity. Altogether, a BR-stabilized formulation of (1.31) reads as

$$\mathcal{B}^{BR}(w_h, w_h') = \sum_{T\in\mathcal{T}_h} \langle \rho w_h, w_h'\rangle_T + b_h^{BR}(w_h, w_h'). \tag{2.57}$$

Defining a suitable norm, we can show coercivity in the usual way. We omit the details and refer to [PE12, Section 5.3.2]. Let us stress again, that we avoid using the inverse inequality directly and do not need to choose λ large enough to ensure coercivity. As such, we can avoid the conditioning issues described above.
Note, that in general both approaches, a generalized eigenvalue problem and a BR-stabilization, are also advantageous for large ratios $\frac{\max \rho|_T}{\min \rho|_T}$ as the penalization parameter has to be adjusted to ρ.

Remark 2.5 (Computational costs)**.** *It has to be mentioned that both modifications come with additional computational costs. In the first case, they are associated with solving the generalized eigenvalue problems. However, due to their local character, these computations are of small dimension and can be parallelized. For the second modification, the implementation of the lifting operator causes the additional computational costs.*

Chapter 3

Numerical experiments

In this chapter, we will investigate how the method performs numerically. We will consider the influence that the mesh has on the convergence rates in the L^2-norm, if the numerically observed convergence rates in the W-norm agree with the result from proposition 2.7 and how the penalization parameter influences the condition number. Furthermore, we use a numerical example to show that the volume term is important for well-posedness of the considered model problem. Finally, we test the method with a non-constant density ρ.
Note that we refer repeatedly to the appendix, in particular to chapter B where the code is display and to chapter C where some additional figures and tables can be found.

3.1 Description of the example problem

The example problem that we want to consider in this section is two dimensional. For the geometry we choose a square $D = [-1, 1]^2 \subset \mathbb{R}^2$. We want to approximate the following exact solution:

$$w := \exp(-6((x + 0.5)^2 + y^2)) - \exp(-6((x - 0.5)^2 + y^2)). \tag{3.1}$$

This function is smooth and has two Gaussian bumps. It is displayed in figure 3.1. We note that this exact solution does not have a physical meaning, but is chosen only for investigating the behaviour of the discretization.

In contrast to the analysis in chapter 2, the Dirichlet boundary condition are not homogeneous for the exact solution. Hence, we have to modify the right-hand side of the discrete problem in the following way:
We define the linear form $F_h(\cdot)$ through

$$F_h(w_h') = \int_D f w_h' \ \mathrm{d}x + \sum_{F \in \mathcal{F}_h^{\partial D}} \int_F u_\nu \partial_u w_h' w \ \mathrm{d}s + \int_F \rho \frac{\lambda}{\rho} |u_\nu|^2 w_h' w \ \mathrm{d}s, \tag{3.2}$$

where $\mathcal{F}_h^{\partial D}$ is the set of all boundary facets.
Then, the discrete problem reads as: Find $w_h \in W_h$ such that

$$\mathcal{B}_h(w_h, w_h') = F_h(w_h') \qquad \forall w_h' \in W_h. \tag{3.3}$$

Given an exact solution, the source term f is calculated automatically with NGSolve as presented in section B.1.

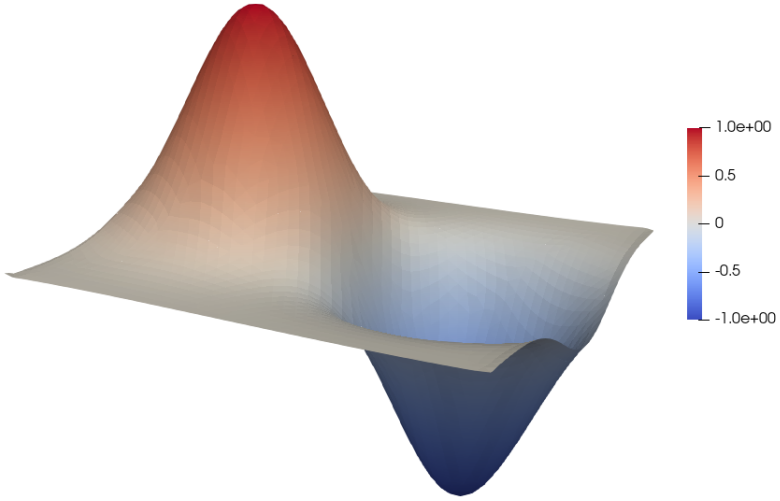


Figure 3.1: Exact solution w on the square geometry

Further, we will consider three different velocity fields:

$$\begin{aligned} u_1 &= (1, 1), \\ u_2 &= (-0.75y, 0.75x) \\ u_3 &= (2y(1-x^2), -2x(1-y^2)) \end{aligned} \tag{3.4}$$

The first velocity field u_1 is simply the constant case. However, as we want to test how the complexity of the velocity field might influence the performance of the discretization, we also consider u_2 and u_3. The former is a rigid body rotation, whereas the latter describes a vortex contained in D. Figures 3.2,3.3 and 3.4 display the velocity fields on a structured mesh.

To evaluate the performance of the method, we considered the error in the L^2- and the W-norm:

$$\begin{aligned} e_{L^2} &:= \|w - w_h\|_D; \\ e_W &:= |||w - w_h|||_W. \end{aligned}$$

Further, we will also compute the error of the best L^2-approximation. This means that we solve for w_{BL^2} such that

$$\langle w_{BL^2}, w_h' \rangle_D = \langle w, w_h' \rangle_D \qquad \forall w_h' \in W_h. \tag{3.5}$$

Then, we denote

$$e_{BL^2} := \|w - w_{BL^2}\|_D. \tag{3.6}$$

We note that e_{BL^2} is naturally a lower bound for e_{L^2}.

In the convergence tables in this section, the estimated order of convergence (eoc) is given. For errors e_i, $2 \leq i \leq N$ where N is the number of mesh refinements, we calculate the eoc with the following formula

$$eoc = \frac{\log(e_{i-1}/e_i)}{\log(2)}. \tag{3.7}$$

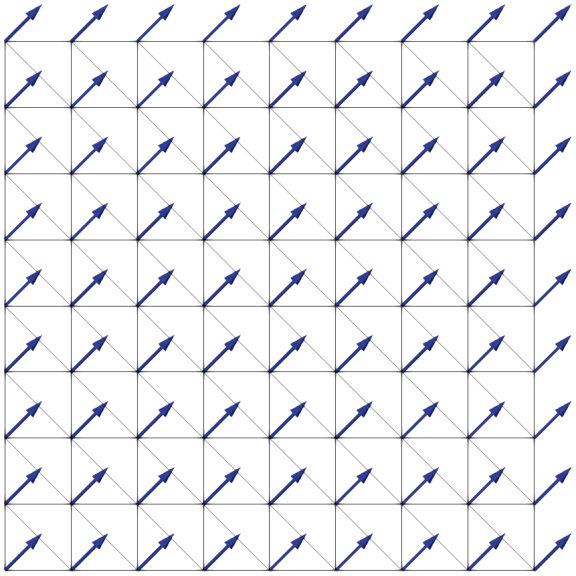


Figure 3.2: Velocity field u_1 on a structured mesh

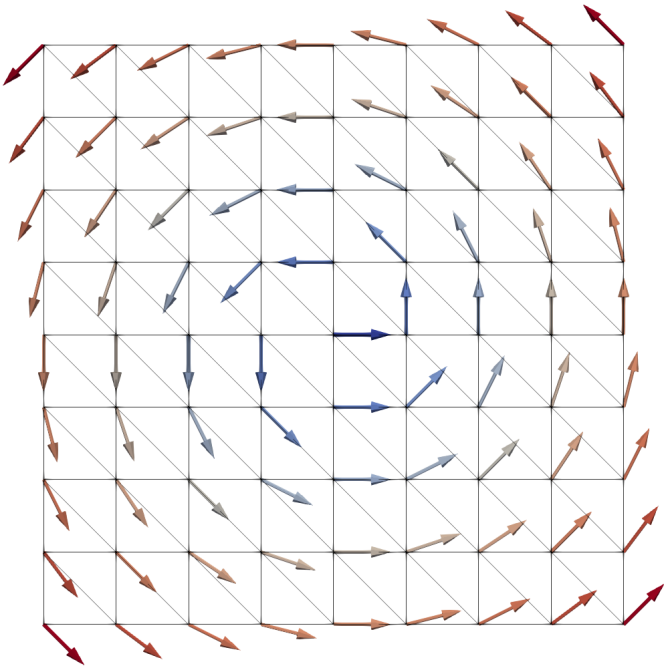


Figure 3.3: Velocity field u_2 on a structured mesh

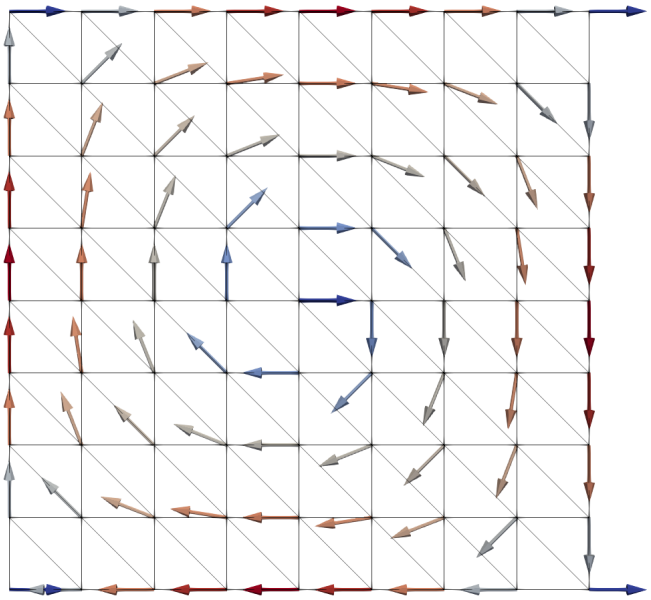


Figure 3.4: Velocity field u_3 on a structured mesh

Furthermore, in the plots displaying the numerical errors we use the abbreviation

$$O(k) := \mathcal{O}(h^k). \tag{3.8}$$

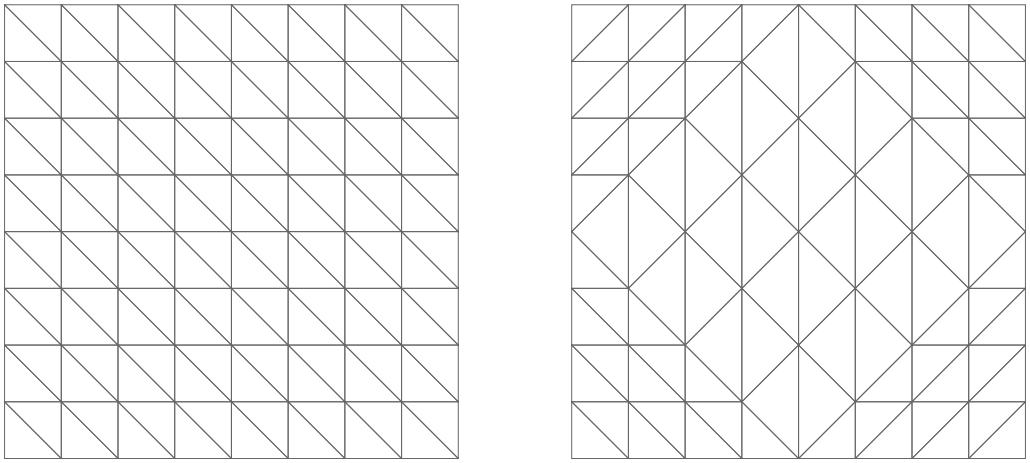
3.2 Structured vs. unstructured Meshes

Before doing convergence studies, we have to choose a triangulation of the domain D. It turns out, that the type of mesh does influence the convergence in the L^2-norm. This section investigates what differences occur between a structured and an unstructured mesh.

The code for generating both, a structured and an unstructured mesh, with NGSolve is shown in section B.2. Further, figure 3.5 shows examples for both types of mesh. We note that in the case of an unstructured mesh[1], varying the initial mesh size does change the appearance of the mesh, cf. figure 3.6.

To compare how the choice of a structured or an unstructured mesh influences the rate of convergence, we test the method with the velocity field u_1, a polynomial degree $k = 1$, a constant density $\rho = 1$, and a penalization parameter $\lambda = 40$. Tables 3.1 display the errors e_{L^2}, e_{BL^2} and e_W for a structured mesh and table 3.2 the same errors for an unstructured mesh with initial mesh size 0.7. We observe that we achieve optimal rates of convergence for all three errors on the structured mesh. In contrast, the rate of convergence in the L^2-norm in not optimal on the unstructured mesh, where the rate tends to be around half an order lower than $k + 1$ and decreases even more with an increasing number of refinements.

[1] note, that we use an unstructured mesh that is regularly refined, which means especially that the angles between u and ν do not change.



(a) Structured Mesh (b) Unstructured Mesh

Figure 3.5: A structured mesh (left) generated with `n = 3` in the code from section B.2 and an unstructured mesh (right) uniformly refined 2 times with an initial mesh size of 1.

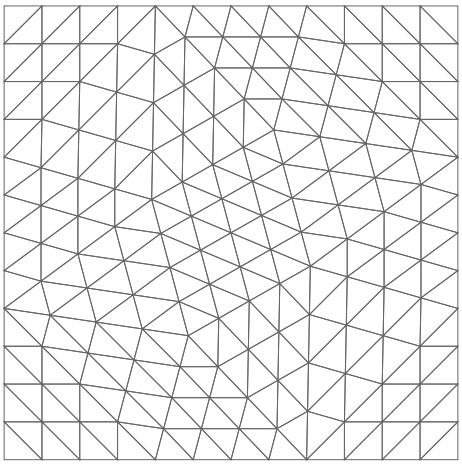


Figure 3.6: An unstructured mesh uniformly refined 2 times with an initial mesh size of 0.7.

refs	e_{L^2}	(eoc)	e_{BL^2}	(eoc)	e_W	(eoc)
$k = 1$						
1	$5.43 \cdot 10^{-1}$		$2.43 \cdot 10^{-1}$		$2.99 \cdot 10^{0}$	
2	$2.63 \cdot 10^{-1}$	(1.05)	$7.11 \cdot 10^{-2}$	(1.77)	$1.43 \cdot 10^{0}$	(1.07)
3	$1.64 \cdot 10^{-1}$	(0.68)	$2.29 \cdot 10^{-2}$	(1.64)	$9.72 \cdot 10^{-1}$	(0.55)
4	$7.22 \cdot 10^{-2}$	(1.18)	$5.85 \cdot 10^{-3}$	(1.97)	$5.30 \cdot 10^{-1}$	(0.88)
5	$2.50 \cdot 10^{-2}$	(1.53)	$1.47 \cdot 10^{-3}$	(1.99)	$2.69 \cdot 10^{-1}$	(0.98)
6	$7.16 \cdot 10^{-3}$	(1.8)	$3.68 \cdot 10^{-4}$	(2.0)	$1.32 \cdot 10^{-1}$	(1.03)
7	$1.87 \cdot 10^{-3}$	(1.94)	$9.21 \cdot 10^{-5}$	(2.0)	$6.47 \cdot 10^{-2}$	(1.03)
8	$4.74 \cdot 10^{-4}$	(1.98)	$2.30 \cdot 10^{-5}$	(2.0)	$3.19 \cdot 10^{-2}$	(1.02)

Table 3.1: Convergence table for u_1 on a structured mesh

refs	e_{L^2}	(eoc)	e_{BL^2}	(eoc)	e_W	(eoc)
$k = 1$						
1	$1.21 \cdot 10^{-1}$		$4.08 \cdot 10^{-2}$		$1.18 \cdot 10^{0}$	
2	$4.11 \cdot 10^{-2}$	(1.56)	$1.11 \cdot 10^{-2}$	(1.88)	$6.06 \cdot 10^{-1}$	(0.96)
3	$1.24 \cdot 10^{-2}$	(1.73)	$2.81 \cdot 10^{-3}$	(1.98)	$3.03 \cdot 10^{-1}$	(1.0)
4	$3.59 \cdot 10^{-3}$	(1.78)	$7.05 \cdot 10^{-4}$	(1.99)	$1.50 \cdot 10^{-1}$	(1.01)
5	$1.11 \cdot 10^{-3}$	(1.7)	$1.76 \cdot 10^{-4}$	(2.0)	$7.47 \cdot 10^{-2}$	(1.01)
6	$4.03 \cdot 10^{-4}$	(1.46)	$4.41 \cdot 10^{-5}$	(2.0)	$3.73 \cdot 10^{-2}$	(1.0)
7	$1.75 \cdot 10^{-4}$	(1.2)	$1.10 \cdot 10^{-5}$	(2.0)	$1.86 \cdot 10^{-2}$	(1.0)
8	$8.35 \cdot 10^{-5}$	(1.07)	$2.76 \cdot 10^{-6}$	(2.0)	$9.31 \cdot 10^{-3}$	(1.0)

Table 3.2: Convergence table for u_1 on an unstructured mesh

To investigate this further, we repeat the test with a structured mesh that has flipped triangles. This means that the hypotenuse of the mesh elements is oriented to the upper right instead of the standard case, where the hypotenuse is oriented to the upper left. Figure 3.7 shows an example of such a mesh and table 3.3 displays the observed rates of convergence.

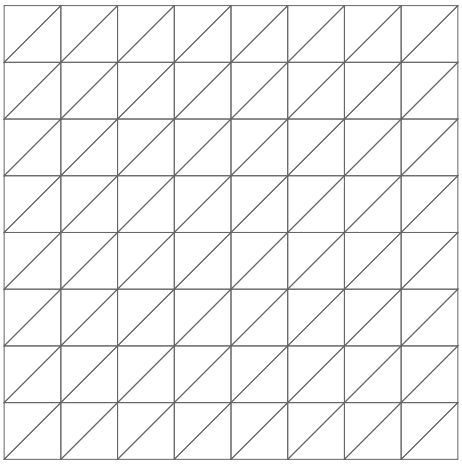


Figure 3.7: A structured mesh as in figure 3.5a with flipped triangles.

While the estimated order of convergence for e_W is still optimal, e_{L^2} does not converge with optimal order. In fact, the rate of convergence is close to $k = 1$ after 5 refinements and thus lower than in the unstructured case. To investigate these results, we plot the velocity field u_1 on the mesh with flipped triangles in figure 3.8. We observed that the edges of the triangles are aligned with velocity field. In particular, this means that $u \cdot \nu_F = 0$ for many facets, where ν_F is the facet normal. In the following, we will explain why the alignment of the facets does, in fact, have an influence on the L^2-convergence rates.

Considering the alignment of the facets, we can give an idea why the rate of convergence is optimal for the first structured mesh, around half an order worse with a downward drift as the mesh gets smaller for the unstructured case, and suboptimal for the second structured mesh with flipped triangles. In remark 2.3 we mentioned that usually second order deriva-

refs	e_{L^2}	(eoc)	e_{BL^2}	(eoc)	e_W	(eoc)
$k = 1$						
1	$4.83 \cdot 10^{-1}$		$2.43 \cdot 10^{-1}$		$2.67 \cdot 10^{0}$	
2	$1.52 \cdot 10^{-1}$	(1.66)	$7.11 \cdot 10^{-2}$	(1.77)	$1.47 \cdot 10^{0}$	(0.86)
3	$6.44 \cdot 10^{-2}$	(1.24)	$2.29 \cdot 10^{-2}$	(1.64)	$8.60 \cdot 10^{-1}$	(0.77)
4	$2.55 \cdot 10^{-2}$	(1.34)	$5.85 \cdot 10^{-3}$	(1.97)	$4.33 \cdot 10^{-1}$	(0.99)
5	$1.17 \cdot 10^{-2}$	(1.12)	$1.47 \cdot 10^{-3}$	(1.99)	$2.17 \cdot 10^{-1}$	(1.0)
6	$5.72 \cdot 10^{-3}$	(1.03)	$3.68 \cdot 10^{-4}$	(2.0)	$1.08 \cdot 10^{-1}$	(1.0)
7	$2.84 \cdot 10^{-3}$	(1.01)	$9.21 \cdot 10^{-5}$	(2.0)	$5.42 \cdot 10^{-2}$	(1.0)
8	$1.42 \cdot 10^{-3}$	(1.0)	$2.30 \cdot 10^{-5}$	(2.0)	$2.71 \cdot 10^{-2}$	(1.0)

Table 3.3: Convergence table for u_1 on a structured mesh with flipped triangles

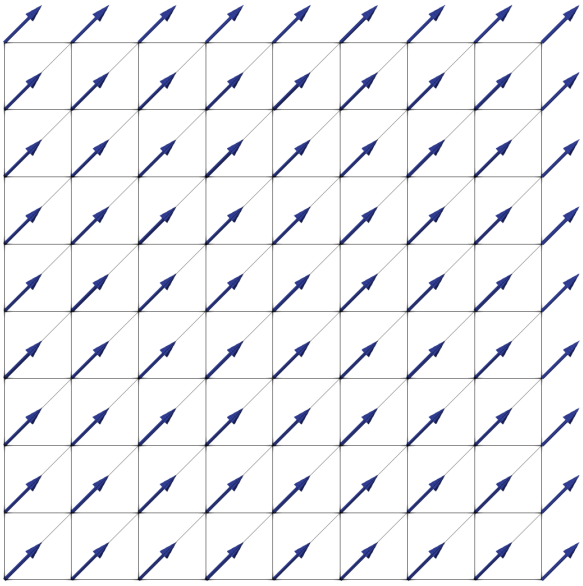


Figure 3.8: Velocity field u_1 on a structured mesh with flipped triangles.

tives have a smoothing effect that is useful for proving optimal convergence rates in the L^2-norm and that we cannot apply such an argument, because the differential operator does not have this effect orthogonal to the velocity field. In the discrete case however, we seem to have a discrete smoothing effect that is captured when the facets are orthogonal to the velocity field, which is the case for the first structured mesh. Consequently, we observe convergence of optimal order in the L^2-norm. In contrast, the structured mesh with flipped triangles does not capture this effect, because the facets are aligned with the velocity field and hence, we only get suboptimal convergence rates. The unstructured mesh lies in between. Some facets might be orthogonal to the velocity field, while others might be aligned, which is displayed in figure 3.9. As the mesh gets smaller there might be a lot of facets that are aligned with the velocity, which explains why the rates in table 3.3 deteriorate with an increasing number of refinements.

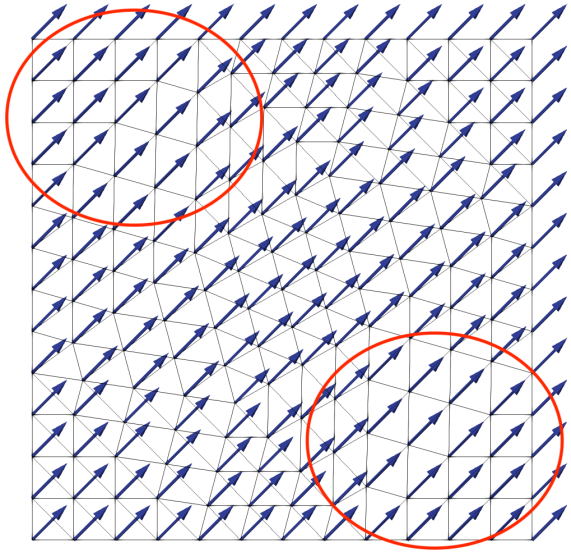


Figure 3.9: Velocity field u_1 on an unstructured mesh with initial mesh size 0.7 and 2 refinements. In the marked areas, the facets are (mostly) aligned with the velocity field.

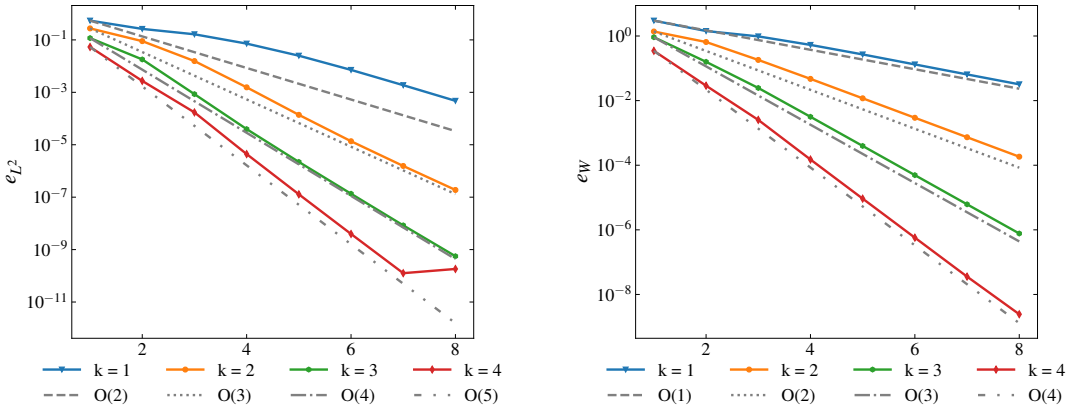
On the basis of the interpretation, we choose to conduct the following experiments on a structured mesh, as the convergence rates on an unstructured mesh can vary. In particular, we choose the first type of structured mesh because it might yield optimal convergence rates in the L^2-norm.

3.3 Convergence studies for different velocity fields

Due to the tests in the previous section, we decided to perform the convergence studies with a structured mesh. We choose a constant density $\rho = 1$ and set the penalization parameter to $\lambda = 10(k+1)^2$. Now we solve the problem for each velocity field u_i, $1 \leq i \leq 3$ and polynomial degrees $1 \leq k \leq 4$. In each case, we refine the mesh eight times.

Our goal is to validate the result (2.33), which indicates convergence in an optimal order of convergence of k for e_W. Further, recall from remark 2.3 that while we are not able to prove an optimal convergence rate of $k+1$ in the L^2, there at least holds $e_{L^2} \lesssim e_W$. We will investigate, if we observe a better rate of convergence for our model problem.

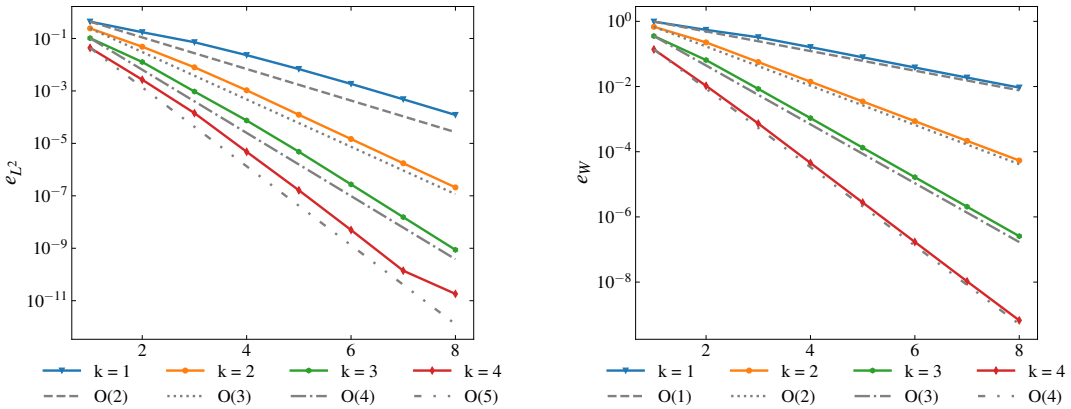
Figures 3.10, 3.11 and 3.12 display the errors e_W and e_{L^2} for the three velocity fields u_1, u_2 and u_3 in a semi-log scale. Additionally, the tables 3.4, 3.5 and 3.6 show the e_W, e_{L^2} and e_{BL^2} together with the respective eoc.



(a) Errors in the L^2-norm.

(b) Errors in the W-norm.

Figure 3.10: Numerical errors for u_1.



(a) Errors in the L^2-norm.

(b) Errors in the W-norm.

Figure 3.11: Numerical errors for u_2.

In accordance with the theoretical result, the errors in the W-norm converge with an optimal order of k for all three velocity fields. Furthermore, the L^2-rate for u_1 and u_2 is approximately $k+1$. For u_3, the rates tends to be half and order below $k+1$. However, similarly as discussed in section 3.2, this might be caused by the combination of the mesh and the non-constant velocity field. In comparison with e_{L^2}, the error of the best L^2-approximation is approximately one order better. Additionally, the rate of convergences drops in the last refinement step, with severity depending on the velocity field. This decline might be an issue of computational machine accuracy, as both the error and the mesh size are small.

In summary, the method seems to perform well.The observed convergence rates in the W-norm are optimal for all three velocity fields. Even though the L^2-rates depend on the velocity field and the mesh, the rates are optimal for u_1 and u_2.

refs	e_{L^2}	(eoc)	e_{BL^2}	(eoc)	e_W	(eoc)
$k=1$						
1	$5.43 \cdot 10^{-1}$		$2.43 \cdot 10^{-1}$		$2.99 \cdot 10^{0}$	
2	$2.63 \cdot 10^{-1}$	(1.05)	$7.11 \cdot 10^{-2}$	(1.77)	$1.43 \cdot 10^{0}$	(1.07)
3	$1.64 \cdot 10^{-1}$	(0.68)	$2.29 \cdot 10^{-2}$	(1.64)	$9.72 \cdot 10^{-1}$	(0.55)
4	$7.22 \cdot 10^{-2}$	(1.18)	$5.85 \cdot 10^{-3}$	(1.97)	$5.30 \cdot 10^{-1}$	(0.88)
5	$2.50 \cdot 10^{-2}$	(1.53)	$1.47 \cdot 10^{-3}$	(1.99)	$2.69 \cdot 10^{-1}$	(0.98)
6	$7.16 \cdot 10^{-3}$	(1.8)	$3.68 \cdot 10^{-4}$	(2.0)	$1.32 \cdot 10^{-1}$	(1.03)
7	$1.87 \cdot 10^{-3}$	(1.94)	$9.21 \cdot 10^{-5}$	(2.0)	$6.47 \cdot 10^{-2}$	(1.03)
8	$4.74 \cdot 10^{-4}$	(1.98)	$2.30 \cdot 10^{-5}$	(2.0)	$3.19 \cdot 10^{-2}$	(1.02)
$k=2$						
1	$2.75 \cdot 10^{-1}$		$1.23 \cdot 10^{-1}$		$1.38 \cdot 10^{0}$	
2	$8.97 \cdot 10^{-2}$	(1.62)	$2.41 \cdot 10^{-2}$	(2.35)	$6.52 \cdot 10^{-1}$	(1.08)
3	$1.56 \cdot 10^{-2}$	(2.53)	$3.09 \cdot 10^{-3}$	(2.96)	$1.82 \cdot 10^{-1}$	(1.84)
4	$1.55 \cdot 10^{-3}$	(3.33)	$3.97 \cdot 10^{-4}$	(2.96)	$4.71 \cdot 10^{-2}$	(1.95)
5	$1.38 \cdot 10^{-4}$	(3.49)	$4.99 \cdot 10^{-5}$	(2.99)	$1.18 \cdot 10^{-2}$	(2.0)
6	$1.36 \cdot 10^{-5}$	(3.34)	$6.25 \cdot 10^{-6}$	(3.0)	$2.94 \cdot 10^{-3}$	(2.0)
7	$1.55 \cdot 10^{-6}$	(3.14)	$7.82 \cdot 10^{-7}$	(3.0)	$7.35 \cdot 10^{-4}$	(2.0)
8	$1.89 \cdot 10^{-7}$	(3.04)	$9.78 \cdot 10^{-8}$	(3.0)	$1.84 \cdot 10^{-4}$	(2.0)
$k=3$						
1	$1.17 \cdot 10^{-1}$		$3.07 \cdot 10^{-2}$		$9.17 \cdot 10^{-1}$	
2	$1.80 \cdot 10^{-2}$	(2.7)	$1.37 \cdot 10^{-3}$	(4.48)	$1.59 \cdot 10^{-1}$	(2.53)
3	$8.53 \cdot 10^{-4}$	(4.4)	$1.11 \cdot 10^{-4}$	(3.63)	$2.46 \cdot 10^{-2}$	(2.69)
4	$3.95 \cdot 10^{-5}$	(4.43)	$6.68 \cdot 10^{-6}$	(4.05)	$3.15 \cdot 10^{-3}$	(2.97)
5	$2.21 \cdot 10^{-6}$	(4.16)	$4.13 \cdot 10^{-7}$	(4.02)	$3.95 \cdot 10^{-4}$	(3.0)
6	$1.35 \cdot 10^{-7}$	(4.03)	$2.57 \cdot 10^{-8}$	(4.0)	$4.94 \cdot 10^{-5}$	(3.0)
7	$8.40 \cdot 10^{-9}$	(4.01)	$1.61 \cdot 10^{-9}$	(4.0)	$6.16 \cdot 10^{-6}$	(3.0)
8	$5.65 \cdot 10^{-10}$	(3.9)	$1.00 \cdot 10^{-10}$	(4.0)	$7.70 \cdot 10^{-7}$	(3.0)
$k=4$						
1	$5.37 \cdot 10^{-2}$		$1.63 \cdot 10^{-2}$		$3.47 \cdot 10^{-1}$	
2	$2.72 \cdot 10^{-3}$	(4.3)	$7.81 \cdot 10^{-4}$	(4.39)	$2.89 \cdot 10^{-2}$	(3.59)
3	$1.72 \cdot 10^{-4}$	(3.99)	$4.08 \cdot 10^{-5}$	(4.26)	$2.53 \cdot 10^{-3}$	(3.52)
4	$4.40 \cdot 10^{-6}$	(5.28)	$1.33 \cdot 10^{-6}$	(4.94)	$1.50 \cdot 10^{-4}$	(4.07)
5	$1.29 \cdot 10^{-7}$	(5.1)	$4.19 \cdot 10^{-8}$	(4.98)	$9.24 \cdot 10^{-6}$	(4.02)
6	$3.95 \cdot 10^{-9}$	(5.03)	$1.31 \cdot 10^{-9}$	(5.0)	$5.73 \cdot 10^{-7}$	(4.01)
7	$1.26 \cdot 10^{-10}$	(4.97)	$4.11 \cdot 10^{-11}$	(5.0)	$3.57 \cdot 10^{-8}$	(4.01)
8	$1.83 \cdot 10^{-10}$	(-0.54)	$1.28 \cdot 10^{-12}$	(5.0)	$2.44 \cdot 10^{-9}$	(3.87)

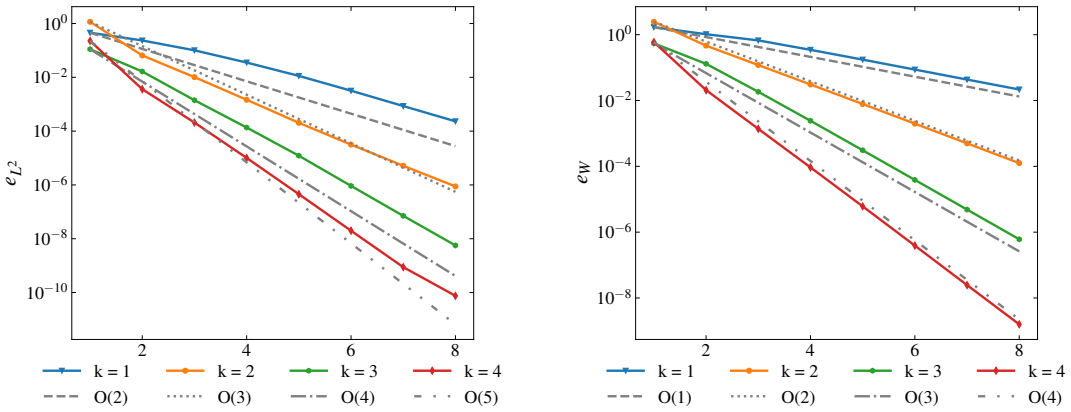
Table 3.4: Convergence table for u_1.

refs	e_{L^2}	(eoc)	e_{BL^2}	(eoc)	e_W	(eoc)
$k = 1$						
1	$4.39 \cdot 10^{-1}$		$2.43 \cdot 10^{-1}$		$9.80 \cdot 10^{-1}$	
2	$1.73 \cdot 10^{-1}$	(1.34)	$7.11 \cdot 10^{-2}$	(1.77)	$5.47 \cdot 10^{-1}$	(0.84)
3	$7.18 \cdot 10^{-2}$	(1.27)	$2.29 \cdot 10^{-2}$	(1.64)	$3.25 \cdot 10^{-1}$	(0.75)
4	$2.32 \cdot 10^{-2}$	(1.63)	$5.85 \cdot 10^{-3}$	(1.97)	$1.62 \cdot 10^{-1}$	(1.01)
5	$6.82 \cdot 10^{-3}$	(1.76)	$1.47 \cdot 10^{-3}$	(1.99)	$7.88 \cdot 10^{-2}$	(1.04)
6	$1.85 \cdot 10^{-3}$	(1.88)	$3.68 \cdot 10^{-4}$	(2.0)	$3.84 \cdot 10^{-2}$	(1.04)
7	$4.80 \cdot 10^{-4}$	(1.95)	$9.21 \cdot 10^{-5}$	(2.0)	$1.88 \cdot 10^{-2}$	(1.03)
8	$1.21 \cdot 10^{-4}$	(1.98)	$2.30 \cdot 10^{-5}$	(2.0)	$9.30 \cdot 10^{-3}$	(1.02)
$k = 2$						
1	$2.43 \cdot 10^{-1}$		$1.23 \cdot 10^{-1}$		$6.80 \cdot 10^{-1}$	
2	$4.84 \cdot 10^{-2}$	(2.33)	$2.41 \cdot 10^{-2}$	(2.35)	$2.26 \cdot 10^{-1}$	(1.59)
3	$7.95 \cdot 10^{-3}$	(2.61)	$3.09 \cdot 10^{-3}$	(2.96)	$5.64 \cdot 10^{-2}$	(2.0)
4	$1.06 \cdot 10^{-3}$	(2.91)	$3.97 \cdot 10^{-4}$	(2.96)	$1.41 \cdot 10^{-2}$	(2.0)
5	$1.24 \cdot 10^{-4}$	(3.1)	$4.99 \cdot 10^{-5}$	(2.99)	$3.48 \cdot 10^{-3}$	(2.01)
6	$1.46 \cdot 10^{-5}$	(3.08)	$6.25 \cdot 10^{-6}$	(3.0)	$8.65 \cdot 10^{-4}$	(2.01)
7	$1.74 \cdot 10^{-6}$	(3.07)	$7.82 \cdot 10^{-7}$	(3.0)	$2.15 \cdot 10^{-4}$	(2.01)
8	$2.09 \cdot 10^{-7}$	(3.06)	$9.78 \cdot 10^{-8}$	(3.0)	$5.37 \cdot 10^{-5}$	(2.0)
$k = 3$						
1	$1.03 \cdot 10^{-1}$		$3.07 \cdot 10^{-2}$		$3.54 \cdot 10^{-1}$	
2	$1.27 \cdot 10^{-2}$	(3.02)	$1.37 \cdot 10^{-3}$	(4.48)	$6.49 \cdot 10^{-2}$	(2.45)
3	$9.49 \cdot 10^{-4}$	(3.74)	$1.11 \cdot 10^{-4}$	(3.63)	$8.46 \cdot 10^{-3}$	(2.94)
4	$7.46 \cdot 10^{-5}$	(3.67)	$6.68 \cdot 10^{-6}$	(4.05)	$1.07 \cdot 10^{-3}$	(2.98)
5	$4.80 \cdot 10^{-6}$	(3.96)	$4.13 \cdot 10^{-7}$	(4.02)	$1.33 \cdot 10^{-4}$	(3.01)
6	$2.73 \cdot 10^{-7}$	(4.14)	$2.57 \cdot 10^{-8}$	(4.0)	$1.65 \cdot 10^{-5}$	(3.01)
7	$1.54 \cdot 10^{-8}$	(4.15)	$1.61 \cdot 10^{-9}$	(4.0)	$2.05 \cdot 10^{-6}$	(3.01)
8	$8.69 \cdot 10^{-10}$	(4.15)	$1.00 \cdot 10^{-10}$	(4.0)	$2.56 \cdot 10^{-7}$	(3.0)
$k = 4$						
1	$4.38 \cdot 10^{-2}$		$1.63 \cdot 10^{-2}$		$1.37 \cdot 10^{-1}$	
2	$2.67 \cdot 10^{-3}$	(4.03)	$7.81 \cdot 10^{-4}$	(4.39)	$1.05 \cdot 10^{-2}$	(3.71)
3	$1.42 \cdot 10^{-4}$	(4.23)	$4.08 \cdot 10^{-5}$	(4.26)	$7.34 \cdot 10^{-4}$	(3.84)
4	$4.78 \cdot 10^{-6}$	(4.9)	$1.33 \cdot 10^{-6}$	(4.94)	$4.51 \cdot 10^{-5}$	(4.02)
5	$1.63 \cdot 10^{-7}$	(4.87)	$4.19 \cdot 10^{-8}$	(4.98)	$2.77 \cdot 10^{-6}$	(4.03)
6	$4.97 \cdot 10^{-9}$	(5.03)	$1.31 \cdot 10^{-9}$	(5.0)	$1.70 \cdot 10^{-7}$	(4.02)
7	$1.39 \cdot 10^{-10}$	(5.16)	$4.11 \cdot 10^{-11}$	(5.0)	$1.05 \cdot 10^{-8}$	(4.01)
8	$1.88 \cdot 10^{-11}$	(2.89)	$1.28 \cdot 10^{-12}$	(5.0)	$6.70 \cdot 10^{-10}$	(3.97)

Table 3.5: Convergence table for u_2.

refs	e_{L^2}	(eoc)	e_{BL^2}	(eoc)	e_W	(eoc)
$k = 1$						
1	$4.56 \cdot 10^{-1}$		$2.43 \cdot 10^{-1}$		$1.69 \cdot 10^{0}$	
2	$2.35 \cdot 10^{-1}$	(0.96)	$7.11 \cdot 10^{-2}$	(1.77)	$1.03 \cdot 10^{0}$	(0.71)
3	$1.01 \cdot 10^{-1}$	(1.22)	$2.29 \cdot 10^{-2}$	(1.64)	$6.71 \cdot 10^{-1}$	(0.62)
4	$3.55 \cdot 10^{-2}$	(1.51)	$5.85 \cdot 10^{-3}$	(1.97)	$3.47 \cdot 10^{-1}$	(0.95)
5	$1.13 \cdot 10^{-2}$	(1.65)	$1.47 \cdot 10^{-3}$	(1.99)	$1.75 \cdot 10^{-1}$	(0.99)
6	$3.19 \cdot 10^{-3}$	(1.83)	$3.68 \cdot 10^{-4}$	(2.0)	$8.71 \cdot 10^{-2}$	(1.01)
7	$8.56 \cdot 10^{-4}$	(1.9)	$9.21 \cdot 10^{-5}$	(2.0)	$4.33 \cdot 10^{-2}$	(1.01)
8	$2.32 \cdot 10^{-4}$	(1.89)	$2.30 \cdot 10^{-5}$	(2.0)	$2.16 \cdot 10^{-2}$	(1.0)
$k = 2$						
1	$1.16 \cdot 10^{0}$		$1.23 \cdot 10^{-1}$		$2.47 \cdot 10^{0}$	
2	$6.50 \cdot 10^{-2}$	(4.16)	$2.41 \cdot 10^{-2}$	(2.35)	$4.65 \cdot 10^{-1}$	(2.41)
3	$1.02 \cdot 10^{-2}$	(2.67)	$3.09 \cdot 10^{-3}$	(2.96)	$1.20 \cdot 10^{-1}$	(1.96)
4	$1.47 \cdot 10^{-3}$	(2.8)	$3.97 \cdot 10^{-4}$	(2.96)	$3.08 \cdot 10^{-2}$	(1.96)
5	$2.04 \cdot 10^{-4}$	(2.84)	$4.99 \cdot 10^{-5}$	(2.99)	$7.83 \cdot 10^{-3}$	(1.97)
6	$3.15 \cdot 10^{-5}$	(2.7)	$6.25 \cdot 10^{-6}$	(3.0)	$1.98 \cdot 10^{-3}$	(1.98)
7	$5.16 \cdot 10^{-6}$	(2.61)	$7.82 \cdot 10^{-7}$	(3.0)	$4.99 \cdot 10^{-4}$	(1.99)
8	$8.81 \cdot 10^{-7}$	(2.55)	$9.78 \cdot 10^{-8}$	(3.0)	$1.25 \cdot 10^{-4}$	(2.0)
$k = 3$						
1	$1.11 \cdot 10^{-1}$		$3.07 \cdot 10^{-2}$		$5.46 \cdot 10^{-1}$	
2	$1.65 \cdot 10^{-2}$	(2.75)	$1.37 \cdot 10^{-3}$	(4.48)	$1.30 \cdot 10^{-1}$	(2.07)
3	$1.41 \cdot 10^{-3}$	(3.54)	$1.11 \cdot 10^{-4}$	(3.63)	$1.84 \cdot 10^{-2}$	(2.82)
4	$1.36 \cdot 10^{-4}$	(3.37)	$6.68 \cdot 10^{-6}$	(4.05)	$2.42 \cdot 10^{-3}$	(2.93)
5	$1.22 \cdot 10^{-5}$	(3.48)	$4.13 \cdot 10^{-7}$	(4.02)	$3.08 \cdot 10^{-4}$	(2.98)
6	$9.20 \cdot 10^{-7}$	(3.73)	$2.57 \cdot 10^{-8}$	(4.0)	$3.87 \cdot 10^{-5}$	(2.99)
7	$7.09 \cdot 10^{-8}$	(3.7)	$1.61 \cdot 10^{-9}$	(4.0)	$4.85 \cdot 10^{-6}$	(3.0)
8	$5.64 \cdot 10^{-9}$	(3.65)	$1.00 \cdot 10^{-10}$	(4.0)	$6.07 \cdot 10^{-7}$	(3.0)
$k = 4$						
1	$2.26 \cdot 10^{-1}$		$1.63 \cdot 10^{-2}$		$5.95 \cdot 10^{-1}$	
2	$3.63 \cdot 10^{-3}$	(5.96)	$7.81 \cdot 10^{-4}$	(4.39)	$2.08 \cdot 10^{-2}$	(4.84)
3	$2.08 \cdot 10^{-4}$	(4.12)	$4.08 \cdot 10^{-5}$	(4.26)	$1.38 \cdot 10^{-3}$	(3.91)
4	$1.01 \cdot 10^{-5}$	(4.37)	$1.33 \cdot 10^{-6}$	(4.94)	$9.29 \cdot 10^{-5}$	(3.89)
5	$4.54 \cdot 10^{-7}$	(4.47)	$4.19 \cdot 10^{-8}$	(4.98)	$6.09 \cdot 10^{-6}$	(3.93)
6	$1.99 \cdot 10^{-8}$	(4.52)	$1.31 \cdot 10^{-9}$	(5.0)	$3.90 \cdot 10^{-7}$	(3.96)
7	$8.82 \cdot 10^{-10}$	(4.49)	$4.11 \cdot 10^{-11}$	(5.0)	$2.46 \cdot 10^{-8}$	(3.98)
8	$7.38 \cdot 10^{-11}$	(3.58)	$1.28 \cdot 10^{-12}$	(5.0)	$1.58 \cdot 10^{-9}$	(3.96)

Table 3.6: Convergence table for u_3.



(a) Errors in the L^2-norm.

(b) Errors in the W-norm.

Figure 3.12: Numerical errors for u_3.

3.4 Influence of the penalization parameter

The penalization parameter λ has to be chosen large enough so that the problem is coercive, but the condition number of the system matrices grow with λ. We already described this problem in section 2.6 along with some possible remedies like the Bassi-Rebay stabilization. In this section, we want to investigate numerically how the penalization parameter influences the condition number in our example problem.
To do this, we will calculate the condition number of the *stiffness matrix* for different λ. For a basis $\Phi = \{\phi_i\}_{1 \leq i \leq \dim W_h}$ of the space W_h, the stiffness matrix $\mathbb{B}$ is defined as

$$\mathbb{B}_{ij} = \mathcal{B}_h(\phi_j, \phi_i). \tag{3.9}$$

We define

$$\bar{\kappa}(\mathbb{B}) := \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{3.10}$$

where $\lambda_{\max}$ is the maximal and $\lambda_{\min}$ the minimal eigenvalue of the matrix $\mathbb{B}$.

This is a common estimate for the condition number, which is exact for symmetric positive definite matrices.
Furthermore, we consider the condition number when the matrix $\mathbb{B}$ is diagonal preconditioned. This means that we calculate $\bar{\kappa}(J\mathbb{B})$, where

$$J := \operatorname{diag}(\mathbb{B}_{FF})^{-1}. \tag{3.11}$$

For this experiment we choose a structured mesh generated by the code B.2 with six refinements. In tables 3.7 and 3.8 the condition numbers for polynomial degree $k = 1$ and $k = 4$ are shown. The results for $k = 2$ and $k = 3$ are similar and can be found in the appendix in section C.1.

We note that to approximate the condition numbers, we calculate the eigenvalues with a computationally simple numerical eigenvalue solver in NGSolve, which might not be very accurate. However, the results do illustrate how the penalization parameter influences the results from the discretization. First of all, we observe that the condition number does in fact grow with the penalization parameter λ. Secondly, we notice that some estimates

velocity field	u_1		u_2		u_3	
λ	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$
1	(-1.96)	(-1.01)	(-2.24)	(-1.07)	(-2.7)	(-2.23)
2	(-5.91)	(-4.43)	(-8.27)	(-7.1)	(-20.55)	(-8.93)
4	8218.48	8601.06	6573.03	2792.32	10349.29	4287.99
8	12291.3	12798.07	7770.09	3660.04	11393.5	5853.17
16	18322.75	17842.76	8388.49	4460.4	10879.98	6169.66
32	25880.81	22168.92	9061.45	5107.72	10716.85	7227.53
64	32346.54	24467.45	9343.06	5499.63	10415.5	8586.66
128	36733.82	31598.81	10199.64	6739.26	10764.01	8733.65
256	43054.67	34331.81	11187.31	5995.68	11008.81	8478.29
512	74125.53	41153.27	12617.2	9744.0	12161.53	8947.2
1024	117410.63	70580.26	13066.6	11694.13	13719.18	12125.62
2048	117454.63	133162.03	13856.81	16724.89	14467.64	21094.17
4096	156701.56	164851.99	13846.94	24457.18	14802.6	26448.93
8192	273717.37	172864.59	14204.49	30718.2	15321.66	29563.64

Table 3.7: Condition numbers with different λ for $k = 1$

velocity field	u_1		u_2		u_3	
λ	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$
1	(-0.66)	(-1.0)	(-0.71)	(-0.44)	(-0.75)	(-1.78)
2	(-0.74)	(-0.99)	(-0.85)	(-0.55)	(-0.94)	(-0.55)
4	(-0.95)	(-1.0)	(-1.25)	(-1.01)	(-1.67)	(-1.23)
8	(-1.66)	(-1.0)	(-2.94)	(-1.01)	(-8.22)	(-5.25)
16	(-7.25)	(-6.01)	(-27.73)	(-13.97)	3615.87	2510.83
32	6269.1	4263.93	3384.01	1975.33	3638.88	2465.41
64	4203.71	3337.48	3556.11	1910.33	3721.05	2472.56
128	3055.45	2068.3	3997.17	2162.34	3934.51	2609.07
256	3329.93	2085.27	4650.05	2491.29	4349.95	2718.85
512	2398.46	2519.34	5876.29	2776.75	5167.0	2997.91
1024	3449.83	2608.66	7870.54	3483.83	6125.91	3607.5
2048	5617.6	2819.21	11701.46	4158.63	8287.57	4225.68
4096	7383.71	5360.18	17969.45	5289.75	13070.98	5442.09
8192	10914.37	10289.04	26455.35	8452.72	20985.54	9032.28

Table 3.8: Condition numbers with different λ for $k = 4$

for the condition numbers are negative. As symmetric positive matrices have only real and positive eigenvalues, this can only happen when the stiffness matrix is not symmetric positive definite, which means that the problem is not coercive. Recall that in 2.2, we showed coercivity under the assumption that

$$\lambda \geq 1 + 4c_{\text{tr}}^2. \tag{3.12}$$

Further, from section 2.5 we know that the constant c_{tr} depends among other things on the polynomial degree k. The results in tables 3.7 and 3.8 shows this dependence: For $k = 1$, the condition number estimate is negative for $\lambda \leq 2$ and hence λ needs to be chosen bigger than two for the problem to be coercive. In contrast, for $k = 4$ we need to choose λ to be at least 16 such that the problem is coercive.

In conclusion, this experiment illustrated how difficult choosing a suitable penalization parameter might be. On the one hand, the condition numbers grow with λ, but on the other hand, λ has to be large enough such that the problem is coercive. Introducing a generalized eigenvalue problem or switching to a Bassi-Rebay type stabilization can solve this problem. The results in section 3.3 show that the discretization performs well for our example problem and the penalization parameter does not seem to cause problems. As such, we see no need to implement one of the remedies here.

3.5 The problem without the volume term

At the beginning of this thesis, we argued that the volume term is added to the diffusion operator to make the problem well-posed. In this section, we will demonstrate this by considering the diffusion operator only. We repeat the convergence studies from section 3.3 for the problem:
Find $w_h \in W_h$ such that

$$-\nabla \cdot (\rho(u \otimes u)\nabla w_h) = f \text{ in } D. \tag{3.13}$$

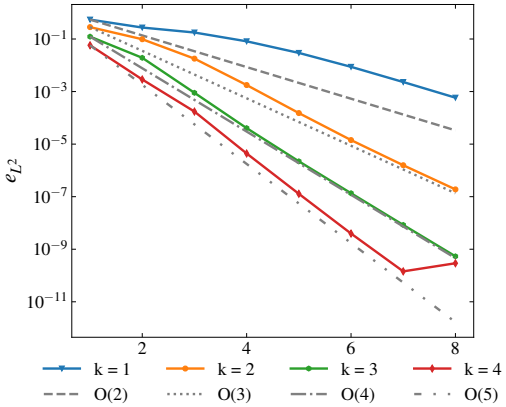
The corresponding bilinear form derived in section 1.3 is

$$\begin{aligned} b_h(w_h, w_h') = \sum_{T\in\mathcal{T}_h} \langle \rho\partial_u w_h, \partial_u w_h'\rangle_T + \sum_{F\in\mathcal{F}_h} \Big\{ &\langle u_\nu \{\!\{-\rho\partial_u w_h \cdot \nu\}\!\}, [\![w_h']\!]\rangle_F \\ &+ \langle u_\nu \{\!\{-\rho\partial_u w_h' \cdot \nu\}\!\}, [\![w_h]\!]\rangle_F \\ &+ \langle \frac{\rho\lambda}{h}|u_\nu|^2 [\![w_h]\!], [\![w_h']\!]\rangle_F \Big\}. \end{aligned} \tag{3.14}$$

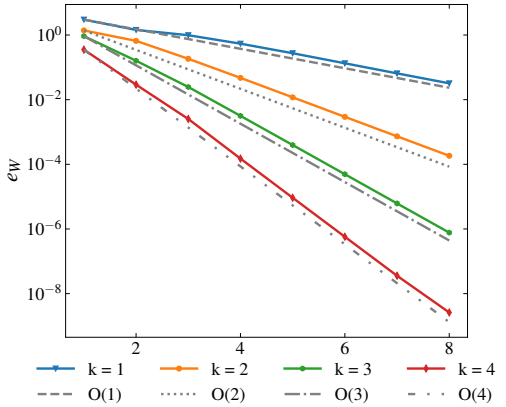
We note that we also have to adjust the code for calculation of the source term f (cf. section B.1) accordingly. Furthermore, we still consider the exact solution w defined in (3.1).

The differential operator ∂_u is linear, so any functions from its kernel that we add to a solution w_h vanishes. As such, the discrete solution is not unique, unless the velocity fields interacts sufficiently with the boundary conditions. In our case, the solution of problem without the volume term is only unique for the constant velocity field u_1 and not for u_2 and u_3. The velocity field u_1 is a straight line crossing the boundary. Hence, the discrete solution will be unique due to the Dirichlet boundary conditions. In contrast, for the other two velocity fields, a rotation, and a vortex, several trajectories do not cross the boundary. Consequently, the exact solution is not unique. Hence, we cannot expect this to be true for the discrete solution either.

The numerical results, displayed in figures 3.13, 3.14 and 3.15, confirm this assertion. For u_1, we observe that the method performs well and the L^2- and the W-norm error converges as expected from the previous experiments. For u_2 and u_3, the errors do not converge. Altogether, this experiment shows that the volume term is indeed vital for the well-posedness of the method.

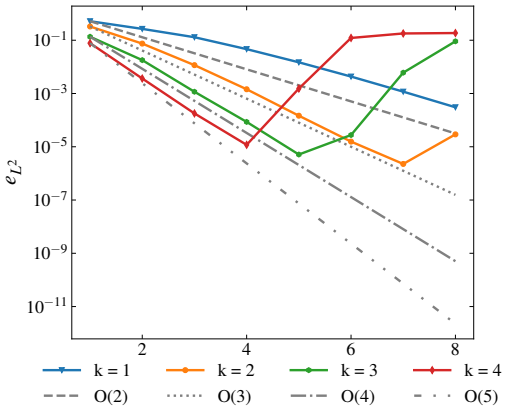


(a) Errors in the L^2-norm.

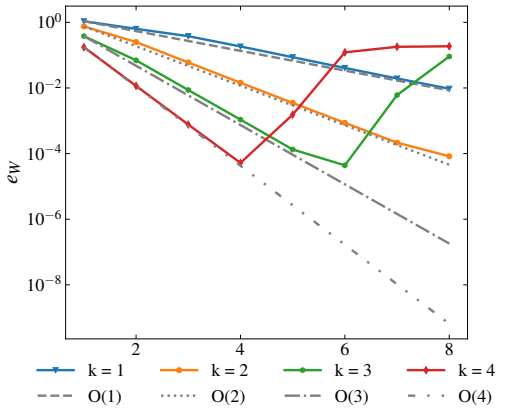


(b) Errors in the W-norm.

Figure 3.13: Numerical errors for u_1 without the volume term.

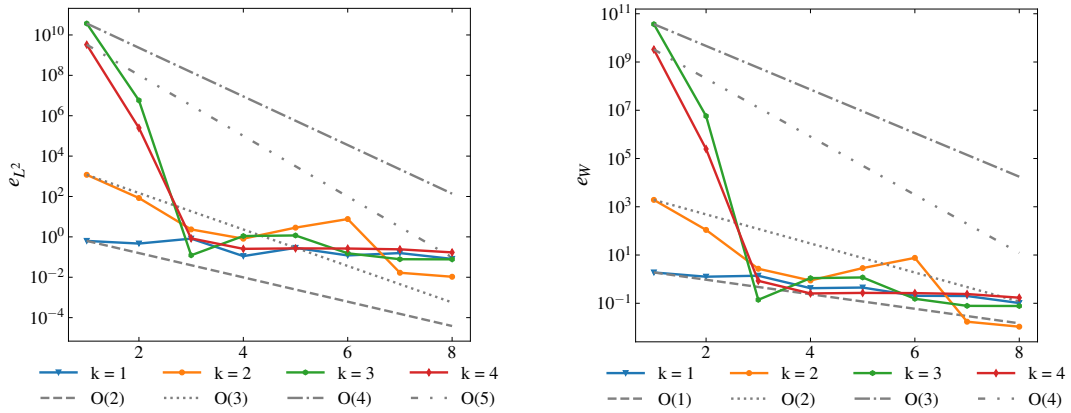


(a) Errors in the L^2-norm.



(b) Errors in the W-norm.

Figure 3.14: Numerical errors for u_2 without the volume term.



(a) Errors in the L^2-norm.

(b) Errors in the W-norm.

Figure 3.15: Numerical errors for u_3 without the volume term.

3.6 Non-constant density

While the theoretical analysis in chapter 2 explicitly allows for ρ to be non-constant, we have only considered a constant density so far. In this section, we want to explore the performance of our method in the case of a non-constant density.
In order to measure the distance of a point to the boundary more conveniently, we switch to a circle geometry:

$$D = \{x \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}, \tag{3.15}$$

Figure 3.16 shows the domain D. The unstructured mesh was generated with the third code block from section B.2. It is possible to generate a structured mesh for the circle, but considering the results from section 3.2 it might not necessarily be beneficial for the velocity fields u_2 and u_3. As such, we choose to conduct the experiments with an unstructured mesh.

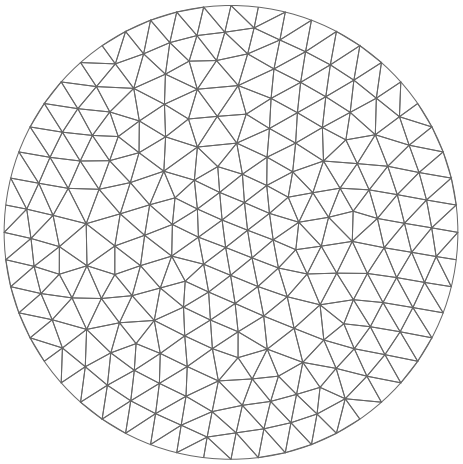


Figure 3.16: Circle geometry triangulated with an unstructured mesh.

The definition of the circle geometry allows us to measure the difference of a point $(x, y) \in \mathbb{R}^2$ to the boundary ∂D in the following way:

$$\text{dist}((x, y), \partial D) = 1 - \sqrt{x^2 + y^2}.$$

Hence, we define a density that has a peak at the origin and gets small close to the boundary:

$$\rho_n(x, y) = (1.75 - (x^2 + y^2))^n \qquad n \in \mathbb{N}. \tag{3.16}$$

Figure 3.17 shows ρ_2. As displayed in table 3.9, the bigger the exponent, the more extreme is the difference between the maximal and the minimal value.

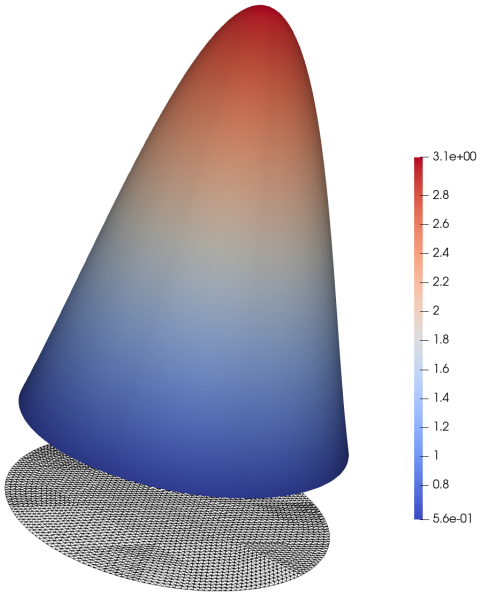


Figure 3.17: Density ρ_2 on the circle geometry.

n	$\max \rho_n$	$\min \rho_n$
2	3.0625	0.5625
4	9.3789	0.3164
8	87.9639	0.1001
12	825.005	0.0317
16	7737.6446	0.01

Table 3.9: Maximal and minimal value of ρ_n for $n \in \{2, 4, 8, 16\}$.

Furthermore, we still use the exact solution from the previous numerical experiments, that is

$$w = \exp(-6((x + 0.5)^2 + y^2)) - \exp(-6((x - 0.5)^2 + y^2)). \tag{3.17}$$

It is displayed on the circle geometry in figure 3.18. Note that the Dirichlet boundary condition are not homogeneous on the circle geometry as well, so we again use the linear form $F_h(\cdot)$ for the implementation.

Remark 3.1 (On the smoothness assumptions on ρ)**.** *Aside from being bounded from below and above, the main assumption on the density ρ in section 1.1 was that it should be sufficiently smooth. In our choice for ρ_n, we purposely used the square of the euclidean norm to fulfil this*

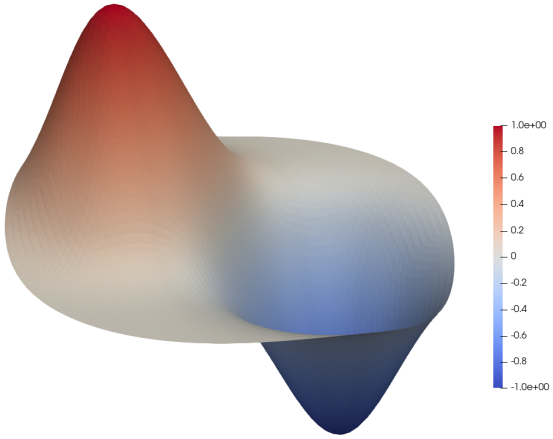
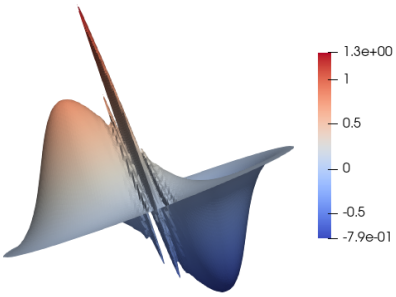


Figure 3.18: Exact solution w on the circle geometry.

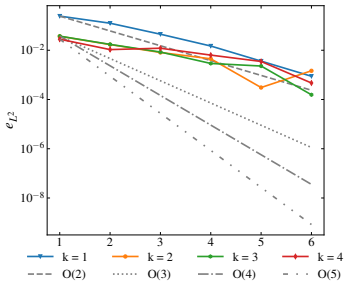
assumption. The following example shows that the smoothness assumption on ρ cannot be left out. We consider the density

$$\rho = (1.75 - \sqrt{x^2 + y^2})^{12}. \tag{3.18}$$

Due to the square root, this density violates the smoothness assumption at the point $(0, 0)$. The velocity field u_1 travels through this singularity.
Figure 3.19 shows the discrete solution and the rates of convergence for this case. The rates are far from optimal, and we can even see the singularity in the discrete solution, which shows that the smoothness assumption on ρ is indeed necessary.



(a) Discrete Solution

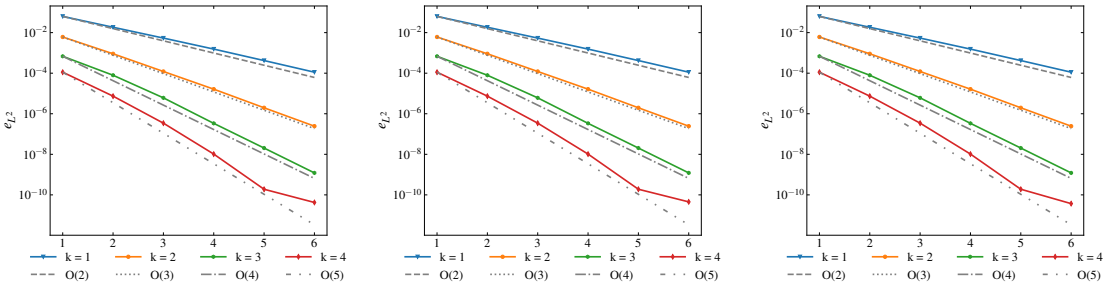


(b) Errors in the L^2-norm

Figure 3.19: Discrete solution (left) with u_1 and the non-smooth density (3.18) with the corresponding L^2-errors (right) in a semi-log scale.

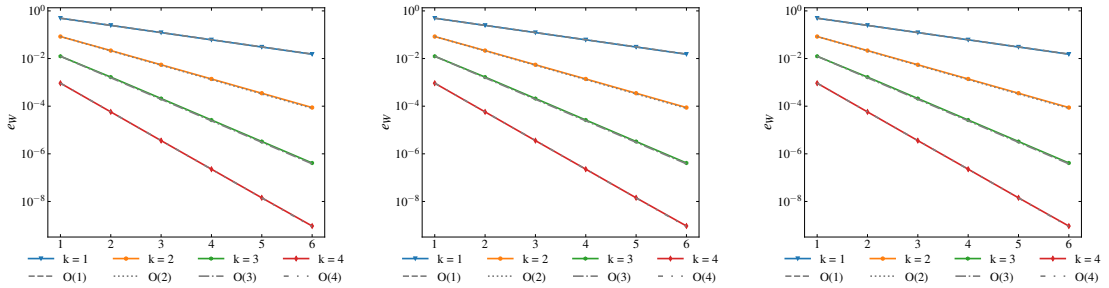
3.6.1 Convergence Studies

Before we present the results of the convergence studies with a non-constant density, we will demonstrate that the geometry switch does not change the observations from section 3.3. To this end, we will first consider the problem with different constant densities $\rho \in \{1, 100, 1000\}$. Figures 3.20 and 3.21 show the errors in the L^2- and the W-norm for velocity field the u_3. In both cases, the errors in the W- and the L^2-norm convergence with an optimal order of k and $k + 1$ respectively. The results for the other two velocity fields are similar and can be found in section C.3.



(a) $\rho = 1$ (b) $\rho = 100$ (c) $\rho = 10000$

Figure 3.20: Numerical errors for u_2 with $\rho \in \{1, 100, 10000\}$ in the L^2-norm.



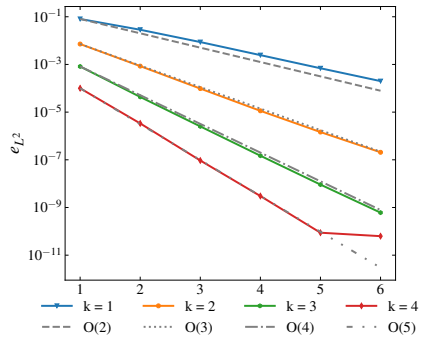
(a) $\rho = 1$ (b) $\rho = 100$ (c) $\rho = 10000$

Figure 3.21: Numerical errors for u_2 with $\rho \in \{1, 100, 10000\}$ in the W-norm.

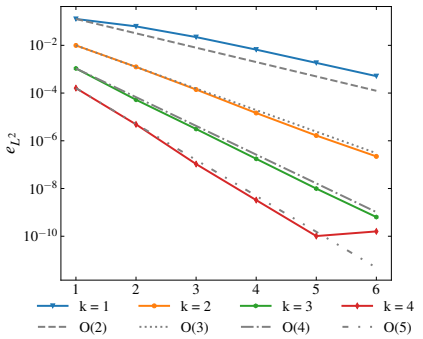
Now, we consider the problem with the non-constant density ρ_n defined by (3.16) with $n \in \{4, 8, 12, 16\}$. Figures 3.22, 3.23 and 3.24 display e_{L^2} for the velocity fields u_1,u_2 and u_3 respectively. We observe that the rates of convergence in the L^2-norm are close to optimal with some variation depending on the velocity field. This makes sense, because as we have seen in section 3.2, the alignment of the facets with respect to the velocity field influences the eoc of e_{L^2}. For u_1 and u_3, the rates in the last refinement step drop as it was the case in section 3.3.
In contrast, the numerical error in the W-norm converges optimally for all three velocity fields. Figure 3.25 displays the errors for u_2; the remaining plots for u_1 and u_3 are similar and can be found in the appendix in C.4.

However, we note that the absolute errors in the W-norm, as seen in table 3.11, increase with increasing exponent n. In contrast, the errors in the $L2$-norm, displayed in table 3.10 do not change with respect to n. This makes sense as the W-norm scales with ρ, while the L^2-norm does not.

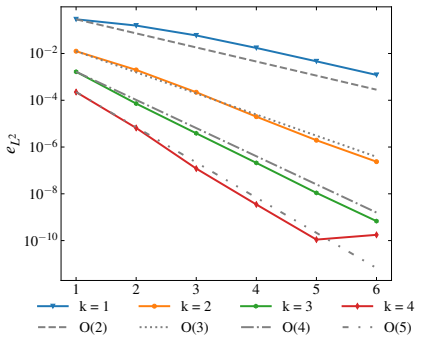
Overall, using a non-constant density ρ does not seem to have a negative impact on the performance of the method, even if the deviation of ρ to a constant inside the domain is large. There are some slight variations in the rate of convergence in the L^2-norm, but this might be caused by the unstructured mesh.



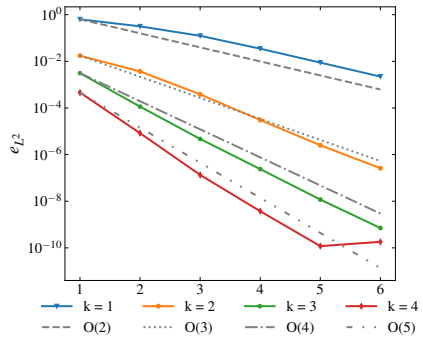
(a) ρ_4



(b) ρ_8

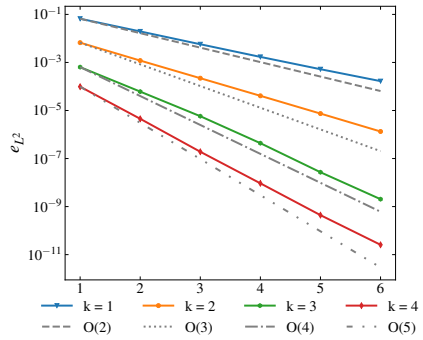


(c) ρ_{12}

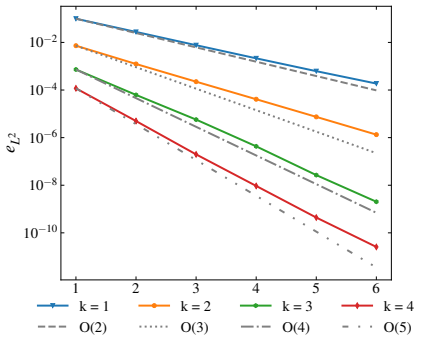


(d) ρ_{16}

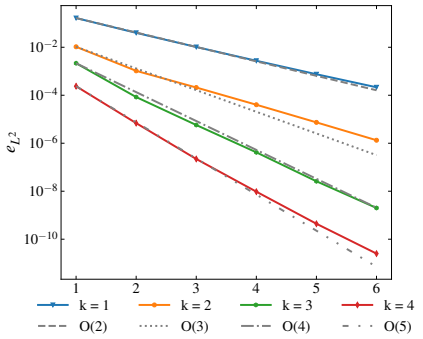
Figure 3.22: Numerical errors for u_1 with and ρ_n with $n \in \{4, 8, 12, 16\}$ in the L^2-norm.



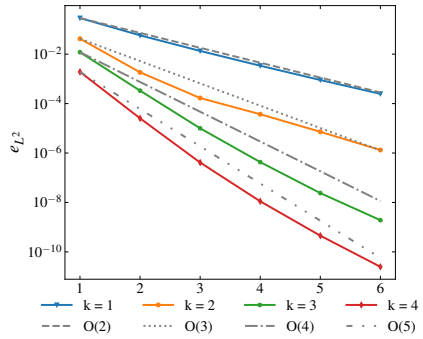
(a) ρ_4



(b) ρ_8

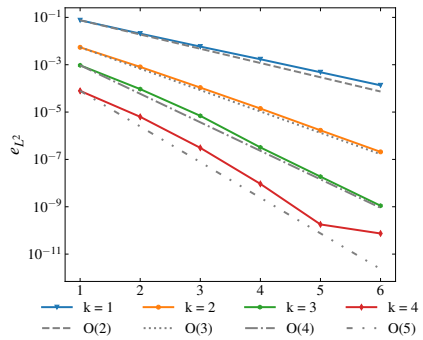


(c) ρ_{12}

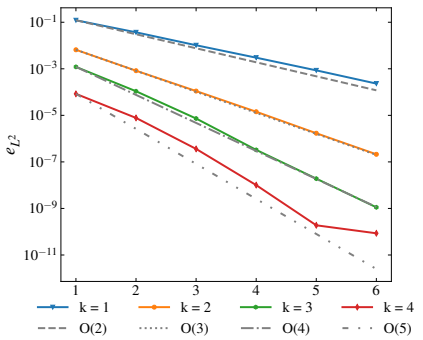


(d) ρ_{16}

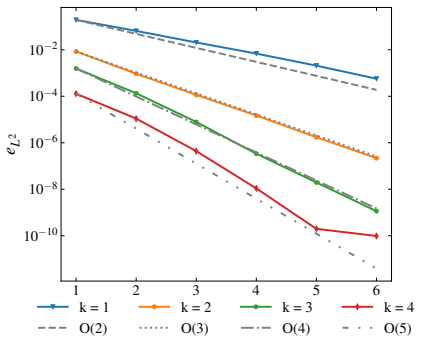
Figure 3.23: Numerical errors for u_2 with and ρ_n with $n \in \{4, 8, 12, 16\}$ in the L^2-norm.



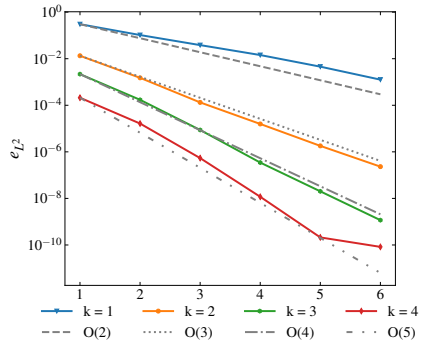
(a) ρ_4



(b) ρ_8

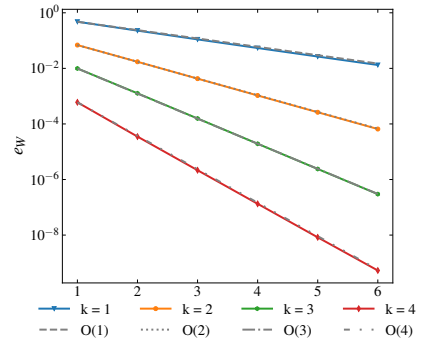


(c) ρ_{12}

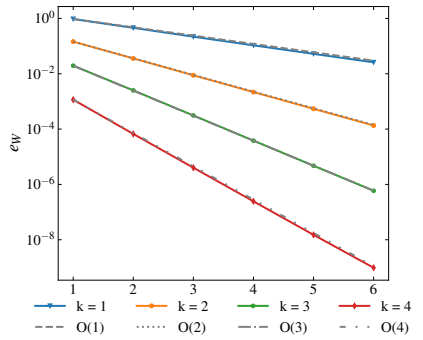


(d) ρ_{16}

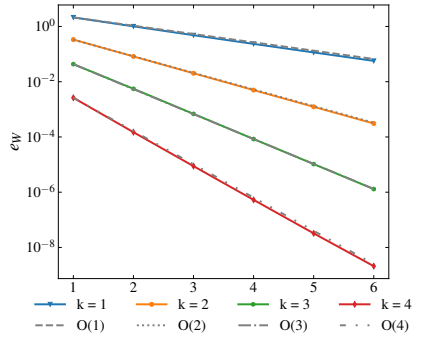
Figure 3.24: Numerical errors for u_3 with and ρ_n with $n \in \{4, 8, 12, 16\}$ in the L^2-norm.



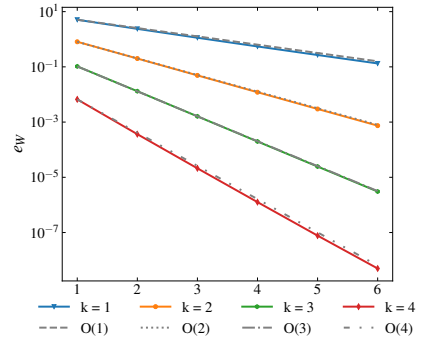
(a) ρ_4



(b) ρ_8



(c) ρ_{12}



(d) ρ_{16}

Figure 3.25: Numerical errors for u_2 with and ρ_n with $n \in \{4, 8, 12, 16\}$ in the W-norm.

	ρ_8		ρ_{16}	
refs	e_{L^2}	(eoc)	e_{L^2}	(eoc)
$k=1$				
1	$9.91 \cdot 10^{-2}$		$2.89 \cdot 10^{-1}$	
2	$2.76 \cdot 10^{-2}$	(1.84)	$5.77 \cdot 10^{-2}$	(2.32)
3	$7.65 \cdot 10^{-3}$	(1.85)	$1.35 \cdot 10^{-2}$	(2.1)
4	$2.15 \cdot 10^{-3}$	(1.83)	$3.40 \cdot 10^{-3}$	(1.99)
5	$6.20 \cdot 10^{-4}$	(1.79)	$8.99 \cdot 10^{-4}$	(1.92)
6	$1.87 \cdot 10^{-4}$	(1.73)	$2.47 \cdot 10^{-4}$	(1.86)
$k=2$				
1	$7.28 \cdot 10^{-3}$		$4.20 \cdot 10^{-2}$	
2	$1.24 \cdot 10^{-3}$	(2.55)	$1.84 \cdot 10^{-3}$	(4.51)
3	$2.24 \cdot 10^{-4}$	(2.47)	$1.68 \cdot 10^{-4}$	(3.46)
4	$4.10 \cdot 10^{-5}$	(2.45)	$3.71 \cdot 10^{-5}$	(2.18)
5	$7.45 \cdot 10^{-6}$	(2.46)	$7.24 \cdot 10^{-6}$	(2.36)
6	$1.34 \cdot 10^{-6}$	(2.48)	$1.33 \cdot 10^{-6}$	(2.45)
$k=3$				
1	$7.29 \cdot 10^{-4}$		$1.20 \cdot 10^{-2}$	
2	$6.24 \cdot 10^{-5}$	(3.55)	$3.34 \cdot 10^{-4}$	(5.16)
3	$5.73 \cdot 10^{-6}$	(3.45)	$1.00 \cdot 10^{-5}$	(5.06)
4	$4.30 \cdot 10^{-7}$	(3.73)	$4.29 \cdot 10^{-7}$	(4.54)
5	$2.67 \cdot 10^{-8}$	(4.01)	$2.39 \cdot 10^{-8}$	(4.16)
6	$2.03 \cdot 10^{-9}$	(3.72)	$1.92 \cdot 10^{-9}$	(3.64)
$k=4$				
1	$1.18 \cdot 10^{-4}$		$1.92 \cdot 10^{-3}$	
2	$4.97 \cdot 10^{-6}$	(4.57)	$2.53 \cdot 10^{-5}$	(6.24)
3	$1.98 \cdot 10^{-7}$	(4.65)	$4.12 \cdot 10^{-7}$	(5.94)
4	$9.34 \cdot 10^{-9}$	(4.4)	$1.11 \cdot 10^{-8}$	(5.21)
5	$4.39 \cdot 10^{-10}$	(4.41)	$4.54 \cdot 10^{-10}$	(4.61)
6	$2.56 \cdot 10^{-11}$	(4.1)	$2.50 \cdot 10^{-11}$	(4.18)

Table 3.10: L^2-convergence table for u_2 with non-constant ρ_n, $n \in \{8, 16\}$.

	ρ_8		ρ_{16}	
refs	e_W	(eoc)	e_W	(eoc)
$k=1$				
1	$9.54 \cdot 10^{-1}$		$5.08 \cdot 10^{0}$	
2	$4.55 \cdot 10^{-1}$	(1.07)	$2.38 \cdot 10^{0}$	(1.1)
3	$2.19 \cdot 10^{-1}$	(1.06)	$1.13 \cdot 10^{0}$	(1.07)
4	$1.06 \cdot 10^{-1}$	(1.04)	$5.48 \cdot 10^{-1}$	(1.05)
5	$5.23 \cdot 10^{-2}$	(1.02)	$2.69 \cdot 10^{-1}$	(1.03)
6	$2.59 \cdot 10^{-2}$	(1.01)	$1.33 \cdot 10^{-1}$	(1.02)
$k=2$				
1	$1.45 \cdot 10^{-1}$		$8.10 \cdot 10^{-1}$	
2	$3.58 \cdot 10^{-2}$	(2.02)	$2.00 \cdot 10^{-1}$	(2.02)
3	$8.81 \cdot 10^{-3}$	(2.02)	$4.91 \cdot 10^{-2}$	(2.03)
4	$2.18 \cdot 10^{-3}$	(2.02)	$1.21 \cdot 10^{-2}$	(2.02)
5	$5.41 \cdot 10^{-4}$	(2.01)	$2.99 \cdot 10^{-3}$	(2.01)
6	$1.35 \cdot 10^{-4}$	(2.01)	$7.42 \cdot 10^{-4}$	(2.01)
$k=3$				
1	$1.95 \cdot 10^{-2}$		$1.04 \cdot 10^{-1}$	
2	$2.51 \cdot 10^{-3}$	(2.96)	$1.31 \cdot 10^{-2}$	(2.99)
3	$3.10 \cdot 10^{-4}$	(3.02)	$1.61 \cdot 10^{-3}$	(3.02)
4	$3.82 \cdot 10^{-5}$	(3.02)	$1.98 \cdot 10^{-4}$	(3.02)
5	$4.73 \cdot 10^{-6}$	(3.01)	$2.45 \cdot 10^{-5}$	(3.01)
6	$5.89 \cdot 10^{-7}$	(3.01)	$3.05 \cdot 10^{-6}$	(3.01)
$k=4$				
1	$1.16 \cdot 10^{-3}$		$6.64 \cdot 10^{-3}$	
2	$6.69 \cdot 10^{-5}$	(4.11)	$3.64 \cdot 10^{-4}$	(4.19)
3	$4.02 \cdot 10^{-6}$	(4.06)	$2.11 \cdot 10^{-5}$	(4.11)
4	$2.45 \cdot 10^{-7}$	(4.04)	$1.26 \cdot 10^{-6}$	(4.06)
5	$1.50 \cdot 10^{-8}$	(4.02)	$7.73 \cdot 10^{-8}$	(4.03)
6	$9.74 \cdot 10^{-10}$	(3.95)	$4.98 \cdot 10^{-9}$	(3.96)

Table 3.11: W-convergence table for u_2 with non-constant ρ_n, $n \in \{8, 16\}$.

3.6.2 Condition numbers for non-constant ρ

To conclude our investigation of the influence of a non-constant density ρ, we will consider the condition numbers of the stiffness matrix as in section 3.4. First, we calculate the conditions numbers for velocity field u_3 with the numerical eigenvalue solver of NGSolve as before. The results are displayed in table 3.12.
Notice that the condition number increases whenever we increase the exponent n in ρ_n. However, the condition number does not grow significantly when refining the mesh. This seems unrealistic and indicates that the condition numbers calculated with NGSolve might not be very exact in this case.

	$\bar{\kappa}(\mathbb{B})$		$\bar{\kappa}(J\mathbb{B})$	
refs	$n=8$	$n=16$	$n=8$	$n=16$
		u_1		
1	8651.06	25290.62	2229.93	3387.13
2	8016.68	21899.12	1861.19	2278.51
3	8679.43	20192.35	1752.74	1854.56
4	8479.85	19137.28	1775.23	1841.85
5	8683.63	18487.89	1883.34	1821.01
6	8734.66	19093.46	1949.56	1903.99
		u_2		
1	6074.63	10109.08	9594.95	10391.57
2	5712.54	10691.52	9679.67	11517.85
3	5327.91	11576.77	8468.4	11316.8
4	5383.58	11261.1	7219.96	12090.28
5	5386.07	11213.81	6511.72	11405.2
6	5393.4	11152.83	6252.85	10038.37
		u_3		
1	5243.02	12600.57	2524.59	6878.67
2	5177.56	12079.49	2424.23	3039.73
3	5299.59	12060.91	2320.72	2451.4
4	5774.93	11743.95	2397.4	2455.26
5	5829.75	11723.35	2509.51	2348.84
6	5892.57	11754.33	2633.9	2492.17

Table 3.12: Condition numbers for $k=4$ with non-constant density ρ_n, $n \in \{8, 16\}$.

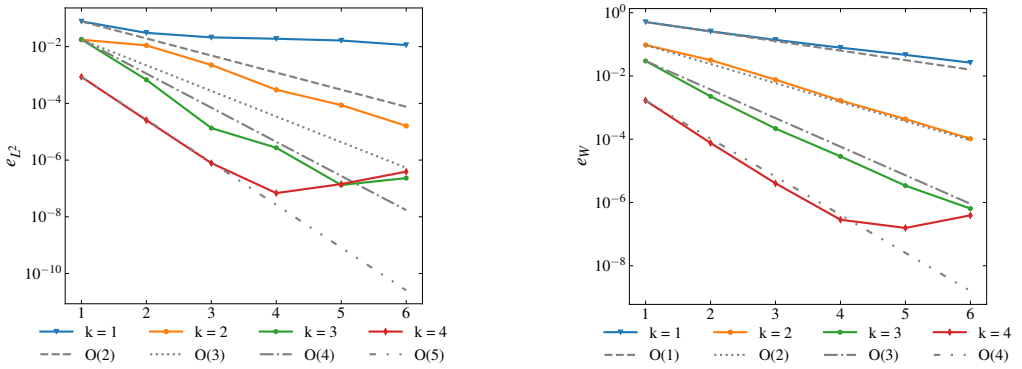
Consequently, in table 3.13 we consider the condition numbers calculated exactly with SciPy. The code can be found in section B.4. Note that this calculation is very expensive, especially for smaller mesh sizes. As such, we only consider the first two refinements levels.

These numbers are significantly greater than the numbers in the previous table. Furthermore, we notice that the condition numbers increase with higher n as well. In particular, when increasing n from 8 to 16, the condition numbers of the matrix $\mathbb{B}$ increase by approximately two orders of magnitude. Using a diagonal preconditioner, this increase gets reduced by approximately an order of magnitude. Nevertheless, using a density that has a large deviation in the domain D might have a negative impact on the condition number.

	$\bar{\kappa}(\mathbb{B})$		$\bar{\kappa}(J\mathbb{B})$	
refs	$n = 8$	$n = 16$	$n = 8$	$n = 16$
1	$2.86 \cdot 10^6$	$1.63 \cdot 10^8$	$6.37 \cdot 10^5$	$3.95 \cdot 10^6$
2	$1.46 \cdot 10^7$	$1.17 \cdot 10^9$	$2.47 \cdot 10^6$	$1.06 \cdot 10^7$

Table 3.13: Condition numbers calculated with SciPy for u_3 and $k = 4$ with a non-constant density ρ_n, $n \in \{8, 16\}$.

Remark 3.2 (Conditioning issues without the ρ-scaling)**.** *Especially when considering either high constant densities or densities that vary highly inside the domain, the ρ-scaling for the volume term becomes vital. For instance, figure 3.26 displays the errors for velocity field u_3 with $\rho = 10000$. Notice that the errors do not convergence optimally, especially for higher polynomial degree.*



(a) Errors in the L^2-norm.

(b) Errors in the W-norm.

Figure 3.26: Numerical errors for u_3 without the ρ-scaling in front of the volume term.

To investigate this result further, we consider the condition numbers of the system matrices with and without the ρ-scaling in front of the volume term. Table 3.14 shows the condition numbers calculated exactly with SciPy. We observe a significant decrease in the condition when the volume term is scaled with ρ. Intuitively, this makes sense. An estimate for the condition numbers is $\frac{\lambda_{\max}}{\lambda_{\min}}$. When only one term of our problem is scaled with ρ, the corresponding eigenvalue also scales with ρ, while the other does not. Hence, the condition numbers are influenced by the density, when the volume term does not scale with ρ.

	$\bar{\kappa}(\mathbb{B})$		$\bar{\kappa}(J\mathbb{B})$	
refs	ρw	w	ρw	w
1	$1.25 \cdot 10^6$	$3.34 \cdot 10^9$	$4.13 \cdot 10^5$	$9.74 \cdot 10^8$
2	$6.08 \cdot 10^6$	$1.53 \cdot 10^{10}$	$1.53 \cdot 10^6$	$6.15 \cdot 10^9$

Table 3.14: Condition numbers calculated with SciPy for u_3 and $k = 4$ with and without the ρ-scaling in front of the volume term.

Chapter 4

Conclusion

To conclude this thesis, this chapter will summarize the achievements and discuss open problems, in terms of the model problem and the broader context.

4.1 Summary

In this thesis we developed and analysed a DG discretization for a degenerate diffusion problem arising in the context of a numerical model for the equations of solar and stellar oscillation.
The developed discretization is consistent, coercive, and continuous, which means that the discrete problem is well-posed. Furthermore, we proved an a priori error estimate that implies an optimal convergence rate of k in an energy-like norm. This rate was also confirmed by the numerical experiments. Because the continuous problem is not necessarily L^2-H^2-regular, we were not able to prove an optimal L^2-error estimate, and we saw in the numerical experiments that it can not be expected in general. However, in the narrow setting of the example consider in this work and with the choice of a suitable triangulation, the numerical experiments yielded an optimal convergence rate in the L^2-norm as well.

Further, we explored, theoretically and numerically, how the penalization parameter might influence the condition number of the system matrices. We presented two possible remedies, a generalized eigenvalue problem and a Bassi-Rebay-stabilization. However, the numerical results indicated that an implementation was not necessary for the considered example problem.
Finally, we investigated if and how the choice of a non-constant density influences the performance of the method. We found that the method performs well with a non-constant density, even if the variation of the density in the domain is large.

4.2 Outlook and open problems

Though we studied the problem intensively, especially in terms of numerical experiments, there are still some issues and improvements left that might be considered in the future. This final section gives an overview over some of those.

First of all, let us start with some straightforward extensions of our problem. For this thesis, we only considered *scalar* functions $w \in W$. In context of the equations of solar and stellar oscillations however, a vectorial model is required. Note, further, that we only considered Dirichlet boundary conditions.

Secondly, we observed that the condition numbers of the system matrices grow depending on the penalization parameter λ and the density ρ. For the former, we already described alternative *stabilization mechanisms* in theory, but did not implement them. However, for different applications this might be beneficial. For the latter, one might consider *preconditioning*. We did not use a diagonal preconditioner for solving the problem, but have already seen in the numerical experiments that it could improve the condition number. Depending on the model problem, other suitable preconditioners might be helpful for solving the problem efficiently.

Furthermore, optimality of convergence rates in the L^2-norm may depend on the mesh and the velocity field. If one considers a different domain D or more complex velocity fields and still wants to achieve optimal convergence in the L^2-norm, one has to be careful when choosing a triangulation of D and even then it is not clear if near best approximation results can be expected.

Finally, there are modifications of DG discretizations that could be considered. One could, for example, formulate the problem as an *Hybrid Discontinuous Galerkin* (HDG) method. In short, HDG methods introduce additional facet unknowns $w_F \in L^2(\mathcal{F}_h)$. This improves computational efficiency as neighbouring elements do not couple directly any more, which means that they can be eliminated by static condensation. For a more extensive overview, we refer to [Leh10].

After giving a short overview over these possible extensions, let us describe what steps are necessary to remove some simplifications that we introduced in the introduction. As described there, the main goal is to develop a numerical model for the equations of solar and stellar oscillations.

Recall that we defined the bilinear form $\tilde{a}_h(\cdot,\cdot)$ through

$$\tilde{a}_h(v_h, v_h) := \langle \rho c^2 \nabla \cdot v_h, \nabla \cdot v_h \rangle + a_h^2(v_h, v_h) \quad \forall v_h \in \mathbf{V}_h, \tag{4.1}$$

where $\mathbf{V}_h := \{\xi_h \in \mathbf{X}_h \mid \nabla \cdot \xi_h = 0\}^\perp$.
The next step is to show inf-sup stability of this bilinear form. Afterwards, one can continue by reintroducing the omitted zero-order terms and considering the case, where the pressure is not constant any more and thus $\mathbf{q} \neq 0$.

Chapter A

On triangulations and elementary inequalities

This chapter contains some elementary results on triangulations and inequalities to which we refer throughout the thesis. All statements are well-known and, thus, given without proofs.

A.1 Triangulations

Triangulations are essential for finite element methods. However, notation is not always consistent. As such, we will give a brief introduction before stating the required results. Note that this chapter largely follows the lecture notes by Lehrenfeld [Leh21, Section 4.1].

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain that can be decomposed into a finite number of subsets T. We denote the collection of these sets as $\mathcal{T}_h = \{T\}$. We call $\mathcal{T}_h$ an admissible triangulation if

- $\overline{D} = \bigcup_{T \in \mathcal{T}_h} T$,
- $\text{int}(T_1) \cap \text{int}(T_2) = \emptyset \qquad \forall T_1, T_2 \in \mathcal{T}_h$,
- for any facet $F \in \partial T_1$ of any $T_1 \in Th$ there holds either $F \in \partial D$ or $F \in \partial T_2$ for a $T_2 \in Th$.

In the following, we will denote the length of an element as h_T and the radius of the largest ball contained in an element as ρ_T. Precisely, this means that we define

$$h_T := \text{diam}(T) = \sup_{x,y \in T} \|x - y\|,$$
$$\rho_T := \sup\{\text{diam}(B) \mid B \text{ is a Ball contained in } T\}.$$

The triangulation $\mathcal{T}_h$ is shape-regular, if there exits a $\sigma > 0$ such that

$$\sigma_T := \frac{h_T}{\rho_T} \leq \sigma \qquad \forall T \in \mathcal{T}_h, \tag{A.1}$$

and quasi-uniform, if there holds

$$h_T \simeq h := \max_{T \in \mathcal{T}_h} h_T. \tag{A.2}$$

In this thesis, we only consider simplex triangulations, but in general other types of elements are possible. For every dimension we can define a reference simplex $\widehat{T} \subset \mathbb{R}^d$.For instance, in two dimensions the reference simplex is given by

$$\widehat{T} = \text{conv}\{(0,0)^T, (1,0)^T, (0,1)^T\}.$$

The following results will be useful when proving the inverse inequality in section 2.5:

Lemma A.1. *Every non-degenerate simplex T in $\mathbb{R}^d$ is affine equivalent to the reference simplex $\widehat{T}$ in $\mathbb{R}^d$, i.e. there exits an affine mapping $\Phi : \widehat{T} \to T$, $\Phi(x) = Ax + b$ with $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$, $\det(A) \neq 0$. Further, Φ is invertible and there holds*

$$\begin{aligned} \|D\Phi\| = \|A\| &\leq h_T/\rho_{\widehat{T}}; \\ \|D\Phi^{-1}\| = \|A^{-1}\| &\leq h_{\widehat{T}}/\rho_T; \\ c\rho_T^d \leq |\det(D\Phi)| = |\det(A)| &\leq Ch_T^d. \end{aligned} \tag{A.3}$$

for some constants $0 < c \leq C$.

Lemma A.2. *Let $T, \widehat{T} \subset D$ be open, bounded and affine equivalent. There holds for $u \in H^m(T)$ and $\widehat{u} := u \circ \Phi \in H^m(\widehat{T})$ that*

$$\begin{aligned} |\widehat{u}|_{H^m(\widehat{T})} &\lesssim \|A\|^m |\det(A)|^{-\frac{1}{2}} |u|_{H^m(T)}, \\ |u|_{H^m(T)} &\lesssim \|A^{-1}\|^m |\det(A)|^{\frac{1}{2}} |\widehat{u}|_{H^m(\widehat{T})}. \end{aligned} \tag{A.4}$$

From the previous lemma we can infer that

$$|\widehat{u}|_{H^m(\widehat{T})} \lesssim \Big(\frac{h_T}{\rho_{\widehat{T}}}\Big)^m \rho_T^{-\frac{d}{2}} |u|_{H^m(T)}, \tag{A.5}$$

and similarly

$$|u|_{H^m(T)} \lesssim \Big(\frac{h_{\widehat{T}}}{\rho_T}\Big)^m h_T^{\frac{d}{2}} |\widehat{u}|_{H^m(\widehat{T})}. \tag{A.6}$$

Note that the implied constants only depend on the shape regularity and on m.

A.2 Inequalities

Especially in the error analysis in chapter 2, we use the following two elementary inequality repeatedly.

Lemma A.3 (Cauchy-Schwarz inequality)**.** *For the L^2-scalar product and $u, v \in L^2(D)$ there holds that*

$$\langle u, v\rangle_D \leq \|u\|_D \|v\|_D. \tag{A.7}$$

A special case for the l^2-scalar product $(a, b)_2 := \sum_{i=1}^n a_i b_i$ where $a, b \in \mathbb{R}^n$ is

$$\sum_{i=1}^n a_i b_i \leq \Big(\sum_{i=1}^n a_i\Big)^{\frac{1}{2}} \Big(\sum_{i=1}^n b_i\Big)^{\frac{1}{2}}. \tag{A.8}$$

Lemma A.4 (Young's inequality)**.** *For $a, b, \gamma \in \mathbb{R}$ there holds that*

$$ab \leq \frac{a^2}{2\gamma} + \frac{\gamma b^2}{2} \tag{A.9}$$

Chapter B

Code

This chapter contains the code used for evaluating the method numerically, cf. chapter 3. We focus on the parts that are tailored to the problem and omit the general parts, such as looping over different levels of refinements or penalization parameters. The implementation relies on the NGSolve library. Note that in the code, the bilinear form is denoted as `a` and the test- and trial functions as `w` and `v`.
Two explanatory Jupyter notebooks containing the code from this section can be found on my GitHub account.

B.1 Calculation of a source term f

This python code calculates the source term f for a given velocity field u, a given exact solution w and a given density ρ for $D \subset \mathbb{R}^2$. Then, f is given by

$$f = \rho w - \nabla \cdot (\rho(u \otimes u)\nabla w)$$

```
from ngsolve import *

def calculate_rhs(u,exact,rho):
    umat = CoefficientFunction(u,dims=(2,1))
    uTen = umat*umat.trans
    exactGrad = (exact.Diff(x),exact.Diff(y))
    exactDiffusion = rho*uTen*exactGrad
    exactDiv = exactDiffusion[0].Diff(x)+exactDiffusion[1].Diff(y)
    rhs = rho*exact-exactDiv
    return rhs, exactGrad
```

B.2 Mesh generation

The following code block generates a structured mesh as seen in figure 3.5a. The number of triangles that the square is decomposed into is controlled by the parameter `n`. Thus, increasing `n` allows refining the mesh.

```
from netgen.geom2d import SplineGeometry
from ngsolve import *
from ngsolve.meshes import *

n = 1
mesh = MakeStructured2DMesh(quads=False, nx=2**n, ny=2**n,mapping =
    ↪ lambda x,y: (2*x-1,2*y-1))
```

To generate an unstructured mesh as seen in figure 3.5b, this code can be used. The parameter `maxh` sets the initial mesh size and the parameter `refs` controls the number of mesh refinements.

```
from netgen.geom2d import SplineGeometry
from ngsolve import *

refs = 1
ngmesh = square.GenerateMesh(maxh=0.7)
mesh = Mesh(ngmesh)
for i in range(refs):
    ngmesh.Refine()
mesh = Mesh(ngmesh)
```

Finally, with this code we can generate the circle geometry used in section 3.6. As before, `maxh` and `refs` can be used to set the initial mesh size and the number of refinements respectively.

```
from netgen.geom2d import SplineGeometry
from ngsolve import *

refs = 1
maxh = 1
geo = SplineGeometry()
geo.AddCircle(c=(0,0),r=1,bc="circle")
ngmesh = geo.GenerateMesh(maxh=maxh)
for i in range(refs):
    ngmesh.Refine()
mesh = Mesh(ngmesh)
mesh.Curve(7)
```

B.3 Solving the problem

The previous code blocks build the preliminary structures which are necessary to solve the problem with the following code. The main function `Solve()` needs the polynomial degree, the exact solution, the density ρ, a velocity field u, a penalization parameter λ and a mesh as input. It offers further the option to calculate the best L^2-approximation on the exact

solution and to estimate the condition numbers of both, the assembled matrix $\mathbb{A}$ and the diagonal preconditioned matrix.
Though we will not describe the specifics of the NGSolve library further, we would like to highlight two peculiarities:

- To reduce the quadrature error, we increased the order of numerical integration with the following code:

  ```
  dX = dx(bonus_intorder=2),
  dS = dx(bonus_intorder=2).
  ```

  Note that this is in particular required when ρ is not constant.

- When defining the bilinear form, we have to add additional terms to include the boundary facets. The integration operator `dX(skeleton=True)` only sums over the interior facets and hence the terms ending with `dS(skeleton=True)` are required to account for the non-homogeneous Dirichlet boundary conditions.

- The condition numbers are estimated as the ratio of the largest and the smallest eigenvalue, which are calculated with NGSolve. To calculate the condition number of the preconditioned stiffness matrix, we decompose $\mathbb{B}$ into degrees of freedom associated with the Dirichlet boundary conditions and the remainder:

  $$\mathbb{B} = \begin{pmatrix} \mathbb{B}_{FF} & \mathbb{B}_{FD} \\ \mathbb{B}_{DF} & \mathbb{B}_{DD} \end{pmatrix}. \tag{B.1}$$

  Then, we define

  $$J := \begin{pmatrix} \operatorname{diag}(\mathbb{B}_{FF})^{-1} & 0 \\ 0 & 0 \end{pmatrix} \tag{B.2}$$

  and consider $\kappa(J\mathbb{B})$.
  Note that the conditions numbers calculated in this way might not be accurate in some cases. For an exact calculation, which is computational expensive, of the condition numbers with SciPy, we can use the code from section B.4.

```
from ngsolve import *
from ngsolve.la import EigenValues_Preconditioner
from calculateRHS import calculate_rhs

du = lambda u,w: InnerProduct(u,grad(w))

def Solve(k,exact,rho,uC,lamb,mesh,bl2 = False,cn = False):
    fes = L2(mesh, order=k, dgjumps=True)
    w,v = fes.TnT()
    gfu = GridFunction(fes)

    n = specialcf.normal(2)
    h = specialcf.mesh_size

    jump_w = w-w.Other()
    jump_v = v-v.Other()
    avg_duw = 0.5*(du(uC,w)+du(uC,w.Other()))
```

```
    avg_duv = 0.5*(du(uC,v)+du(uC,v.Other()))

    dX = dx(bonus_intorder = 2)
    dS = ds(bonus_intorder = 2)

    rhs,exactGrad = calculate_rhs(uC,exact,rho,rhoVol)

    a = BilinearForm(fes,symmetric=True)
    a += rho*w*v*dx
    a += rho*du(uC,w)*du(uC,v)*dX
    a += uC*n*-rho*avg_duw*jump_v*dX(skeleton=True)
    a += uC*n*-rho*avg_duv*jump_w*dX(skeleton=True)
    a += rho*lamb*1/h*(uC*n)*(uC*n)*jump_w*jump_v*dX(skeleton=True)
    a += uC*n*-rho*du(uC,w)*v*dS(skeleton=True)
    a += uC*n*-rho*du(uC,v)*w*dS(skeleton=True)
    a += rho*lamb*1/h*(uC*n)*(uC*n)*w*v*dS(skeleton=True)

    f = LinearForm(fes)
    f += rhs*v*dX
    f += uC*n*-rho*du(uC,v)*exact*dS(skeleton=True)
    f += rho*lamb*1/h*(uC*n)*(uC*n)*exact*v*dS(skeleton=True)

    a.Assemble()
    f.Assemble()

    aInv = a.mat.Inverse(freedofs=fes.FreeDofs(),inverse="
        ↪ sparsecholesky")
    gfu.vec[:] = 0.0
    gfu.vec.data = aInv * f.vec

    if cn == True:
        lams = EigenValues_Preconditioner(a.mat,IdentityMatrix(a.
            ↪ space.ndof))
        Prelams = EigenValues_Preconditioner(a.mat,a.mat.
            ↪ CreateSmoother())
        cond = lams[-1]/lams[0]
        condPre = Prelams[-1]/Prelams[0]
    else:
        cond = 'N/A'
        condPre = 'N/A'

    if bl2 == True:
        w1,v1 = fes.TnT()
        gfu2 = GridFunction(fes)

        b = BilinearForm(fes,symmetric=True)
        b += w1*v1*dX

        l = LinearForm(fes)
```

```
        l += exact*v1*dX

        b.Assemble()
        l.Assemble()

        gfu2.vec[:] = 0.0
        gfu2.vec.data = b.mat.Inverse(freedofs=fes.FreeDofs(),inverse
            ↪ ="sparsecholesky")*l.vec
        return gfu,gfu2,exactGrad, cond, condPre
    else:
        return gfu, exactGrad, cond, condPre
```

B.4 Condition numbers with SciPy

This code block calculates the condition numbers using the software package SciPy[1]. The first part calculated the condition number system matrix $\mathbb{A}$, which is obtained from NGSolve. Afterwards, a diagonal preconditioner J is constructed and the condition number of $J\mathbb{A}$ is calculated.

```
import scipy
import scipy.sparse as sp
import scipy.sparse.linalg
import numpy as np

rows,cols,vals = a.mat.COO()
A = sp.csr_matrix((vals,(rows,cols)))
m = A.todense()
cond = np.linalg.cond(m)

diags = np.diagonal(m)
mdiag = sp.diags(diags,offsets=0)
mInv = sp.linalg.inv(mdiag)
PA = np.dot(mInv,A)
precond= np.linalg.cond(PA.todense())
```

[1] can be found at https://www.scipy.org.

Chapter C

Convergence tables and Plots

This section contains supplementary material to the numerical experiments in chapter 3. Often, when the results are similar for all three velocity fields, we only display them for one velocity field to keep the chapter more readable, but refer to this chapter for the remaining results.

C.1 Condition numbers for different penalization parameter λ

In 3.4 we investigated the influence of the penalization parameter λ on the condition numbers of the matrix $\mathbb{B}$ and the diagonal preconditioned matrix $J\mathbb{B}$. Here, the results for the remaining polynomial degrees $k = 2$ and $k = 3$ are given.

velocity field	u_1		u_2		u_3	
λ	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$
1	(-0.96)	(-0.99)	(-1.09)	(-1.77)	(-1.3)	(-1.97)
2	(-1.34)	(-1.0)	(-1.77)	(-7.82)	(-2.66)	(-1.73)
4	(-3.28)	(-2.44)	(-6.41)	(-0.24)	(-34.34)	(-13.81)
8	(-92.62)	(-85.59)	4854.36	3050.74	6206.03	3382.45
16	15177.8	13494.93	4641.7	2582.09	5731.83	3256.88
32	15187.75	12670.07	4630.54	2500.5	5449.14	3321.56
64	10809.05	9078.97	4752.74	2629.86	5318.47	3272.27
128	7771.74	8207.32	5037.85	2572.77	5508.3	2901.07
256	3174.39	5249.23	5819.9	2780.41	5862.26	3134.53
512	4693.75	2504.93	7522.26	3066.24	6285.11	3735.4
1024	6893.6	3573.18	10866.9	3804.44	7956.02	3916.73
2048	7266.88	6885.37	16787.84	5251.21	12424.41	5463.92
4096	10528.21	13344.03	24229.23	8722.69	19762.32	8950.7
8192	19697.0	23272.96	32307.2	13627.69	29394.07	16125.51

Table C.1: Condition numbers for $k = 2$ with different λ.

velocity field	u_1		u_2		u_3	
λ	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$	$\bar{\kappa}(\mathbb{B})$	$\bar{\kappa}(J\mathbb{B})$
1	(-0.77)	(-124.45)	(-0.86)	(-1.33)	(-0.86)	(-0.95)
2	(-0.93)	(-1.0)	(-1.13)	(-1.54)	(-1.17)	(-0.93)
4	(-1.33)	(-4.18)	(-2.03)	(-3.34)	(-2.83)	(-2.63)
8	(-3.94)	(-2.89)	(-9.46)	(-1.73)	(-89.07)	(-36.3)
16	10822.41	8440.66	3797.05	2152.13	4372.28	2843.0
32	10307.64	6868.96	3798.41	2230.3	4181.91	2853.16
64	6891.96	5041.76	3808.07	2427.73	4197.38	2674.81
128	4343.08	2984.17	4150.5	2473.97	4320.57	3043.0
256	3773.07	2015.01	4972.59	2788.08	4523.8	3065.25
512	4545.36	2610.42	6533.09	3296.44	5066.18	3352.56
1024	7039.35	3724.68	9159.1	3464.68	6889.09	3553.75
2048	9499.21	6717.17	12890.19	4307.31	10192.29	4310.72
4096	11460.26	12672.27	17038.71	6941.39	14701.37	7246.98
8192	19787.87	20911.2	20715.95	12098.72	19788.33	11982.5

Table C.2: Condition numbers for $k = 3$ with different λ.

C.2 Convergence tables for the problem without the volume term

This section contains the convergence tables for the problem in section 3.5, where we only consider the diffusion term. These tables supplement the plots 3.13, 3.14, and 3.15 and are in accordance with the conclusion, that the problem is only well-posed for u_1.

refs	e_{L^2}	(eoc)	e_W	(eoc)
$k = 1$				
1	$5.47 \cdot 10^{-1}$		$3.00 \cdot 10^{0}$	
2	$2.73 \cdot 10^{-1}$	(1.0)	$1.44 \cdot 10^{0}$	(1.06)
3	$1.77 \cdot 10^{-1}$	(0.63)	$9.86 \cdot 10^{-1}$	(0.54)
4	$8.17 \cdot 10^{-2}$	(1.11)	$5.40 \cdot 10^{-1}$	(0.87)
5	$2.96 \cdot 10^{-2}$	(1.47)	$2.73 \cdot 10^{-1}$	(0.98)
6	$8.73 \cdot 10^{-3}$	(1.76)	$1.33 \cdot 10^{-1}$	(1.03)
7	$2.31 \cdot 10^{-3}$	(1.92)	$6.51 \cdot 10^{-2}$	(1.04)
8	$5.85 \cdot 10^{-4}$	(1.98)	$3.20 \cdot 10^{-2}$	(1.02)
$k = 2$				
1	$2.84 \cdot 10^{-1}$		$1.39 \cdot 10^{0}$	
2	$9.80 \cdot 10^{-2}$	(1.53)	$6.60 \cdot 10^{-1}$	(1.07)
3	$1.79 \cdot 10^{-2}$	(2.45)	$1.84 \cdot 10^{-1}$	(1.84)
4	$1.77 \cdot 10^{-3}$	(3.34)	$4.73 \cdot 10^{-2}$	(1.96)
5	$1.52 \cdot 10^{-4}$	(3.54)	$1.18 \cdot 10^{-2}$	(2.0)
6	$1.43 \cdot 10^{-5}$	(3.42)	$2.94 \cdot 10^{-3}$	(2.0)
7	$1.57 \cdot 10^{-6}$	(3.18)	$7.35 \cdot 10^{-4}$	(2.0)
8	$1.89 \cdot 10^{-7}$	(3.05)	$1.84 \cdot 10^{-4}$	(2.0)
$k = 3$				
1	$1.23 \cdot 10^{-1}$		$9.23 \cdot 10^{-1}$	
2	$1.92 \cdot 10^{-2}$	(2.68)	$1.60 \cdot 10^{-1}$	(2.53)
3	$8.96 \cdot 10^{-4}$	(4.42)	$2.47 \cdot 10^{-2}$	(2.7)
4	$4.06 \cdot 10^{-5}$	(4.46)	$3.15 \cdot 10^{-3}$	(2.97)
5	$2.21 \cdot 10^{-6}$	(4.2)	$3.95 \cdot 10^{-4}$	(3.0)
6	$1.35 \cdot 10^{-7}$	(4.03)	$4.94 \cdot 10^{-5}$	(3.0)
7	$8.40 \cdot 10^{-9}$	(4.01)	$6.16 \cdot 10^{-6}$	(3.0)
8	$5.37 \cdot 10^{-1}$	(3.97)	$7.70 \cdot 10^{-7}$	(3.0)
$k = 4$				
1	$5.78 \cdot 10^{-2}$		$3.52 \cdot 10^{-1}$	
2	$2.88 \cdot 10^{-3}$	(4.32)	$2.91 \cdot 10^{-2}$	(3.6)
3	$1.73 \cdot 10^{-4}$	(4.06)	$2.53 \cdot 10^{-3}$	(3.52)
4	$4.41 \cdot 10^{-6}$	(5.3)	$1.50 \cdot 10^{-4}$	(4.07)
5	$1.29 \cdot 10^{-7}$	(5.1)	$9.24 \cdot 10^{-6}$	(4.02)
6	$3.95 \cdot 10^{-9}$	(5.03)	$5.73 \cdot 10^{-7}$	(4.01)
7	$1.44 \cdot 10^{-1}$	(4.78)	$3.57 \cdot 10^{-8}$	(4.0)
8	$2.93 \cdot 10^{-1}$	(-1.03)	$2.61 \cdot 10^{-9}$	(3.77)

Table C.3: Convergence table for u_1 without the volume term.

refs	e_{L^2}	(eoc)	e_W	(eoc)
$k = 1$				
1	$5.18 \cdot 10^{-1}$		$1.08 \cdot 10^{0}$	
2	$2.67 \cdot 10^{-1}$	(0.96)	$6.33 \cdot 10^{-1}$	(0.77)
3	$1.30 \cdot 10^{-1}$	(1.04)	$3.79 \cdot 10^{-1}$	(0.74)
4	$4.63 \cdot 10^{-2}$	(1.49)	$1.84 \cdot 10^{-1}$	(1.04)
5	$1.48 \cdot 10^{-2}$	(1.64)	$8.66 \cdot 10^{-2}$	(1.09)
6	$4.33 \cdot 10^{-3}$	(1.78)	$4.08 \cdot 10^{-2}$	(1.09)
7	$1.17 \cdot 10^{-3}$	(1.88)	$1.95 \cdot 10^{-2}$	(1.06)
8	$3.04 \cdot 10^{-4}$	(1.95)	$9.48 \cdot 10^{-3}$	(1.04)
$k = 2$				
1	$3.29 \cdot 10^{-1}$		$7.51 \cdot 10^{-1}$	
2	$7.48 \cdot 10^{-2}$	(2.14)	$2.50 \cdot 10^{-1}$	(1.59)
3	$1.16 \cdot 10^{-2}$	(2.69)	$5.99 \cdot 10^{-2}$	(2.06)
4	$1.44 \cdot 10^{-3}$	(3.01)	$1.44 \cdot 10^{-2}$	(2.05)
5	$1.47 \cdot 10^{-4}$	(3.3)	$3.50 \cdot 10^{-3}$	(2.04)
6	$1.56 \cdot 10^{-5}$	(3.23)	$8.66 \cdot 10^{-4}$	(2.02)
7	$2.26 \cdot 10^{-6}$	(2.79)	$2.16 \cdot 10^{-4}$	(2.0)
8	$2.92 \cdot 10^{-5}$	(-3.69)	$8.28 \cdot 10^{-5}$	(1.38)
$k = 3$				
1	$1.35 \cdot 10^{-1}$		$3.80 \cdot 10^{-1}$	
2	$1.79 \cdot 10^{-2}$	(2.91)	$6.97 \cdot 10^{-2}$	(2.44)
3	$1.16 \cdot 10^{-3}$	(3.95)	$8.67 \cdot 10^{-3}$	(3.01)
4	$8.58 \cdot 10^{-5}$	(3.76)	$1.09 \cdot 10^{-3}$	(3.0)
5	$5.12 \cdot 10^{-6}$	(4.07)	$1.34 \cdot 10^{-4}$	(3.02)
6	$2.76 \cdot 10^{-5}$	(-2.43)	$4.38 \cdot 10^{-5}$	(1.61)
7	$6.06 \cdot 10^{-3}$	(-7.78)	$6.06 \cdot 10^{-3}$	(-7.11)
8	$9.11 \cdot 10^{-2}$	(-3.91)	$9.11 \cdot 10^{-2}$	(-3.91)
$k = 4$				
1	$7.77 \cdot 10^{-2}$		$1.75 \cdot 10^{-1}$	
2	$3.66 \cdot 10^{-3}$	(4.41)	$1.15 \cdot 10^{-2}$	(3.92)
3	$1.79 \cdot 10^{-4}$	(4.36)	$7.67 \cdot 10^{-4}$	(3.91)
4	$1.18 \cdot 10^{-5}$	(3.92)	$5.21 \cdot 10^{-5}$	(3.88)
5	$1.53 \cdot 10^{-3}$	(-7.02)	$1.53 \cdot 10^{-3}$	(-4.88)
6	$1.22 \cdot 10^{-1}$	(-6.31)	$1.22 \cdot 10^{-1}$	(-6.31)
7	$1.80 \cdot 10^{-1}$	(-0.56)	$1.80 \cdot 10^{-1}$	(-0.56)
8	$1.86 \cdot 10^{-1}$	(-0.05)	$1.86 \cdot 10^{-1}$	(-0.05)

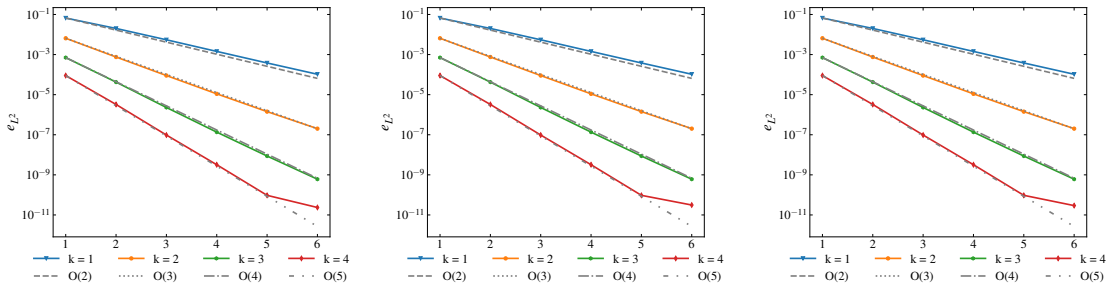
Table C.4: Convergence table for u_2 without the volume term.

refs	e_{L^2}	(eoc)	e_W	(eoc)
$k = 1$				
1	$6.31 \cdot 10^{-1}$		$1.89 \cdot 10^{0}$	
2	$4.66 \cdot 10^{-1}$	(0.44)	$1.26 \cdot 10^{0}$	(0.58)
3	$8.17 \cdot 10^{-1}$	(-0.81)	$1.39 \cdot 10^{0}$	(-0.13)
4	$1.11 \cdot 10^{-1}$	(2.87)	$4.23 \cdot 10^{-1}$	(1.71)
5	$2.85 \cdot 10^{-1}$	(-1.35)	$4.49 \cdot 10^{-1}$	(-0.09)
6	$1.20 \cdot 10^{-1}$	(1.24)	$2.04 \cdot 10^{-1}$	(1.13)
7	$1.58 \cdot 10^{-1}$	(-0.39)	$2.00 \cdot 10^{-1}$	(0.03)
8	$8.11 \cdot 10^{-2}$	(0.96)	$1.02 \cdot 10^{-1}$	(0.97)
$k = 2$				
1	$1.19 \cdot 10^{3}$		$1.93 \cdot 10^{3}$	
2	$8.41 \cdot 10^{1}$	(3.82)	$1.10 \cdot 10^{2}$	(4.13)
3	$2.34 \cdot 10^{0}$	(5.17)	$2.70 \cdot 10^{0}$	(5.35)
4	$8.13 \cdot 10^{-1}$	(1.52)	$8.80 \cdot 10^{-1}$	(1.62)
5	$2.86 \cdot 10^{0}$	(-1.81)	$2.87 \cdot 10^{0}$	(-1.71)
6	$7.62 \cdot 10^{0}$	(-1.41)	$7.62 \cdot 10^{0}$	(-1.41)
7	$1.67 \cdot 10^{-2}$	(8.83)	$1.72 \cdot 10^{-2}$	(8.79)
8	$1.06 \cdot 10^{-2}$	(0.66)	$1.07 \cdot 10^{-2}$	(0.69)
$k = 3$				
1	$3.70 \cdot 10^{1}$		$3.70 \cdot 10^{1}$	
2	$5.75 \cdot 10^{6}$	(12.65)	$5.75 \cdot 10^{6}$	(12.65)
3	$1.22 \cdot 10^{-1}$	(25.49)	$1.39 \cdot 10^{-1}$	(25.3)
4	$1.10 \cdot 10^{0}$	(-3.17)	$1.10 \cdot 10^{0}$	(-2.98)
5	$1.18 \cdot 10^{0}$	(-0.1)	$1.18 \cdot 10^{0}$	(-0.09)
6	$1.53 \cdot 10^{-1}$	(2.94)	$1.54 \cdot 10^{-1}$	(2.94)
7	$7.81 \cdot 10^{-2}$	(0.97)	$7.82 \cdot 10^{-2}$	(0.97)
8	$7.71 \cdot 10^{-2}$	(0.02)	$7.71 \cdot 10^{-2}$	(0.02)
$k = 4$				
1	$3.30 \cdot 10^{9}$		$3.30 \cdot 10^{9}$	
2	$2.48 \cdot 10^{5}$	(13.7)	$2.48 \cdot 10^{5}$	(13.7)
3	$8.40 \cdot 10^{-1}$	(18.17)	$8.52 \cdot 10^{-1}$	(18.15)
4	$2.55 \cdot 10^{-1}$	(1.72)	$2.56 \cdot 10^{-1}$	(1.74)
5	$2.69 \cdot 10^{-1}$	(-0.08)	$2.69 \cdot 10^{-1}$	(-0.07)
6	$2.64 \cdot 10^{-1}$	(0.03)	$2.64 \cdot 10^{-1}$	(0.03)
7	$2.41 \cdot 10^{-1}$	(0.13)	$2.41 \cdot 10^{-1}$	(0.13)
8	$1.70 \cdot 10^{-1}$	(0.5)	$1.70 \cdot 10^{-1}$	(0.5)

Table C.5: Convergence table for u_3 without the volume term.

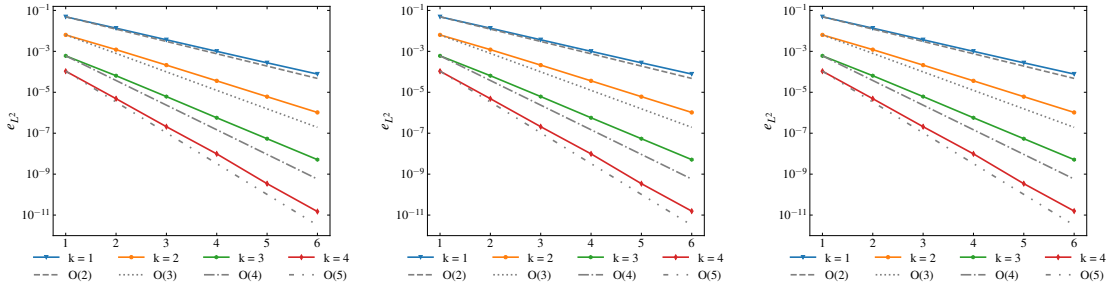
C.3 Convergence plots for constant densities on the circle geometry

In section 3.6.1, we demonstrated that the geometry switch to the circle does not influence the performance of the method for constant densities. There, we only displayed the errors for the velocity field u_3. Here, the errors for the remaining velocity fields can be found.



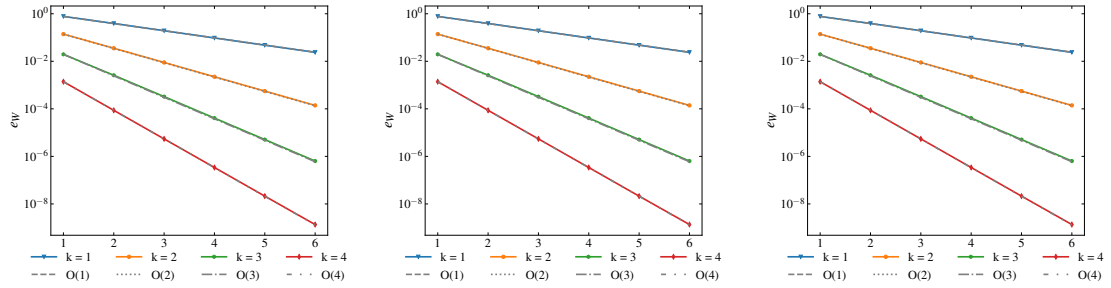
(a) $\rho = 1$ (b) $\rho = 100$ (c) $\rho = 10000$

Figure C.1: Numerical errors for u_1 with $\rho \in \{1, 100, 10000\}$ in the L^2-norm.



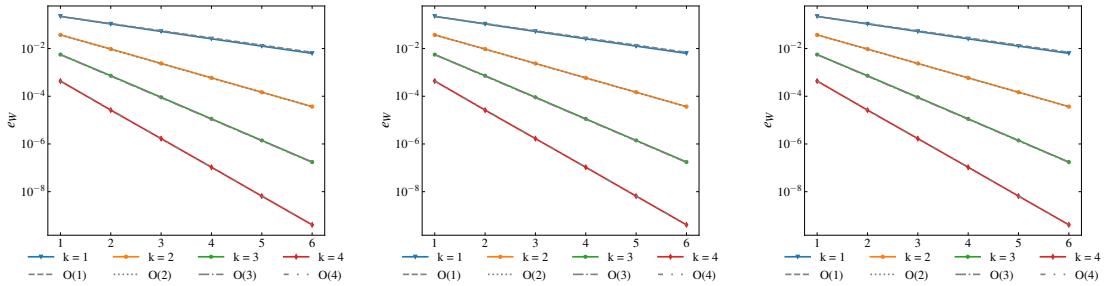
(a) $\rho = 1$ (b) $\rho = 100$ (c) $\rho = 10000$

Figure C.2: Numerical errors for u_2 with $\rho \in \{1, 100, 10000\}$ in the L^2-norm.



(a) $\rho = 1$ (b) $\rho = 100$ (c) $\rho = 10000$

Figure C.3: Numerical errors for u_1 with $\rho \in \{1, 100, 10000\}$ in the W-norm.

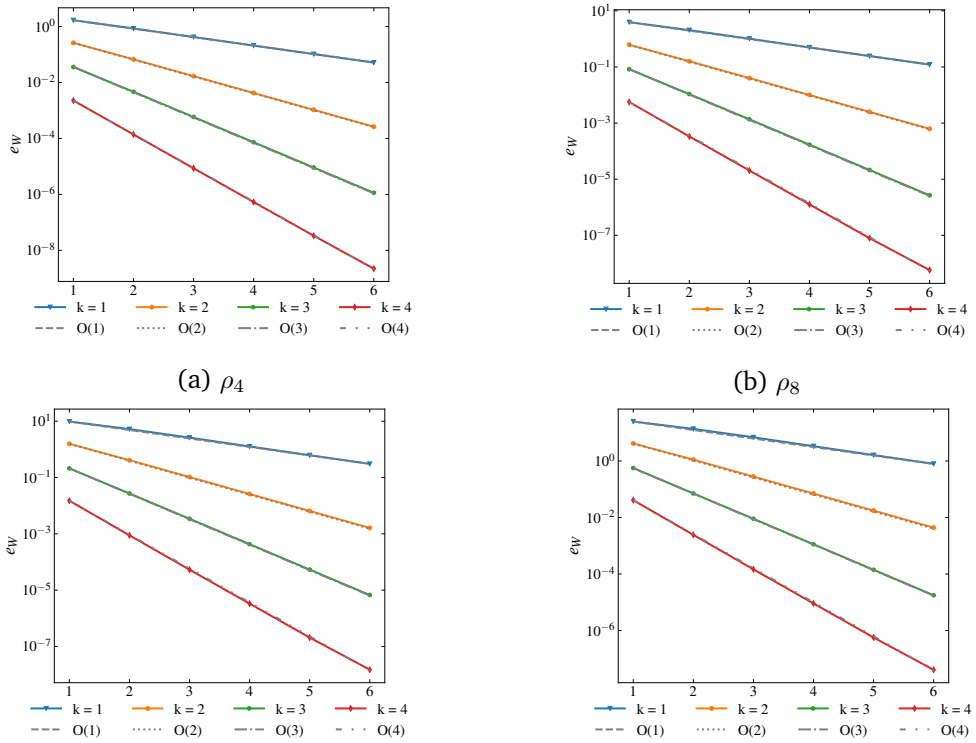


(a) $\rho = 1$ (b) $\rho = 100$ (c) $\rho = 10000$

Figure C.4: Numerical errors for u_2 with $\rho \in \{1, 100, 10000\}$ in the W-norm.

C.4 Convergence plots for a non-constant density

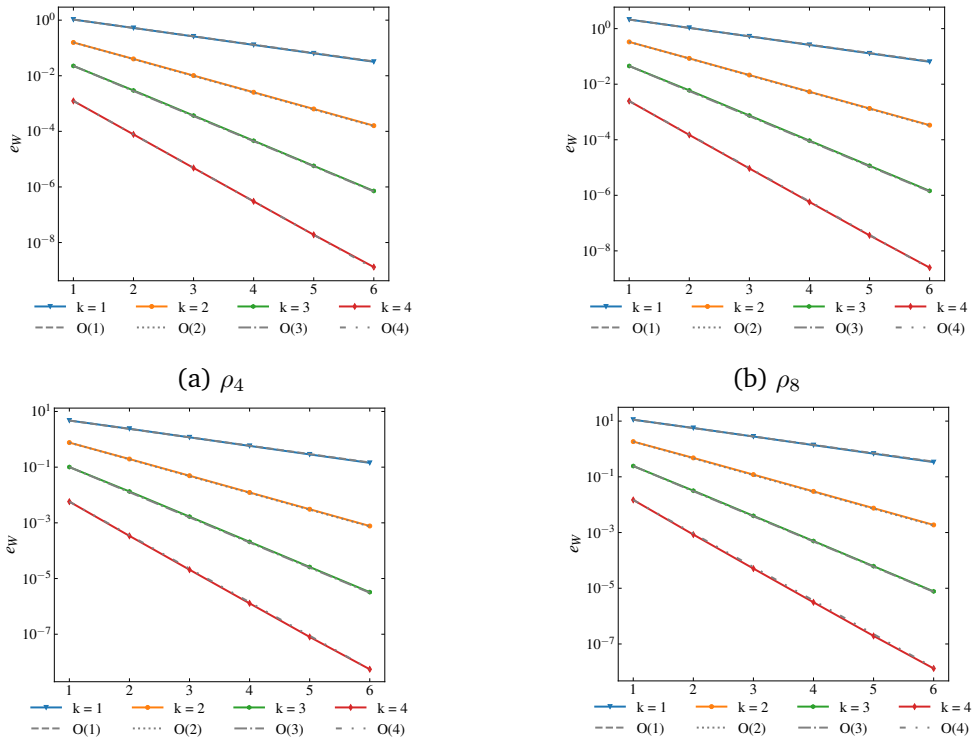
The results in the W-norm for the convergence studies in section 3.6.1 are quite similar. Hence, we only showed the convergence plot for u_2 in figure 3.25. The plots for u_1 and u_3 are displayed below.



(a) ρ_4 (b) ρ_8

(c) ρ_{12} (d) ρ_{16}

Figure C.5: Numerical errors in the W-norm for u_1 with and ρ_n,$n \in \{4, 8, 12, 16\}$.



(a) ρ_4 (b) ρ_8

(c) ρ_{12} (d) ρ_{16}

Figure C.6: Numerical errors in the W-norm for u_3 with and ρ_n,$n \in \{4, 8, 12, 16\}$.

C.5 Convergence tables for non-constant ρ

In section 3.6.1 we further considered the performance of the method for a non-constant density ρ. To describe the results, we used the errors for the velocity field u_2. Here, the errors for the remaining velocity fields u_1 and u_3 are displayed.

	ρ_8		ρ_{16}	
refs	e_{L^2}	(eoc)	e_{L^2}	(eoc)
$k=1$				
1	$1.28 \cdot 10^{-1}$		$6.35 \cdot 10^{-1}$	
2	$6.24 \cdot 10^{-2}$	(1.03)	$3.16 \cdot 10^{-1}$	(1.01)
3	$2.22 \cdot 10^{-2}$	(1.49)	$1.24 \cdot 10^{-1}$	(1.35)
4	$6.61 \cdot 10^{-3}$	(1.75)	$3.48 \cdot 10^{-2}$	(1.84)
5	$1.84 \cdot 10^{-3}$	(1.84)	$8.80 \cdot 10^{-3}$	(1.98)
6	$5.12 \cdot 10^{-4}$	(1.85)	$2.21 \cdot 10^{-3}$	(1.99)
$k=2$				
1	$9.91 \cdot 10^{-3}$		$1.74 \cdot 10^{-2}$	
2	$1.25 \cdot 10^{-3}$	(2.99)	$3.69 \cdot 10^{-3}$	(2.24)
3	$1.41 \cdot 10^{-4}$	(3.15)	$3.85 \cdot 10^{-4}$	(3.26)
4	$1.46 \cdot 10^{-5}$	(3.27)	$3.03 \cdot 10^{-5}$	(3.66)
5	$1.66 \cdot 10^{-6}$	(3.13)	$2.52 \cdot 10^{-6}$	(3.59)
6	$2.22 \cdot 10^{-7}$	(2.91)	$2.62 \cdot 10^{-7}$	(3.26)
$k=3$				
1	$1.08 \cdot 10^{-3}$		$3.11 \cdot 10^{-3}$	
2	$5.25 \cdot 10^{-5}$	(4.36)	$1.14 \cdot 10^{-4}$	(4.77)
3	$3.12 \cdot 10^{-6}$	(4.07)	$4.70 \cdot 10^{-6}$	(4.6)
4	$1.76 \cdot 10^{-7}$	(4.15)	$2.39 \cdot 10^{-7}$	(4.3)
5	$9.92 \cdot 10^{-9}$	(4.15)	$1.18 \cdot 10^{-8}$	(4.34)
6	$6.37 \cdot 10^{-10}$	(3.96)	$7.10 \cdot 10^{-10}$	(4.05)
$k=4$				
1	$1.61 \cdot 10^{-4}$		$4.58 \cdot 10^{-4}$	
2	$4.86 \cdot 10^{-6}$	(5.05)	$8.28 \cdot 10^{-6}$	(5.79)
3	$1.06 \cdot 10^{-7}$	(5.52)	$1.35 \cdot 10^{-7}$	(5.94)
4	$3.27 \cdot 10^{-9}$	(5.02)	$3.74 \cdot 10^{-9}$	(5.17)
5	$1.02 \cdot 10^{-10}$	(5.0)	$1.19 \cdot 10^{-10}$	(4.98)
6	$1.59 \cdot 10^{-10}$	(-0.65)	$1.82 \cdot 10^{-10}$	(-0.62)

Table C.6: L^2-convergence table for u_1 with non-constant ρ_n, $n \in \{8, 16\}$.

	ρ_8		ρ_{16}	
refs	e_{L^2}	(eoc)	e_{L^2}	(eoc)
$k=1$				
1	$1.22 \cdot 10^{-1}$		$3.04 \cdot 10^{-1}$	
2	$3.70 \cdot 10^{-2}$	(1.72)	$1.04 \cdot 10^{-1}$	(1.55)
3	$1.03 \cdot 10^{-2}$	(1.85)	$3.84 \cdot 10^{-2}$	(1.44)
4	$3.00 \cdot 10^{-3}$	(1.78)	$1.44 \cdot 10^{-2}$	(1.42)
5	$8.57 \cdot 10^{-4}$	(1.81)	$4.56 \cdot 10^{-3}$	(1.66)
6	$2.31 \cdot 10^{-4}$	(1.89)	$1.26 \cdot 10^{-3}$	(1.85)
$k=2$				
1	$6.56 \cdot 10^{-3}$		$1.34 \cdot 10^{-2}$	
2	$8.28 \cdot 10^{-4}$	(2.99)	$1.51 \cdot 10^{-3}$	(3.15)
3	$1.11 \cdot 10^{-4}$	(2.9)	$1.33 \cdot 10^{-4}$	(3.51)
4	$1.43 \cdot 10^{-5}$	(2.95)	$1.58 \cdot 10^{-5}$	(3.07)
5	$1.70 \cdot 10^{-6}$	(3.07)	$1.80 \cdot 10^{-6}$	(3.13)
6	$2.12 \cdot 10^{-7}$	(3.01)	$2.32 \cdot 10^{-7}$	(2.96)
$k=3$				
1	$1.21 \cdot 10^{-3}$		$2.18 \cdot 10^{-3}$	
2	$1.09 \cdot 10^{-4}$	(3.48)	$1.72 \cdot 10^{-4}$	(3.66)
3	$7.32 \cdot 10^{-6}$	(3.89)	$8.70 \cdot 10^{-6}$	(4.3)
4	$3.28 \cdot 10^{-7}$	(4.48)	$3.46 \cdot 10^{-7}$	(4.65)
5	$1.91 \cdot 10^{-8}$	(4.1)	$2.01 \cdot 10^{-8}$	(4.11)
6	$1.12 \cdot 10^{-9}$	(4.09)	$1.16 \cdot 10^{-9}$	(4.11)
$k=4$				
1	$8.39 \cdot 10^{-5}$		$2.11 \cdot 10^{-4}$	
2	$7.75 \cdot 10^{-6}$	(3.44)	$1.64 \cdot 10^{-5}$	(3.68)
3	$3.61 \cdot 10^{-7}$	(4.42)	$5.38 \cdot 10^{-7}$	(4.93)
4	$1.01 \cdot 10^{-8}$	(5.17)	$1.18 \cdot 10^{-8}$	(5.52)
5	$1.89 \cdot 10^{-10}$	(5.73)	$2.11 \cdot 10^{-10}$	(5.8)
6	$8.56 \cdot 10^{-11}$	(1.14)	$8.29 \cdot 10^{-11}$	(1.35)

Table C.7: L^2-converge table for u_3 with non-constant ρ_n, $n \in \{8, 16\}$.

	ρ_8		ρ_{16}	
refs	e_W	(eoc)	e_W	(eoc)
$k=1$				
1	$3.94 \cdot 10^{0}$		$2.50 \cdot 10^{1}$	
2	$2.06 \cdot 10^{0}$	(0.94)	$1.37 \cdot 10^{1}$	(0.87)
3	$1.02 \cdot 10^{0}$	(1.02)	$6.90 \cdot 10^{0}$	(0.99)
4	$4.98 \cdot 10^{-1}$	(1.03)	$3.34 \cdot 10^{0}$	(1.05)
5	$2.45 \cdot 10^{-1}$	(1.02)	$1.62 \cdot 10^{0}$	(1.04)
6	$1.21 \cdot 10^{-1}$	(1.01)	$7.95 \cdot 10^{-1}$	(1.03)
$k=2$				
1	$6.18 \cdot 10^{-1}$		$4.18 \cdot 10^{+}$	
2	$1.60 \cdot 10^{-1}$	(1.95)	$1.12 \cdot 10^{+}$	(1.9)
3	$4.03 \cdot 10^{-2}$	(1.99)	$2.82 \cdot 10^{-1}$	(1.99)
4	$1.01 \cdot 10^{-2}$	(2.0)	$7.05 \cdot 10^{-2}$	(2.0)
5	$2.51 \cdot 10^{-3}$	(2.0)	$1.76 \cdot 10^{-2}$	(2.0)
6	$6.27 \cdot 10^{-4}$	(2.0)	$4.40 \cdot 10^{-3}$	(2.0)
$k=3$				
1	$8.33 \cdot 10^{-2}$		$5.62 \cdot 10^{-1}$	
2	$1.08 \cdot 10^{-2}$	(2.95)	$7.11 \cdot 10^{-2}$	(2.98)
3	$1.37 \cdot 10^{-3}$	(2.98)	$9.02 \cdot 10^{-3}$	(2.98)
4	$1.71 \cdot 10^{-4}$	(3.0)	$1.13 \cdot 10^{-3}$	(3.0)
5	$2.13 \cdot 10^{-5}$	(3.0)	$1.41 \cdot 10^{-4}$	(3.0)
6	$2.67 \cdot 10^{-6}$	(3.0)	$1.77 \cdot 10^{-5}$	(3.0)
$k=4$				
1	$5.64 \cdot 10^{-3}$		$4.11 \cdot 10^{-2}$	
2	$3.35 \cdot 10^{-4}$	(4.07)	$2.44 \cdot 10^{-3}$	(4.08)
3	$2.04 \cdot 10^{-5}$	(4.04)	$1.46 \cdot 10^{-4}$	(4.06)
4	$1.27 \cdot 10^{-6}$	(4.01)	$9.09 \cdot 10^{-6}$	(4.01)
5	$7.92 \cdot 10^{-8}$	(4.0)	$5.67 \cdot 10^{-7}$	(4.0)
6	$5.77 \cdot 10^{-9}$	(3.78)	$4.05 \cdot 10^{-8}$	(3.81)

Table C.8: W-convergence table for u_1 with non-constant ρ_n, $n \in \{8, 16\}$.

	ρ_8		ρ_{16}	
refs	e_W	(eoc)	e_W	(eoc)
$k = 1$				
1	$2.12 \cdot 10^{0}$		$1.12 \cdot 10^{1}$	
2	$1.07 \cdot 10^{0}$	(0.99)	$5.68 \cdot 10^{0}$	(0.98)
3	$5.26 \cdot 10^{-1}$	(1.02)	$2.80 \cdot 10^{0}$	(1.02)
4	$2.60 \cdot 10^{-1}$	(1.02)	$1.37 \cdot 10^{0}$	(1.03)
5	$1.29 \cdot 10^{-1}$	(1.01)	$6.78 \cdot 10^{-1}$	(1.02)
6	$6.39 \cdot 10^{-2}$	(1.01)	$3.37 \cdot 10^{-1}$	(1.01)
$k = 2$				
1	$3.33 \cdot 10^{-1}$		$1.84 \cdot 10^{0}$	
2	$8.49 \cdot 10^{-2}$	(1.97)	$4.78 \cdot 10^{-1}$	(1.94)
3	$2.14 \cdot 10^{-2}$	(1.99)	$1.20 \cdot 10^{-1}$	(1.99)
4	$5.36 \cdot 10^{-3}$	(2.0)	$3.01 \cdot 10^{-2}$	(2.0)
5	$1.34 \cdot 10^{-3}$	(2.0)	$7.51 \cdot 10^{-3}$	(2.0)
6	$3.35 \cdot 10^{-4}$	(2.0)	$1.88 \cdot 10^{-3}$	(2.0)
$k = 3$				
1	$4.52 \cdot 10^{-2}$		$2.44 \cdot 10^{-1}$	
2	$5.99 \cdot 10^{-3}$	(2.92)	$3.18 \cdot 10^{-2}$	(2.94)
3	$7.48 \cdot 10^{-4}$	(3.0)	$3.98 \cdot 10^{-3}$	(3.0)
4	$9.30 \cdot 10^{-5}$	(3.01)	$4.96 \cdot 10^{-4}$	(3.01)
5	$1.16 \cdot 10^{-5}$	(3.0)	$6.18 \cdot 10^{-5}$	(3.0)
6	$1.45 \cdot 10^{-6}$	(3.0)	$7.72 \cdot 10^{-6}$	(3.0)
$k = 4$				
1	$2.47 \cdot 10^{-3}$		$1.48 \cdot 10^{-2}$	
2	$1.50 \cdot 10^{-4}$	(4.04)	$8.45 \cdot 10^{-4}$	(4.13)
3	$9.27 \cdot 10^{-6}$	(4.01)	$5.09 \cdot 10^{-5}$	(4.05)
4	$5.78 \cdot 10^{-7}$	(4.0)	$3.13 \cdot 10^{-6}$	(4.03)
5	$3.61 \cdot 10^{-8}$	(4.0)	$1.94 \cdot 10^{-7}$	(4.01)
6	$2.49 \cdot 10^{-9}$	(3.86)	$1.32 \cdot 10^{-8}$	(3.88)

Table C.9: W-converge table for u_3 with non-constant ρ_n, $n \in \{8, 16\}$.

References

[BR97] Fancesco Bassi and Stefano Rebay. “A high-order accurate Discontinuous Finite Element Method for the Numerical Solution of the compressible Navier–Stokes equations”. In: *Journal of Computational Physics* 131.2 (1997), pp. 267–279. ISSN: 0021-9991. DOI: https://doi.org/10.1006/jcph.1996.5572. URL: https://www.sciencedirect.com/science/article/pii/S0021999196955722 (cit. on p. 26).

[Cas02] Paul Castillo. “Performance of discontinuous Galerkin methods for elliptic PDEs”. In: *SIAM J. Sci. Comput.* (2002). URL: https://www.ima.umn.edu/sites/default/files/1764.pdf (cit. on p. 23).

[Chr03] Jørgen Christensen-Dalsgaard. *Lecture notes on stellar oscillations*. Tech. rep. Institut for Fysik og Astronomi, Aarhus Universitet, Denmark, 2003. URL: http://w.astro.berkeley.edu/~eliot/Astro202/2009_Dalsgaard.pdf (cit. on p. 3).

[Eva10] Lawrence C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 2010. ISBN: 9780821849743 (cit. on pp. 4, 9, 18).

[GBS10] Laurent Gizon, Aaron C. Birch, and Henk C. Spruit. “Local Helioseismology: Three-Dimensional Imaging of the Solar Interior”. In: *Annual Review of Astronomy and Astrophysics* 48.1 (2010), pp. 289–338. DOI: 10.1146/annurev-astro-082708-101722 (cit. on p. 2).

[Hal19] Martin Halla. *Galerkin approximation of holomorphic eigenvalue problems: weak T-coercivity and T-compatibility*. 2019. arXiv: 1908.05029 [math.NA] (cit. on p. 4).

[HH21] Martin Halla and Thorsten Hohage. *On the Well-posedness of the Damped Time-harmonic Galbrun Equation and the Equations of Stellar Oscillations*. 2021. DOI: 10.1137/20M1348558. eprint: https://doi.org/10.1137/20M1348558. URL: https://doi.org/10.1137/20M1348558 (cit. on pp. 3, 4, 8).

[Leh10] Christoph Lehrenfeld. “Hybrid Discontinuous Galerkin methods for solving incompressible flow problems”. PhD thesis. May 2010 (cit. on p. 54).

[Leh21] Christoph Lehrenfeld. *Numerics of partial differential equations I-IV, Lecture Notes*. 2019-2021 (cit. on pp. 23, 55).

[LO67] Donald Lynden-Bell and Jeremiah P. Ostriker. *On the stability of differentially rotating bodies*. 1967 (cit. on p. 3).

[PE12] Daniele A. Di Petro and Alexandre Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Springer-Verlag Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-22980-0 (cit. on pp. 12, 19, 26).

[WH03] Tim Warburton and Jan Hesthaven. “On the constants hp-finite element trace inverse inequalities”. In: *Computer Methods in Applied Mechanics and Engineering* 192 (June 2003), pp. 2765–2773. DOI: 10.1016/S0045-7825(03)00294-9 (cit. on p. 23).