Computer-assisted Existence Proofs for Navier-Stokes Equations on an Unbounded Strip with Obstacle

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Jonathan Matthias Wunderlich

Tag der mündlichen Prüfung: 22.02.2022

Referent:	Prof. Dr. Michael Plum
Korreferent:	Prof. Dr. Christian Wieners

Preface and Acknowledgment

Before going into the topic, I would like to seize the chance to express my gratitude to many people who supported me at several stages of this thesis. Without the support of so many people, writing this thesis would not be possible.

First of all, I would like to thank my supervisor Prof. Dr. Michael Plum who inspired me for the field of computer-assisted proofs and therefore, laid the foundation for the success of this thesis. At any time he had an open ear and with his patience and helpful advice in several discussions many issues could be solved successfully.

I am also very grateful to Prof. Dr. Christian Wieners who agreed to work as my co-referee and supported me in several numerical issues. Moreover, I am very thankful for giving me the opportunity to use the cluster of his working group to execute the calculations for my results.

Besides both of my referees I would like to thank Dr. Kaori Nagato-Plum and Dr. Elena Queirolo for introducing me to the field of fluid mechanics and suggesting the topic of this thesis. Additionally, I am grateful to Prof. Dr. Mitsuhiro T. Nakao, Prof. Dr. Yoshitaka Watanabe and Prof. Dr. Xuefeng Liu for fruitful discussions, especially at the BIRS workshop “Rigorous Numerics for Infinite Dimensional Nonlinear Dynamics” where the topic of this thesis first was discussed.

Furthermore, I would like to express my gratitude to Dr. Gabriele Brüll, Dr. Janina Gärtner, Niklas Baumgarten, Kevin Drescher, Julia Henninger and Sebastian Ohrem for carefully proofreading parts of this thesis.

Moreover, I would like to thank all members of my working group for various pleasant lunch and coffee breaks. I very much appreciate the hospitality and pleasant atmosphere in our group. Especially, the discussions with Dr. Gerd Herzog and Dr. Christoph Schmoeger have always given me a lot of pleasure. Special thanks goes to Marion Ewald who always had an open ear for mathematical, organizational as well as private issues over the last years and therefore contributed an immense part in creating a warm and pleasant atmosphere each day in office.

Concerning programming, I am very grateful to Dr. Johannes Ernesti who took a lot of time to explain the structure and usage of the finite element software M++. Moreover, I would like to thank Niklas Baumgarten, Daniele Corallo and Jonathan Fröhlich for various hours of coding, bug fixing as well as developing and organizing new ideas. I really enjoyed being part of this team.

Last but not least, I am very grateful to my parents and friends for their invaluable support. Without their encouragement and confidence over the last years finishing this thesis would not have been possible.

Contents

1 Introduction

Over the last decades, fluid mechanics became a crucial part of modern research and development in several fields. Modern transport systems like trucks, trains, ships, and air planes, but also cars are not produced without simulating their fluid mechanical behavior in advance. Moreover, skyscrapers became larger and larger over the last years which would not have been possible without analyzing their fluid mechanical behavior to ensure stability and safety for millions of people. Certainly, there are many more applications of fluid mechanics, but we want to close this listing with one final application which is certainly indispensable to everyday life. Modern water supply and sewerage systems would not be so efficient and resilient without fluid mechanical analysis. Thus, today's engineers heavily exploit the achievements of fluid mechanics to increase comfort of millions of people and safety of dozens of products in our daily life.

Even though the studies of fluid mechanics have developed over the last century, the beginnings actually go back to the time before Christ. Already Archimedes (287-212 B.C.) studied the forces acting on a body (partially or fully) immersed in a fluid and claimed that "the weight of the fluid displaced by the body is equal to the upward buoyant force" (see [79, p. 22]). Up to the present day, Archimedes' principle is still one of the central statements in fluid mechanics. Furthermore, Galileo Galilei (1564-1642), who was inspired by publications of Leonardo da Vinci (1452-1519), mainly worked on experimental fluid mechanics and developed several drafts and prototypes for machines to treat fluid mechanical problems.

Navier-Stokes Equations

Nevertheless, a theoretical analysis of the motion of a fluid first became possible after Isaac Newton (1642-1727) stated his laws of motion, which later inspired Leonhard Euler (1707-1783) to develop a mathematical model which describes the flow of perfect fluids (without friction) by the balance of momentum equation. Later, on the basis of Euler's equation, Louis Marie Henri Navier (1785-1836) developed a version which also takes friction effects (described by the dynamic viscosity constant η and Lamé's first parameter λ) into account. Then, the balance of momentum equation reads as

$$\rho\left(\frac{\partial v}{\partial t}+(v\cdot\nabla)v\right)+\nabla\bar{q}=\eta\Delta v+(\lambda+\eta)\,\nabla\mathrm{div}\,v+\bar{f}, \tag{1.1}$$

where v denotes the velocity field associated to the flow, ρ describes the density of the fluid, $\bar{q}$ is the pressure and $\bar{f}$ models external forces acting on the fluid. For a detailed derivation of the balance of momentum and more information about the physical background we refer the reader to [19, Chapter 1] or [24, Section 5.3].

Moreover, in the 19^{th} century Georg Gabriel Stokes (1816-1903) established analytic techniques for fluid mechanics of viscous flows. Especially, he developed Stokes' law describing the force of viscosity on a small sphere moving through a viscous fluid (see [97]).

The combination of the balance of momentum on the one hand together with the conservation of mass on the other hand yields the Navier-Stokes equations. Using the Divergence Theorem and two different representations of the rate of change of mass (on arbitrary fixed subregions) the continuity equation (which is the law of conservation of mass in fluid mechanics) can be derived on a general level (see for instance [19, Section 1.1]). Hence, in the field of fluid mechanics the continuity equation is given by

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho v) = 0. \tag{1.2}$$

Here, the historical overview of key developments in fluid mechanics is by no means complete. A more detailed overview can be found for instance in [24, Chapter 1] and the references therein.

Incompressible Navier-Stokes Equations

In the following, we consider incompressible flows which is characterized by the equation $\frac{\partial \rho}{\partial t} + v \cdot \nabla \rho = 0$. This condition together with the general continuity equation (1.2) implies that the continuity equation in the case of the incompressible Navier-Stokes equations now reads as $\operatorname{div} v = 0$ (note that the density function ρ is positive). Thus, the balance of momentum equation (1.1) reduces to

$$\rho \left(\frac{\partial v}{\partial t} + (v \cdot \nabla) v \right) + \nabla \bar{q} = \eta \Delta v + \bar{f}. \tag{1.3}$$

Both equations combined with appropriate boundary and/or initial conditions are called (time-dependent) incompressible Navier-Stokes equations.

Additionally, in the further course we make the assumption that the density ρ is given by a positive constant (which fits into the case of incompressible flows, i.e., $\frac{\partial \rho}{\partial t} + v \cdot \nabla \rho = 0$). In this setting the balance of momentum equation (1.3) can be transformed into

$$\frac{\partial v}{\partial t} + (v \cdot \nabla) v + \nabla q = \nu \Delta v + f, \tag{1.4}$$

with $q = \frac{\bar{q}}{\rho}$, $f = \frac{\bar{f}}{\rho}$, and $\nu = \frac{\eta}{\rho}$ denoting the kinematic viscosity. We want to emphasize that in the literature also this balance of momentum equation (1.4) together with $\operatorname{div} v = 0$ and appropriate boundary and/or initial conditions are called incompressible Navier-Stokes equations. Since we assumed the density ρ to be constant throughout this thesis, in the further course this version will be referred as the (time-dependent) incompressible Navier-Stokes equations.

In several works the (time-dependent) incompressible Navier-Stokes equations are considered on the entire space $\mathbb{R}^n \times (0, \infty)$ with initial condition $v(\cdot, 0) = v_0$ and the question of global and local well-posedness depending on the initial datum v_0 is discussed. In physical examples usually the dimension n takes values 2 or 3, whereas in purely analytical settings higher dimensions can be considered as well.

If the initial datum is smooth enough, the Navier-Stokes equations are locally well-posed if suitable "boundary" conditions on the velocity field and the pressure at ∞ are introduced. For instance in [51] the existence of a local unique solution for suitable initial datum was proven by Kato and Ponce. In [54] the authors claim that for initial datum $v_0 \in H^s(\mathbb{R}^n)$ with $s > \frac{n}{2}$ there exists a unique local solution $v \in C([0,t], H^s(\mathbb{R}^n))$ with associated pressure $q \in C([0,t], H^s(\mathbb{R}^n))$. Moreover, in [48] Kato answered the question of global well-posedness for small initial data and local well-posedness for large data if $v_0 \in L^n(\mathbb{R}^n)$. A few years later, analogue results have been proved by Giga and Miyakawa [35], Taylor [99], and Kato [49] if the initial datum is an element of a suitable Morrey space. Moreover, Cannone [16] and Planchon [83] obtained similar results if the initial datum lies in certain Besov spaces. Later, Koch and Tataru proved in [54] the existence of a global solution of the Navier-Stokes equations if the initial datum is "small" in a suitable subset of tempered distributions.

In [100] Temam considered the time-dependent Navier-Stokes equations and proved existence, uniqueness and regularity results on a bounded Lipschitz domain. Moreover, some Remarks concerning unbounded domains are given by Temam. In [32] Galdi, Maremonti, and Zhou proved existence and uniqueness of a regular solution of the Navier-Stokes initial-boundary value problem on a smooth exterior domain of $\mathbb{R}^n$ (with $n \geq 3$) in space and $(0,T)$ in time if the initial datum is bounded and non-decaying. Moreover, they give a sufficient condition on the spatial growth of ∇q which provides boundedness of the solution v for all times.

The stability of the Orr-Sommerfeld equation with plane Poiseuille flow was investigated by Watanabe, Plum and Nakao in [106] using computer-assisted means. The authors proved an instability result in the case of a two-dimensional flow between two infinite plates. In [104], Watanabe, Nagatou, Plum and Nakao presented an instability result for the Orr-Sommerfeld problem with Poiseuille flow for some interval domain. Moreover, in [57] Lahmann and Plum considered the Orr-Sommerfeld equation with Blasius profile on the unbounded domain $[0,\infty)$ and proved an instability result in this setting.

Furthermore, the Navier-Stokes equations on $\mathbb{T}^3 \times \mathbb{R}$ with periodic boundary conditions have been investigated by Bruckmaster and Vicol in [15], where a non-uniqueness result for weak solutions with bounded energy has been proved. We note that the work by Bruckmaster and Vicol was inspired by earlier results of Leray [59] and Hopf [44] dealing with this setting as well.

Concerning the pressure, in [95] Sohr and van Wahl showed some regularity results for the pressure on bounded and exterior domains in $\mathbb{R}^n$ with $n \geq 3$.

Moreover, we want to mention results by Galdi and Silvestre on the Navier-Stokes equations on a domain which is exterior to a rigid body that rotates with constant angular velocity. For instance, in [34] they proved existence of a global solution for a certain class of initial data.

In [11] Boukir, Maday, Métivet and Razafindrakoto propose a high-order-time splitting scheme for the (time-dependent) incompressible Navier-Stokes equations for bounded domains in $\mathbb{R}^2$ and $\mathbb{R}^3$. We close our overview about the (time-dependent) Navier-Stokes equations with an approximation result by Chorin. In [18], he proved the convergence of approximations to the exact solution in the case of a bounded domain in space and time. Moreover, he provided an estimate for the rate of convergence.

Stationary Navier-Stokes Equations

In the following we drop the time-dependency of the velocity field as well as for the pressure, i.e., we will only consider time-independent solutions. Hence, the balance of momentum (1.4) results in the stationary (i.e., time-independent) version

$$-\nu\Delta v + (v\cdot\nabla)v + \nabla q = f.$$

Together with the continuity equation $\operatorname{div} v = 0$ (which remains the same in the time-independent case) we obtain the stationary Navier-Stokes equations.

Working on bounded domains, certain boundary conditions need to be implemented. In the literature several different boundary conditions are considered. For instance in [66], Mucha investigated the stationary Navier-Stokes equations with slip boundary conditions on an infinite pipe (i.e., $v\cdot\hat{\nu} = 0$ on the boundary, where $\hat{\nu}$ denotes the outer normal) which can be perturbed by a compact obstacle. In the work by Mucha, the velocity is supposed to be driven by a constant background flow at infinity which is possible due to the slip boundary conditions.

Watanabe, in [102] and [103], considered the Navier-Stokes equations on a two-dimensional flat torus in view of Kolmogorov flows. In his papers he presented a computer-assisted proof for the steady-state solutions for a given Reynolds number and a prescribed aspect ratio which somehow characterizes the torus under investigation. Moreover, in [67], Nagatou presented a computer-assisted approach to prove the stability of the Kolmogorov flow on a two-dimensional flat torus. In all these results a stream function formulation of the problem is used.

One of the most popular stationary problem considered by several authors over the last decades is Leray's problem

$$\begin{aligned} \left.\begin{aligned} -\nu\Delta v + (v\cdot\nabla)v + \nabla q &= f \\ \operatorname{div} v &= 0 \end{aligned}\right\} &\text{ in } \Omega \\ v = v_0 \quad &\text{ on } \partial\Omega \end{aligned} \tag{1.5}$$

which appears either with inhomogeneous or homogeneous (i.e., $v_0 = 0$) Dirichlet boundary conditions.

Before stating famous results for Leray's problem we would like to mention two simplified versions by Stokes and Oseen which in both cases result in a linear problem. First, Stokes in [97] considered the equations without convection term, i.e., the equation of momentum reads as $-\nu\Delta v + \nabla q = f$. Moreover, Temam in [100] proved existence and uniqueness results of the Stokes equations on bounded domains as well as on certain unbounded domains using a corresponding weak formulation of the problem. A detailed investigation of Stokes' problem for bounded as well as for unbounded domains can be found for instance in [31, Chapters IV-VI]. Additionally, Nakao, Yamamoto and Watanabe in [77] considered Stokes' problem on a two-dimensional, convex, polygonal domain and presented a method to obtain (constructive) a priori error bounds for the Stokes problem.

Moreover, for exterior domains in [81] Oseen replaced the convection term by $(v_1\cdot\nabla)v$ where v_1 is a constant velocity field such that

$$\lim_{|x|\to\infty} v(x) = v_1. \tag{1.6}$$

For further information on Oseen's problem we refer the reader again to [31, Chapters VII-VIII].

Back to Leray's problem: We want to point out that the existence of solutions heavily depends on the choice of the domain Ω. For bounded domains Ω results by several authors are known. For instance in [56] Kozono and Yanagisawa considered Leray's problem on a bounded simply connected domain with smooth boundary and inhomogeneous boundary condition. Moreover, in [55] Korobkov, Pileckas, and Russo investigated the inhomogeneous Leray's problem on a bounded domain with a C^2-boundary and proved existence of a solution under the sole necessary condition of zero total flux through the boundary.

In [28] and [27] Farwig, Galdi, and Sohr developed a larger class, the so-called class of very weak solutions, in which they proved existence and uniqueness of solutions to Leray's problem with inhomogeneous boundary condition on bounded domains in $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively.

Furthermore, in the bounded domain case some existence proofs using computer-assisted techniques exist. In [112] Wieners proved existence of a solenoidal (in the L^2-sense) solution to Leray's problem on a bounded Lipschitz domain $\Omega \subseteq \mathbb{R}^2$ if the Reynolds number is sufficiently small, i.e., since we have the relation $\nu = \frac{1}{Re}$, if the kinematic viscosity ν is sufficiently large. We note that the existence proof for the pressure is neglected in his investigation. However, since the domain is bounded the existence of an associated pressure can be obtained as described in [31, Chapter IX]. In [68] Nagatou, Hashimoto and Nakao proposed an improvement which also remains applicable for high Reynolds numbers. However, in their approach Nagatou, Hashimoto and Nakao used a stream function formulation of the problem which requires a simply connected (bounded) domain.

Moreover, in [107] Watanabe, Yamamoto and Nakao considered the stationary Navier-Stokes equations on a bounded, convex polygonal domain in $\mathbb{R}^2$ and proved the existence of a solution using computer-assisted techniques. In contrast to that, the stationary Navier-Stokes equations on a bounded, nonconvex polygonal domain in $\mathbb{R}^2$ was investigated by Nakao, Hashimoto and Kobayashi in [72]. Especially, a bounded L-shaped domain, which is a mathematical model for step flow problems, is considered. Using Nakao's method (cf. [69], [71], [73]) the authors proved the existence and local uniqueness of a solution of the stationary Navier-Stokes equations. Additionally, in [105] Watanabe, Nakao and Nagatou proved a compactness result for a non-linear operator related to a stream function-vorticity formulation of the Navier-Stokes equations for two-dimensional rectangular domains (which by a remark of the authors can be extended to convex polygonal domains). The compactness of such an operator for instance is required for a computer-assisted proof based on Schauder's Fixed-point theorem.

Whereas for bounded domains there are quite a lot of results concerning existence and uniqueness of solutions to Leray's problem, in case of unbounded domains several difficulties come into play. The case of exterior domains was investigated by several authors. For instance, in [33] Galdi and Rabier considered the stationary Navier-Stokes equations with inhomogeneous boundary conditions on an exterior domain in $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively, with a non-zero velocity v_1 at infinity, i.e., (1.6) holds with a non-trivial and constant velocity v_1. In this situation they proved existence of a solution if the boundary data is in a suitable Sobolev space and the forcing term f lies in an appropriate Lebesgue space.

Later, in the case of a smooth boundary and $v_1 = 0$, i.e., the velocity equals zero at infinity, Kim and Kozono in [52] proved existence of a weak solution to Leray's problem provided the forcing term is sufficiently small in a suitable Lebesgue space. Moreover, the authors obtained a uniqueness result in the class of solutions satisfying the energy inequality. In [91] Russo dropped the constraint at infinity completely and showed that there exists a solution to Leray's problem if the Reynolds number is sufficiently small.

Moreover, in [115] Wittwer considered the stationary Navier-Stokes equations on a half-space domain with a background flow at "infinity" and appropriate boundary conditions. In [40] Hillairet and Wittwer considered a half-space domain but now perturbed by a compact smooth obstacle and proved the existence of solutions for Dirichlet boundary conditions at the boundary of the half-space and suitable boundary conditions at infinity and the obstacle. Additionally, Hillairet and Wittwer in [41] proved the existence of solutions to the stationary Navier-Stokes equations on a two-dimensional exterior domain of a disc with non-zero Dirichlet boundary conditions on the boundary of the disc and zero boundary conditions at infinity.

Finally, for more general domains with unbounded boundary some results are also known. However, the proof of existence becomes a more challenging task. For instance in [4] and [5] Amick considered simply connected cylindrical domains with smooth boundary and proved existence and regularity results provided the Reynolds number is "small".

Symmetrical channels in $\mathbb{R}^2$ are investigated also by Fraenkel and Eagles in [29] and [30], where they also proved the existence of solutions under certain constraints on the Reynolds number. Moreover, in [39] Heywood considered flows through certain apertures and ducts in $\mathbb{R}^3$ that widen strongly at infinity.

For a more detailed overview about the results for general domains with unbounded boundary we refer the reader to [6] and especially to [31, Chapter XIII], where Galdi proved an existence result in a very general framework for the domain if the Reynolds number is below some critical (and in general not explicitly known) value provided the solution has a fixed flux through a suitable intersection of the domain (cf. [31, Theorem XIII.3.2]).

At the end of this overview, we would like to mention some numerical results: For instance, in [98] Taylor and Hood proposed (and compared) two methods to obtain approximate solutions of the Navier-Stokes equations via finite element methods. Their first method uses the velocity and pressure as variables whereas the second one is based on the stream function formulation, i.e., the stream function and the vorticity are used as unknowns. Moreover, in [101] Tobiska and Verfürth investigated a streamline diffusion finite element method on bounded, polyhedral domains in dimension 2 and 3.

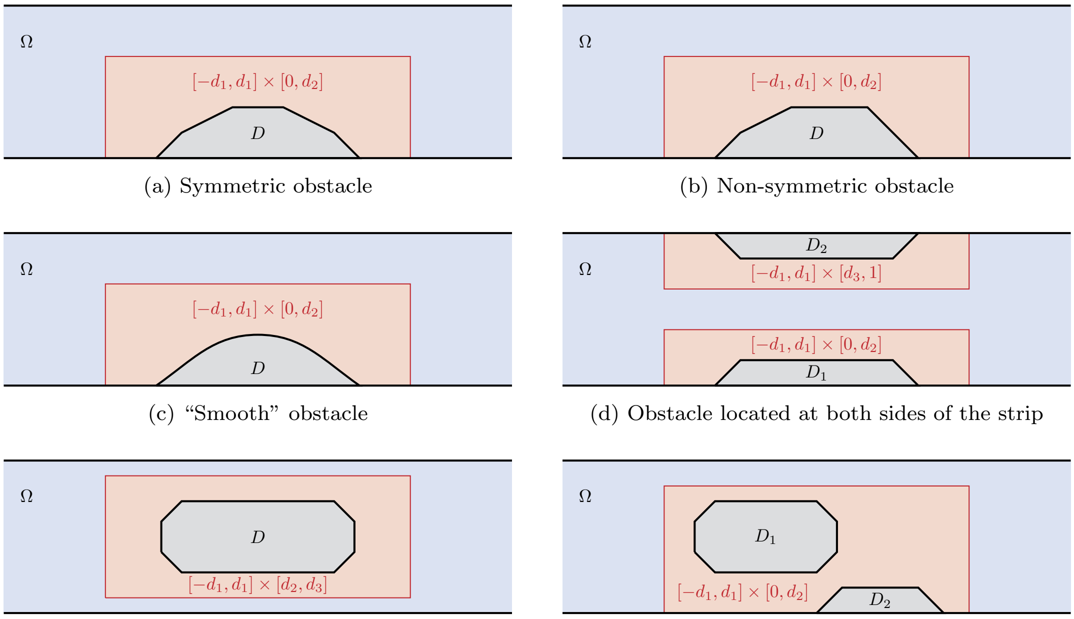
Aim of this Thesis

We would like to point out that all previous results are based on a certain smallness assumption on the Reynolds number. To the best of our knowledge there are neither purely analytical nor computer-assisted existence results for arbitrarily large Reynolds numbers provided the flux through a suitable intersection of the domain remains the same (cf. [31, Remark XIII.3.4]). Since to this day purely analytical proofs failed to answer the question if solutions to Leray's problem exist also for large Reynolds numbers, in this thesis we establish an abstract setting to apply computer-assisted techniques to the Navier-Stokes equations which (at least) theoretically results in an answer to this question. It is clear that these techniques cannot cover the whole range of possible Reynolds numbers since we have to fix concrete values in our proofs. However, if our computer-assisted existence theorem (cf. Theorem 3.4) provides the existence of a solution for large Reynolds numbers this would be a first step towards a statement for large Reynolds numbers, i.e., in the affirmative case one might expect that there exist solutions to the Navier-Stokes equations for larger Reynolds numbers as well. We note that the restriction to single Reynolds numbers in Theorem 3.4 can be weakened using interval arithmetic, i.e., our methods can also provide the existence of a solution for all Reynolds numbers lying in a compact interval (cf. Section 8.2).

Assumptions on the Domain Ω

In the further course we restrict our considerations and examples to domains in $\mathbb{R}^2$. However, we want to emphasize that the analytical setting directly applies to higher dimensions, where at several stages some adaptions are required. In Section 8.5 we give some remarks on the higher dimensional case.

In the following we finally fix the domain Ω as the infinite strip $S := \mathbb{R} \times (0,1)$ perturbed by a compact obstacle $D \subseteq \overline{S}$, i.e, we set $\Omega := S \setminus D$. Additionally, we assume that the obstacle is chosen such that the (unbounded) boundary of Ω is Lipschitz. We distinguish between two different types of obstacle. On the one hand, we investigate obstacles located at the boundary of the strip and, on the other hand, we consider the case where the obstacle is detached from the boundary. Thus, in the further course, we suppose that there exist constants $d_1, d_2, d_3 > 0$ with $d_2 < d_3 < 1$ such that either $D \subseteq [-d_1, d_1] \times ([0, d_2] \cup [d_3, 1])$ (which describes the case where the obstacle is located at the boundary of the strip) or $D \subseteq [-d_1, d_1] \times [d_2, d_3]$ (in the case with obstacles detached from the boundary) holds true. For some example domains see Figure 1.1.


Ω
$[-d_1, d_1] \times [0, d_2]$
D
Ω
$[-d_1, d_1] \times [0, d_2]$
D
Ω
$[-d_1, d_1] \times [0, d_2]$
D
Ω
D_2
$[-d_1, d_1] \times [d_3, 1]$
$[-d_1, d_1] \times [0, d_2]$
D_1
Ω
D
$[-d_1, d_1] \times [d_2, d_3]$
Ω
D_1
$[-d_1, d_1] \times [0, d_2]$
D_2

(a) Symmetric obstacle

(b) Non-symmetric obstacle

(c) “Smooth” obstacle

(d) Obstacle located at both sides of the strip

(e) Obstacle detached from strip boundary

(f) Obstacle located on one side of the strip

Figure 1.1: Example domains with different obstacles

At this stage, we want to emphasize that the theory developed in this thesis also allows obstacles which are detached from the boundary of S, i.e., obstacles which are completely flowed around by the fluid, which results in a domain Ω that is not simply connected (cf. Figure 1.1 (e)). Therefore, in all our approaches we avoid the use of stream functions since for the existence of a stream function associated to the velocity field a simply connected domain is required.

On our domain Ω we consider the stationary incompressible Navier-Stokes equations (1.5) with no-slip boundary conditions, i.e., homogeneous Dirichlet boundary conditions ($v_0 = 0$). Using the definition of the Reynolds number $Re = \frac{1}{\nu}$, we obtain the equivalent formulation of the Navier-Stokes equations

$$\begin{aligned} \left.\begin{aligned} -\Delta v + Re\,[(v\cdot\nabla)v + \nabla q] &= f \\ \operatorname{div} v &= 0 \end{aligned}\right\} &\text{ in } \Omega, \\ v = 0 \quad &\text{ on } \partial\Omega. \end{aligned} \tag{1.7}$$

We note that on the right-hand side of the first equation the factor $\frac{1}{\nu}$ is absorbed into the external force f.

Remark 1.1. (i) *If for the first problem type the obstacle is only located at one side of the boundary (without loss of generality we assume that the obstacle is located at the bottom of the strip), i.e., $D \subseteq [-d_1, d_1] \times [0, d_2]$ is satisfied, the constant d_3 is actually not needed for the description of the obstacle. In this setting we (formally) set $d_3 = 1$ in the further course which in particular comes into play in our numerical algorithms.*

(ii) *If the obstacle is located at opposite sides of the boundary (cf. Figure 1.1 (d)) we have to assume that there is "some space" between the two parts which can be measured by the difference $d_3 - d_2$. In practice one can expect that for "small" differences $d_3 - d_2$ our computer-assisted proof becomes more challenging or even fails.*

(iii) *We note that in our applications it turned out that it makes sense to choose the constants d_2 and d_3 such that $\operatorname{dist}([-d_1, d_1] \times \{d_2, d_3\}, D)$ is "small", i.e., the set containing the obstacle should be chosen minimal in some sense to obtain optimal results in the further course (cf. part* (ii)*). The vertical parts are not that critical in that sense. However, we are interested in a "small" computational domain which suggests to choose d_1 not "too" large as well (cf. Section 4.1).*

(iv) *The theory developed in this thesis is not restricted to a connected obstacle, i.e., we can also treat obstacles which consist of two or even finitely many compact parts. Moreover, we can also consider problems where both obstacle types occur (cf. Figure 1.1 (f)) which can also be considered by the techniques presented in the further course.*

To transform the Navier-Stokes equations into a weak setting on a suitable Sobolev space, we require some background flow which somehow "models" the solution for $|x| \to \infty$. Therefore, in the following Section we introduce a "simple" solution on the strip S.

1.1 A "Simple" Solution on the Strip

For the moment, we suppose that the Navier-Stokes equations (1.7) are formulated on the whole strip S (instead of Ω) without forcing term, i.e., $f = 0$. Hence, we consider the Navier-Stokes equations

$$\left.\begin{aligned} -\Delta v + Re\,[(v\cdot\nabla)v + \nabla q] &= 0 \\ \operatorname{div} v &= 0 \end{aligned}\right\} \text{ in } S, \qquad v = 0 \quad \text{on } \partial S. \tag{1.8}$$

In this setting, we can construct a "simple" solution (U, P) which is of the form

$$U(x,y) = \begin{pmatrix} U_1(x,y) \\ 0 \end{pmatrix} \quad \text{for all } (x,y) \in S. \tag{1.9}$$

Using the divergence condition in (1.8) and the structure of U (see (1.9)) we obtain $\frac{\partial U_1}{\partial x}(x,y) = 0$ for all $(x,y) \in S$, implying $U_1(x,y) = U_1(y)$ for all $(x,y) \in S$. Inserting this result into the first equation of (1.8) yields

$$\begin{pmatrix} -\frac{\partial^2 U_1}{\partial y^2}(y) \\ 0 \end{pmatrix} + Re\left[U_1(y)\frac{\partial}{\partial x}\begin{pmatrix} U_1(y) \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{\partial P}{\partial x}(x,y) \\ \frac{\partial P}{\partial y}(x,y) \end{pmatrix}\right] = 0 \quad \text{for all } (x,y) \in S.$$

Hence, the second equation gives $P(x,y) = P(x)$ for all $(x,y) \in S$. Since $\frac{\partial U_1}{\partial x}(x,y) = 0$ for all $(x,y) \in S$, we get

$$\frac{\partial^2 U_1}{\partial y^2}(y) = Re\frac{\partial P}{\partial x}(x) = -2\alpha = \text{const} \quad \text{for all } (x,y) \in S$$

for a constant $\alpha \in \mathbb{R}$.

Solving the first ordinary differential equation and inserting the boundary conditions $U_1(0) = U_1(1) = 0$ (see (1.8)) yields $U_1(y) = \alpha y(1-y)$ for all $y \in (0,1)$. The pressure is given by $P(x) = -\frac{2\alpha}{Re}x + \beta$ for all $x \in \mathbb{R}$. Choosing $\alpha = 1$ and $\beta = 0$ yields the following solution of (1.8)

$$U(x,y) = U(y) = \begin{pmatrix} y(1-y) \\ 0 \end{pmatrix} \quad \text{and} \quad P(x,y) = P(x) = -\frac{2}{Re}x \quad \text{for all } (x,y) \in S, \tag{1.10}$$

called the Poiseuille flow and its corresponding pressure (see Figure 1.2).

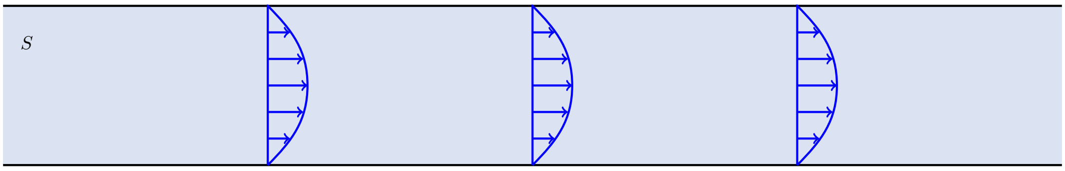


Figure 1.2: Strip with Poiseuille flow

Remark 1.2. *The choice $\alpha = 1$ is somehow arbitrary. However, in the context of the Navier-Stokes problem considered in [31, Chapter XIII] it fixes the flux of the Poiseuille flow through the intersection $\Sigma \coloneqq \{x\} \times [0,1]$ (for arbitrary $x \in \mathbb{R}$) with respect to the normal $n \coloneqq (1,0)^T$, i.e., we have*

$$\int_\Sigma U \cdot n \,\mathrm{d}\sigma = \frac{1}{6}.$$

1.2 Transforming the Equations

Since we consider the Navier-Stokes equations (1.7) on the unbounded domain Ω, we have to clarify the type of solutions we are interested in. In the following we look for velocity solutions containing the Poiseuille flow U defined in (1.10) as background flow, i.e., we consider velocities of the form $v \coloneqq U + \bar{u}$ where $\bar{u}$ decays to zero as $|x| \to \infty$. For the pressure we consider the corresponding form $q \coloneqq P + p$. Thus, inserting the ansatz for the velocity and pressure into (1.7) and using the fact that (U, P) solves (1.8) yields

$$\begin{aligned} \left.\begin{aligned} -\Delta\bar{u} + Re\,[(\bar{u}\cdot\nabla)\bar{u} + (\bar{u}\cdot\nabla)U + (U\cdot\nabla)\bar{u} + \nabla p] = f \\ \operatorname{div}\bar{u} = 0 \end{aligned}\right\} &\quad \text{in } \Omega \\ \bar{u} = 0 &\quad \text{on } \partial\Omega \setminus \partial D \\ \bar{u} = -U &\quad \text{on } \partial\Omega \cap \partial D \end{aligned} \tag{1.11}$$

Remark 1.3. *Since the Poiseuille flow is non-zero on the boundary of the obstacle, the transformed boundary condition splits into two parts, where on ∂D the condition $\bar{u} = -U$ yields $v(x,y) = U(x,y) + \bar{u}(x,y) = 0$ for all $(x,y) \in \partial D$ for the solution v of* (1.7).

To avoid the splitting of the boundary conditions (cf. (1.11)), we perform a second transformation. Therefore, we first construct a divergence-free and compactly supported function $V\colon S \to \mathbb{R}^2$ with $\operatorname{supp}(V) \subseteq [-d_0, d_0] \times [0,1]$, where $d_0 > d_1$ is a constant to be specified later on. Moreover, we demand $V = U$ on ∂D as well as $V = 0$ on $\partial\Omega\setminus\partial D$. In the classical sense one can think of a function which is sufficiently smooth, for example we could choose $V \in C^2(\Omega, \mathbb{R}^2)$. In the context of weak solutions (cf. Section 1.3), it becomes clear that it is sufficient to choose $V \in H^2(\Omega, \mathbb{R}^2)$ which will be the case in the further reading. For more details about the construction of V we refer the reader to Section 4.1.

With such a function V in hand, we introduce the second transformation $\bar{u} = u - V$ where u decays to zero as $|x| \to \infty$

Direct calculations show

$$\begin{aligned} (\bar{u}\cdot\nabla)\bar{u} &= (u\cdot\nabla)u - (u\cdot\nabla)V - (V\cdot\nabla)u + (V\cdot\nabla)V, \\ (\bar{u}\cdot\nabla)U &= (u\cdot\nabla)U - (V\cdot\nabla)U, \\ (U\cdot\nabla)\bar{u} &= (U\cdot\nabla)u - (V\cdot\nabla)V, \end{aligned}$$

and, together with

$$\Gamma \coloneqq U - V, \tag{1.12}$$

we obtain the following Navier-Stokes equations considered in this thesis:

$$\left.\begin{aligned} -\Delta u + Re\,[(u\cdot\nabla)u + (u\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)u + \nabla p] &= g \\ \operatorname{div} u &= 0 \end{aligned}\right\} \text{ in } \Omega \qquad (1.13)$$
$$u = 0 \quad \text{on } \partial\Omega$$

where the modified right-hand side is given by

$$\begin{aligned} g :=& f - \Delta V - Re\,[(V\cdot\nabla)V - (V\cdot\nabla)U - (U\cdot\nabla)V] \\ =& f - \Delta V + Re\,[(V\cdot\nabla)\Gamma + (U\cdot\nabla)V] \qquad (1.14) \\ =& f - \Delta V - Re\,(\Gamma\cdot\nabla)\Gamma. \end{aligned}$$

Remark 1.4. (i) *We note that the assumption $V(x,y) = U(x,y)$ for all $(x,y) \in \partial D$ implies $u(x,y) = \bar{u}(x,y) + V(x,y) = 0$ for all $(x,y) \in \partial D$, i.e., u satisfies the Dirichlet boundary condition on ∂D (cf.* (1.13)*).*

(ii) *In the further considerations we set the forcing term f to zero, however all calculations and proofs can be slightly modified to obtain corresponding results when f is non-zero. Moreover, we note that g defined in* (1.14) *is non-zero although f vanishes.*

1.3 Weak Formulation

Since we are interested in weak solutions of the transformed Navier-Stokes equations (1.13) the Sobolev space $H^1_0(\Omega, \mathbb{R}^2)$ is the canonical choice for the velocity. Following the lines in [92] and [94] in the case of bounded domains, as well as [31, Chapter XIII] we split our problem into two separate parts.

In the first part, we consider velocity fields lying in the subspace

$$H(\Omega) := \left\{ u \in H^1_0(\Omega, \mathbb{R}^2) \colon \operatorname{div} u = 0 \right\} \subseteq H^1_0(\Omega, \mathbb{R}^2).$$

of divergence-free functions. Thus, by our choice of the velocity space the second equation of the transformed Navier-Stokes equations (1.13) is satisfied by construction.

Moreover, $(\nabla p)[\varphi] := -\int_\Omega p \operatorname{div} \varphi \,\mathrm{d}(x,y)$ for all $p \in L^2_{loc}(\overline{\Omega})$ and $\varphi \in H^1_0(\Omega, \mathbb{R}^2)$ with compact support in $\overline{\Omega}$ leads to the weak formulation:

Find $u \in H(\Omega)$ such that

$$\int_\Omega \left(\left[\sum_{i=1}^{2} \nabla u_i \cdot \nabla \varphi_i \right] + Re\,[(u\cdot\nabla)u + (u\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)u] \cdot \varphi \right) \mathrm{d}(x,y)$$
$$= \int_\Omega g \cdot \varphi \,\mathrm{d}(x,y) \quad \text{for all } \varphi \in H(\Omega) \text{ with } \operatorname{supp}(\varphi) \subseteq \overline{\Omega} \text{ compact.}$$

From [31, Section III.4.3 (conclusions after Theorem III.4.3)] we conclude that

$$\left\{ u \in H^1_0(\Omega, \mathbb{R}^2) \colon \operatorname{supp}(u) \subseteq \overline{\Omega} \text{ is compact} \right\}$$

is dense in $H(\Omega)$ for our type of domain. Thus, we obtain the following weak formulation of our transformed Navier-Stokes equations:

Find $u \in H(\Omega)$ such that

$$\begin{aligned}\int_\Omega \left(\left[\sum_{i=1}^{2} \nabla u_i \cdot \nabla \varphi_i \right] + Re \left[(u \cdot \nabla) u + (u \cdot \nabla) \Gamma + (\Gamma \cdot \nabla) u \right] \cdot \varphi \right) \mathrm{d}(x,y) \\ = \int_\Omega g \cdot \varphi \, \mathrm{d}(x,y) \quad \text{for all } \varphi \in H(\Omega).\end{aligned} \tag{1.15}$$

After having proved the existence of an exact solution of (1.15), in the second part we reconstruct the corresponding pressure. For more details we refer the reader to Chapter 7.

1.4 Outline of this Thesis

In this thesis we establish an existence proof using computer-assisted techniques to obtain a solution $u^* \in H(\Omega)$ of (1.15). Therefore, in Chapter 3 we present an existence and enclosure theorem based on an abstract theorem by Plum (cf. [85, Theorem 1]. Moreover, in Chapters 4 - 6 we explain the crucial assumptions of our main theorem in a more detailed way.

In applications where our computer-assisted proof of a solution u^* of the Navier-Stokes equation (1.15) was successful, in a second step we prove existence of the corresponding pressure p^* such that (u^*, p^*) is a weak solution of our transformed Navier-Stokes equations (1.13) in a sense described in Chapter 7. For more details about the reconstruction procedure for the pressure we refer the reader to Chapter 7. Finally, in Chapter 8 we present the results obtained for several domains and Reynolds numbers.

At the end of this thesis, in Chapter 9, we give an overview about crucial extensions for the finite element software M++ (Meshes, Multigrid and More) which were developed to treat the problems of this thesis successfully.

2 Preliminaries and Basic Notations

Before going into details about our computer-assisted proof for the Navier-Stokes equations (1.15) we clarify some basic notations like norms and spaces needed at several stages of this thesis. Furthermore, in Section 2.4 we introduce a Fourier transform on the unbounded strip S which uses the well-known Fourier transform in x-direction and Fourier series expansion in y-direction.

2.1 Spaces and Norms

For fixed $p \in [1, \infty]$ let $L^p(\Omega, \mathbb{R}^2)$ (with Ω defined in Chapter 1) denote the Lebesgue space consisting of L^p-integrable functions with values in $\mathbb{R}^2$. On $L^p(\Omega, \mathbb{R}^2)$ we define the L^p-norm by

$$\|u\|_{L^p(\Omega,\mathbb{R}^2)} := \left(\sum_{i=1}^{2} \|u_i\|^2_{L^p(\Omega)} \right)^{\frac{1}{2}} \quad \text{for all } u \in L^p(\Omega, \mathbb{R}^2), \tag{2.1}$$

where $\| \cdot \|_{L^p(\Omega)}$ on the right-hand side denotes the usual L^p-norm on $L^p(\Omega)$, i.e., we have $\|w\|^p_{L^p(\Omega)} = \int_\Omega |w|^p \,\mathrm{d}(x,y)$ for all $w \in L^p(\Omega)$, $p \in [1, \infty)$ and $\|w\|_{L^\infty(\Omega)} = \operatorname{ess\,sup}_\Omega |w|$ for all $w \in L^\infty(\Omega)$. We note that our definition differs from that one in other works (e.g. [112]).

If $p = 2$, we can use the Euclidean norm $| \cdot |$ on $\mathbb{R}^2$ to obtain the following equivalent formulation: $\|u\|_{L^2(\Omega,\mathbb{R}^2)} = (\int_\Omega |u|^2 \,\mathrm{d}(x,y))^{\frac{1}{2}}$ for all $u \in L^2(\Omega, \mathbb{R}^2)$. Since $L^2(\Omega)$ together with the usual inner product $\langle \cdot\,, \cdot \rangle_{L^2(\Omega)}$ is a Hilbert space, $L^2(\Omega, \mathbb{R}^2)$ endowed with the inner product

$$\langle u, v \rangle_{L^2(\Omega,\mathbb{R}^2)} := \sum_{i=1}^{2} \langle u_i, v_i \rangle_{L^2(\Omega)} = \int_\Omega u \cdot v \,\mathrm{d}(x,y) \quad \text{for all } u, v \in L^2(\Omega, \mathbb{R}^2) \tag{2.2}$$

becomes a Hilbert space as well. Obviously, the inner product in (2.2) corresponds to the L^2-norm stated above in (2.1) for the case $p = 2$.

In almost the same manner, for $p \in [1, \infty]$ we endow the Lebesgue space $L^p(\Omega, \mathbb{R}^{2\times 2})$ (which consists of L^p-integrable functions with values in $\mathbb{R}^{2\times 2}$) with the norm

$$\|A\|_{L^p(\Omega,\mathbb{R}^{2\times 2})} := \left(\sum_{i,j=1}^{2} \|A_{ij}\|^2_{L^p(\Omega)} \right)^{\frac{1}{2}} \quad \text{for all } A \in L^p(\Omega, \mathbb{R}^{2\times 2}).$$

In the case $p = 2$ we also define the corresponding inner product

$$\langle A, B \rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} := \sum_{i,j=1}^{2} \langle A_{ij}, B_{ij} \rangle_{L^2(\Omega)} \quad \text{for all } A, B \in L^2(\Omega, \mathbb{R}^{2\times 2}). \tag{2.3}$$

In the further course we denote the rows of a function $A \in L^2(\Omega, \mathbb{R}^{2\times 2})$ by $A_1 \in L^2(\Omega, \mathbb{R}^2)$ and $A_2 \in L^2(\Omega, \mathbb{R}^2)$, i.e., we have

$$A = \begin{pmatrix} -A_1- \\ -A_2- \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Moreover, for two functions $A, B \in L^2(\Omega, \mathbb{R}^{2\times 2})$ we define the Frobenius product

$$A \bullet B := \sum_{i,j=1}^{2} A_{ij} B_{ij} = \sum_{i=1}^{2} A_i \cdot B_i.$$

Thereby we can rewrite the inner product on $L^2(\Omega, \mathbb{R}^{2\times 2})$ defined in (2.3) as follows

$$\langle A, B \rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} = \sum_{i=1}^{2} \langle A_i, B_i \rangle_{L^2(\Omega,\mathbb{R}^2)} = \int_\Omega A \bullet B \,\mathrm{d}(x,y) \quad \text{for all } A, B \in L^2(\Omega, \mathbb{R}^{2\times 2}).$$

In the same way, for $p \in [1, \infty]$ we obtain

$$\|A\|_{L^p(\Omega,\mathbb{R}^{2\times 2})} = \left(\sum_{i=1}^{2} \|A_i\|^2_{L^p(\Omega,\mathbb{R}^2)} \right)^{\frac{1}{2}} \quad \text{for all } A \in L^p(\Omega, \mathbb{R}^{2\times 2}) \tag{2.4}$$

for the corresponding norm.

Next, we endow the velocity space $H_0^1(\Omega, \mathbb{R}^2)$ (and thus its subspace $H(\Omega)$) with the inner product

$$\langle u, v \rangle_{H_0^1(\Omega,\mathbb{R}^2)} := \langle \nabla u, \nabla v \rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \sigma \langle u, v \rangle_{L^2(\Omega,\mathbb{R}^2)} \quad \text{for all } u, v \in H_0^1(\Omega, \mathbb{R}^2), \tag{2.5}$$

where $\nabla u \in L^2(\Omega, \mathbb{R}^{2\times 2})$ is defined as the row-wise gradients of the components of the velocity $u \in H_0^1(\Omega, \mathbb{R}^2)$, i.e.,

$$\nabla u = \begin{pmatrix} -\nabla u_1- \\ -\nabla u_2- \end{pmatrix} = \begin{pmatrix} \frac{\partial u_1}{\partial x} & \frac{\partial u_1}{\partial y} \\ \frac{\partial u_2}{\partial x} & \frac{\partial u_2}{\partial y} \end{pmatrix},$$

and σ is a non-negative constant to be specified later (see Chapter 6). It is clear that $H_0^1(\Omega, \mathbb{R}^2)$ together with the inner product defined in (2.5) is a Hilbert space. Since $\operatorname{div}\colon H_0^1(\Omega, \mathbb{R}^2) \to L^2(\Omega)$ is a bounded linear operator (cf. Lemma A.1), $H(\Omega) = \ker(\operatorname{div})$ forms a closed subspace of $H_0^1(\Omega, \mathbb{R}^2)$, and hence, it is a Hilbert space with the same inner product. Note that $\sigma = 0$ is a possible choice since our domain Ω is contained in the unbounded strip S and Poincaré's inequality holds for S and thus also for Ω (see Lemma A.4 and Remark A.5 in Appendix A.1).

Remark 2.1. *Using the representation for the inner product introduced above together with this definition of the derivative, our Navier-Stokes equations* (1.15) *can be written in the following form:*

Find $u \in H(\Omega)$ such that

$$\int_\Omega \left(\nabla u \bullet \nabla \varphi + \mathit{Re}\,[(u \cdot \nabla)u + (u \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u] \cdot \varphi \right) \mathrm{d}(x,y)$$
$$= \int_\Omega g \cdot \varphi \,\mathrm{d}(x,y) \quad \textit{for all } \varphi \in H(\Omega).$$

In addition to the inner product, we get the corresponding norm

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} := \left(\|\nabla u\|^2_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \sigma\|u\|^2_{L^2(\Omega,\mathbb{R}^2)}\right)^{\frac{1}{2}} \quad \text{for all } u \in H_0^1(\Omega,\mathbb{R}^2) \tag{2.6}$$

on $H_0^1(\Omega,\mathbb{R}^2)$ and $H(\Omega)$ respectively.

Using the definitions (2.2) and (2.3) respectively, and rearranging the terms appearing in the definition of the inner product on $H(\Omega)$ (see (2.5)) we obtain

$$\begin{aligned}\langle u,v\rangle_{H_0^1(\Omega,\mathbb{R}^2)} &= \sum_{i=1}^{2}\left(\langle \nabla u_i, \nabla v_i\rangle_{L^2(\Omega,\mathbb{R}^2)} + \sigma\langle u_i, v_i\rangle_{L^2(\Omega)}\right)\\ &= \sum_{i=1}^{2}\langle u_i, v_i\rangle_{H_0^1(\Omega)} \quad \text{for all } u,v \in H_0^1(\Omega,\mathbb{R}^2),\end{aligned} \tag{2.7}$$

which gives an alternative representation of the inner product on $H_0^1(\Omega,\mathbb{R}^2)$ using the inner product $\langle\cdot,\cdot\rangle_{H_0^1(\Omega)} = \langle\cdot,\cdot\rangle_{L^2(\Omega,\mathbb{R}^2)} + \sigma\langle\cdot,\cdot\rangle_{L^2(\Omega)}$ on $H_0^1(\Omega)$ suggested for the setting of computer-assisted proofs by Plum in [85, p. 34]. In almost the same manner we get an alternative representation for the $\|\cdot\|_{H_0^1(\Omega,\mathbb{R}^2)}$-norm using the $\|\cdot\|_{H_0^1(\Omega)}$-norm.

Finally, we will need the subspace of $L^2(\Omega,\mathbb{R}^{2\times 2})$ where the row-wise divergence is an element of $L^2(\Omega)$, i.e., we define

$$H(\operatorname{div},\Omega,\mathbb{R}^{2\times 2}) := \left\{A \in L^2(\Omega,\mathbb{R}^{2\times 2})\colon \operatorname{div} A_1, \operatorname{div} A_2 \in L^2(\Omega)\right\},$$

where $A_1, A_2 \in L^2(\Omega,\mathbb{R}^2)$ again denote the rows of $A \in L^2(\Omega,\mathbb{R}^{2\times 2})$. Given a function $A \in H(\operatorname{div},\Omega,\mathbb{R}^{2\times 2})$, we set

$$\operatorname{div}\colon H(\operatorname{div},\Omega,\mathbb{R}^{2\times 2}) \to L^2(\Omega,\mathbb{R}^2),\ \operatorname{div} A := \begin{pmatrix}\operatorname{div} A_1\\ \operatorname{div} A_2\end{pmatrix}. \tag{2.8}$$

2.2 Topological Dual Space of $H(\Omega)$

In the further course for a (bounded) linear operator $F\colon X \to Y$ between two normed spaces X and Y we denote its usual operator norm by $\|F\|_{\mathcal{B}(X,Y)} = \|F\|_{\mathcal{B}}$, where we omit the spaces in the index if they are clear from the context.

According to the abstract setting described in [85, p. 34], we introduce the topological dual space of $H(\Omega)$ which will be denoted by $H(\Omega)'$ and is endowed with the usual dual norm $\|\cdot\|_{H(\Omega)'}$.

Next, we show that $-\Delta$ can be identified with an operator mapping $H(\Omega)$ into its dual space $H(\Omega)'$, i.e., we have $-\Delta u \in H(\Omega)'$ for any function $u \in H(\Omega)$. Therefore, using formal integration by parts, for any $A \in L^2(\Omega,\mathbb{R}^{2\times 2})$ we first define the functional

$$\operatorname{div} A\colon H(\Omega) \to \mathbb{R},\ (\operatorname{div} A)[\varphi] := -\int_\Omega A \bullet \nabla\varphi \,\mathrm{d}(x,y) \quad \text{for all } \varphi \in H(\Omega). \tag{2.9}$$

Cauchy-Schwarz' inequality and the definition of the norm on $H(\Omega)$ (see (2.6)) imply

$$\begin{aligned}|(\operatorname{div} A)[\varphi]| &\le \int_\Omega |A \bullet \nabla\varphi| \,\mathrm{d}(x,y)\\ &\le \|A\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}\|\nabla\varphi\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \le \|A\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}\|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega).\end{aligned}$$

Hence, we obtain

$$\|\operatorname{div} A\|_{H(\Omega)'} \leq \|A\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}, \tag{2.10}$$

i.e., $\operatorname{div} A$ is indeed a bounded linear functional on $H(\Omega)$ and thus, $\operatorname{div} A$ is an element in $H(\Omega)'$ for all $A \in L^2(\Omega,\mathbb{R}^{2\times 2})$. Moreover, since $\nabla u \in L^2(\Omega,\mathbb{R}^{2\times 2})$ for $u \in H(\Omega)$ the definition above applied to ∇u yields

$$(-\Delta u)[\varphi] \coloneqq -(\operatorname{div} \nabla u)[\varphi] = \int_\Omega \nabla u \bullet \nabla \varphi \,\mathrm{d}(x,y) \quad \text{for all } \varphi \in H(\Omega). \tag{2.11}$$

Next, we analyze some embeddings needed at several stages in this thesis. Sobolev's Embedding Theorem [2, Theorem 5.4] yields $H_0^1(\Omega) \subseteq L^p(\Omega)$ with a bounded embedding $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$ for every $p \in [2,\infty)$, i.e., there exists some constant $C_p > 0$ satisfying $\|u\|_{L^p(\Omega)} \leq C_p \|u\|_{H_0^1(\Omega)}$ for all $u \in H_0^1(\Omega)$ (recall that we are in space dimension 2). Thus, using the definition of the $L^p(\Omega,\mathbb{R}^2)$-norm (see (2.1)) and the embedding described before as well as the alternative representation formula of the inner product on X (see (2.7)) we obtain an embedding $H_0^1(\Omega,\mathbb{R}^2) \hookrightarrow L^p(\Omega,\mathbb{R}^2)$ which satisfies

$$\begin{aligned} \|u\|_{L^p(\Omega,\mathbb{R}^2)} &= \left(\sum_{i=1}^{2} \|u_i\|_{L^p(\Omega)}^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{2} {C_p}^2 \|u_i\|_{H_0^1(\Omega)}^2\right)^{\frac{1}{2}} \\ &\leq C_p \left(\sum_{i=1}^{2} \|u_i\|_{H_0^1(\Omega)}^2\right)^{\frac{1}{2}} = C_p \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H_0^1(\Omega,\mathbb{R}^2) \end{aligned} \tag{2.12}$$

for every $p \in [2,\infty)$, where the embedding constant C_p remains the same as in the usual case of Sobolev's Embedding Theorem for functions with values in $\mathbb{R}$.

Thus, for every $u \in L^q(\Omega,\mathbb{R}^2)$ with $q \in (1,2]$ these embeddings together with Hölder's inequality (with $\frac{1}{p} + \frac{1}{q} = 1$) imply

$$\begin{aligned} \left| \int_\Omega u \cdot \varphi \,\mathrm{d}(x,y) \right| &\leq \|u\|_{L^q(\Omega,\mathbb{R}^2)} \|\varphi\|_{L^p(\Omega,\mathbb{R}^2)} \\ &\leq C_p \|u\|_{L^q(\Omega,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H_0^1(\Omega,\mathbb{R}^2). \end{aligned}$$

Hence, since $H(\Omega) \subseteq H_0^1(\Omega,\mathbb{R}^2)$, $u\colon H(\Omega) \to \mathbb{R}$ defined by

$$u[\varphi] \coloneqq \int_\Omega u \cdot \varphi \,\mathrm{d}(x,y) \quad \text{for all } \varphi \in H(\Omega) \tag{2.13}$$

is a bounded linear functional on $H(\Omega)$ (satisfying $\|u\|_{H(\Omega)'} \leq C_p \|u\|_{L^q(\Omega,\mathbb{R}^2)}$), i.e., for all $q \in (1,2]$ we can identify $u \in L^q(\Omega,\mathbb{R}^2)$ with an element in $H(\Omega)'$.

Finally, we have a closer look at the divergence operator again. Thus, for functions with higher regularity, which especially is the case for functions in $A \in H(\operatorname{div},\Omega,\mathbb{R}^{2\times 2})$ or $A \in H_0^1(\Omega,\mathbb{R}^2)$, we can use integration by parts and obtain

$$\begin{aligned} -\int_\Omega A \bullet \nabla \varphi \,\mathrm{d}(x,y) &= -\sum_{i=1}^{2} \int_\Omega A_i \cdot \nabla \varphi_i \,\mathrm{d}(x,y) = \sum_{i=1}^{2} \int_\Omega \operatorname{div}(A_i) \varphi_i \,\mathrm{d}(x,y) \\ &= \int_\Omega \begin{pmatrix} \operatorname{div} A_1 \\ \operatorname{div} A_2 \end{pmatrix} \cdot \varphi \,\mathrm{d}(x,y) \quad \text{for all } \varphi \in H(\Omega), \end{aligned}$$

hence, by (2.13) the weak divergence defined in (2.9) coincides with the "classical" row-wise divergence (2.8) (which is an element of $L^2(\Omega,\mathbb{R}^2)$) read as an element in $H(\Omega)'$.

2.3 Isometric Isomorphism $\Phi\colon H(\Omega) \to H(\Omega)'$

Using the techniques presented above (cf. (2.11) and (2.13)), the operator

$$\Phi\colon H(\Omega) \to H(\Omega)',\ \Phi u := -\Delta u + \sigma u, \tag{2.14}$$

is well defined and we get

$$(\Phi u)[\varphi] = \int_\Omega (\nabla u \bullet \nabla\varphi + \sigma u \cdot \varphi)\,\mathrm{d}(x,y) = \langle u, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u, v \in H(\Omega). \tag{2.15}$$

Moreover, Riesz' Representation Lemma for bounded linear functionals implies that Φ defines an isometric isomorphism from $H(\Omega)$ into its dual space $H(\Omega)'$. Hence, the identity $\|\Phi u\|_{H(\Omega)'} = \|u\|_{H_0^1(\Omega,\mathbb{R}^2)}$ holds true for all $u \in H(\Omega)$ (see Lemma A.6).

Hence, by (2.15), we obtain

$$\langle \Phi^{-1} f, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} = (\Phi(\Phi^{-1} f))[\varphi] = f[\varphi] \quad \text{for all } f \in H(\Omega)', \varphi \in H(\Omega). \tag{2.16}$$

2.4 Fourier Transform on the Unbounded Strip

In Section 6.2.1.4 we require a Fourier transform defined on $L^2(S, \mathbb{R}^2)$ or $L^2(S, \mathbb{C}^2)$ respectively. Recall that S denotes the infinite strip $\mathbb{R} \times (0,1)$ (cf. Chapter 1). For the reader's convenience, we first construct a version on $L^2(S, \mathbb{C})$ which can easily be extended to a Fourier transform on $L^2(S, \mathbb{C}^2)$ later on. To define such a variant of the Fourier transform, we first consider the Fourier series basis $\{\varphi_n\}_{n\in\mathbb{Z}}$ on $(0,1)$ defined by

$$\varphi_n\colon (0,1) \to \mathbb{C},\ \varphi_n(y) = \mathrm{e}^{-2i\pi n y} \quad \text{for all } n \in \mathbb{Z}. \tag{2.17}$$

Using the usual Schwartz space $\mathcal{S}(\mathbb{R}, \mathbb{C})$ on $\mathbb{R}$, we define

$$\mathcal{D} := \left\{ \sum_{n\in\mathbb{Z}} u_n \varphi_n \colon u_n \in \mathcal{S}(\mathbb{R}, \mathbb{C}),\ u_n = 0 \text{ for almost all } n \in \mathbb{Z} \right\} \subseteq L^2(S, \mathbb{C}). \tag{2.18}$$

Remark 2.2. *Having a closer look at the definition of $\mathcal{D}$ again, we directly obtain*

$$\mathcal{D} = \left\{ \sum_{n=-N}^{N} u_n \varphi_n \colon u_n \in \mathcal{S}(\mathbb{R}, \mathbb{C}),\ N \in \mathbb{N}_0 \right\}.$$

Next, we define a "new" Fourier transform "in x-direction" (which will be indicated by the index x) on the subspace $\mathcal{D}$ using the usual Fourier transform $\mathcal{F}$ on $L^2(\mathbb{R}, \mathbb{C})$ (or on $\mathcal{S}(\mathbb{R}, \mathbb{C})$ respectively). Therefore, for $u \in \mathcal{D}$ we set

$$\mathcal{F}_x[u](\xi, y) := \sum_{n=-N}^{N} \mathcal{F}[u_n](\xi)\varphi_n(y) \quad \text{for all } (\xi, y) \in S, \tag{2.19}$$

which, since $\mathcal{F}[u_n] \in \mathcal{S}(\mathbb{R}, \mathbb{C})$ for all $n \in \mathbb{Z}$, directly implies $\mathcal{F}_x[u] \in \mathcal{D}$. Moreover, direct calculations show that

$$\mathcal{F}_x^{-1}\colon \mathcal{D} \to \mathcal{D},\ \mathcal{F}_x^{-1}[u](\xi, y) := \sum_{n=-N}^{N} \mathcal{F}^{-1}[u_n](\xi)\varphi_n(y) \tag{2.20}$$

defines the inverse Fourier transform on $\mathcal{D}$, i.e., $\mathcal{F}_x\colon \mathcal{D} \to \mathcal{D}$ is bijective (cf. Lemma A.7).

Remark 2.3. *By definition of the well-known Fourier transform on the Schwartz space we can use its integral representation formula to rewrite the definition of our new Fourier transform on the strip and its inverse as follows:*

$$\mathcal{F}_x[u](\xi,y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x,y)\mathrm{e}^{-\mathrm{i}x\xi}\,\mathrm{d}x \quad \textit{and} \quad \mathcal{F}_x^{-1}[v](\xi,y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} v(\xi,y)\mathrm{e}^{\mathrm{i}\xi x}\,\mathrm{d}\xi$$

for all $u, v \in \mathcal{D}$ and $(\xi, y) \in S$.

Using the fact that $\mathcal{F}$ satisfies

$$\int_{-\infty}^{\infty} \mathcal{F}[u](x)v(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} u(x)\mathcal{F}[v](x)\,\mathrm{d}x \quad \text{for all } u,v \in \mathcal{S}(\mathbb{R},\mathbb{C}),$$

(see [93, proof of Corollary 16.12, p. 141]) we can transfer this property to our new Fourier transform $\mathcal{F}_x$ and obtain

$$\begin{aligned} \int_S \mathcal{F}_x[u](x,y)v(x,y)\,\mathrm{d}(x,y) &= \sum_{n=-N}^{N}\sum_{m=-M}^{M} \int_0^1 \left(\int_{-\infty}^{\infty} \mathcal{F}[u_n](x)v_m(x)\,\mathrm{d}x\right)\varphi_n(y)\varphi_m(y)\,\mathrm{d}y \\ &= \sum_{n=-N}^{N}\sum_{m=-M}^{M} \int_0^1 \left(\int_{-\infty}^{\infty} u_n(x)\mathcal{F}[v_m](x)\,\mathrm{d}x\right)\varphi_n(y)\varphi_m(y)\,\mathrm{d}y \\ &= \int_S u(x,y)\mathcal{F}_x[v](x,y)\,\mathrm{d}(x,y) \quad \text{for all } u,v \in \mathcal{D}. \end{aligned} \tag{2.21}$$

Moreover, employing the integral representation of $\mathcal{F}_x$ mentioned in Remark 2.3, we calculate

$$\begin{aligned} \overline{\mathcal{F}_x[u](x,y)} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \overline{u(x,y)\mathrm{e}^{-\mathrm{i}x\xi}}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \overline{u(x,y)}\mathrm{e}^{\mathrm{i}x\xi}\,\mathrm{d}x = \mathcal{F}_x^{-1}[\overline{u}](x,y) \quad \text{for all } u \in \mathcal{D}. \end{aligned} \tag{2.22}$$

Thus, (2.21) together with (2.22) yields

$$\begin{aligned} \|\mathcal{F}_x[u]\|_{L^2(S,\mathbb{C})}^2 &= \int_S \mathcal{F}_x[u](x,y)\overline{\mathcal{F}_x[u](x,y)}\,\mathrm{d}(x,y) \\ &= \int_S \mathcal{F}_x[u](x,y)\mathcal{F}_x^{-1}[\overline{u}](x,y)\,\mathrm{d}(x,y) \\ &= \int_S u(x,y)\mathcal{F}_x\left[\mathcal{F}_x^{-1}[\overline{u}]\right](x,y)\,\mathrm{d}(x,y) \\ &= \int_S u(x,y)\overline{u(x,y)}\,\mathrm{d}(x,y) \\ &= \|u\|_{L^2(S,\mathbb{C})}^2 \quad \text{for all } u \in \mathcal{D}, \end{aligned} \tag{2.23}$$

hence, $\mathcal{F}_x$ defines an isometric isomorphism (with respect to the $L^2(S,\mathbb{C})$-norm) from $\mathcal{D}$ into $\mathcal{D}$.

Similar to the usual Fourier transform on $\mathcal{S}(\mathbb{R},\mathbb{C})$ the following Lemma provides some properties concerning derivatives for the new Fourier transform $\mathcal{F}_x$.

Lemma 2.4. *Let $j, k \in \mathbb{N}_0$. Then, for $u \in \mathcal{D}$ the following assertions hold true:*

(i) $\mathcal{F}_x\left[\frac{\partial^{j+k}u}{\partial x^j \partial y^k}\right](\xi, y) = (\mathrm{i}\xi)^j \left(\frac{\partial^k}{\partial y^k}\mathcal{F}_x[u]\right)(\xi, y)$ *for all* $(\xi, y) \in S$.

(ii) *For* $v := (-\mathrm{i}x)^j u$ *we have* $\mathcal{F}_x\left[\frac{\partial^k v}{\partial y^k}\right](\xi, y) = \left(\frac{\partial^{j+k}}{\partial \xi^j \partial y^k}\mathcal{F}_x[u]\right)(\xi, y)$ *for all* $(\xi, y) \in S$.

Since the proof of Lemma 2.4 is rather technical it will be postponed to the Appendix (see Proof of Lemma 2.4 in Appendix A.2).

Next, we are going to extend the definition of the new Fourier transform $\mathcal{F}_x$ on the subspace $\mathcal{D}$ (see (2.19)) to the larger space $L^2(S, \mathbb{C})$. Therefore, we prove the following density result first.

Proposition 2.5. *$\mathcal{D}$ is dense in $L^2(S, \mathbb{C})$ (with respect to the $L^2(S, \mathbb{C})$-norm).*

Proof. Let $u \in L^2(S, \mathbb{C})$. Then, for almost every fixed $x \in \mathbb{R}$ we define

$$u_n(x) := \langle u(x, \cdot), \varphi_n \rangle_{L^2((0,1),\mathbb{C})} \quad \text{for all } n \in \mathbb{Z},$$

and consider the Fourier series representation

$$u(x, \cdot) = \sum_{n \in \mathbb{Z}} u_n(x) \varphi_n,$$

which converges with respect to the $L^2((0,1), \mathbb{C})$-norm. By Parseval's theorem we obtain

$$\|u(x, \cdot)\|^2_{L^2((0,1),\mathbb{C})} = \sum_{n \in \mathbb{Z}} |u_n(x)|^2$$

for almost every $x \in \mathbb{R}$, implying

$$\begin{aligned} \|u_n\|^2_{L^2(\mathbb{R},\mathbb{C})} = \int_{-\infty}^{\infty} |u_n(x)|^2 \, \mathrm{d}x &\leq \int_{-\infty}^{\infty} \sum_{n \in \mathbb{Z}} |u_n(x)|^2 \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \|u(x, \cdot)\|^2_{L^2((0,1),\mathbb{C})} \, \mathrm{d}x = \|u\|^2_{L^2(S,\mathbb{C})} \quad \text{for all } n \in \mathbb{Z}, \end{aligned}$$

hence, $u_n \in L^2(S, \mathbb{C})$ for all $n \in \mathbb{Z}$. Moreover, applying the monotone convergence theorem to the (monotone) series $(\sum_{n=-N}^{N} |u_n|^2)_{N \in \mathbb{N}}$, we get

$$\|u\|^2_{L^2(S,\mathbb{C})} = \int_{-\infty}^{\infty} \sum_{n \in \mathbb{Z}} |u_n(x)|^2 \, \mathrm{d}x = \sum_{n \in \mathbb{Z}} \int_{-\infty}^{\infty} |u_n(x)|^2 \, \mathrm{d}x = \sum_{n \in \mathbb{Z}} \|u_n\|^2_{L^2(\mathbb{R},\mathbb{C})} \tag{2.24}$$

Now, let $\varepsilon > 0$. Since $\mathcal{S}(\mathbb{R}, \mathbb{C})$ is dense in $L^2(\mathbb{R}, \mathbb{C})$ and u_n is square integrable for all $n \in \mathbb{Z}$, we find $v_n \in \mathcal{S}(\mathbb{R}, \mathbb{C})$ such that

$$\|u_n - v_n\|^2_{L^2(\mathbb{R},\mathbb{C})} \leq \frac{\varepsilon}{2^{|n|+2}} \quad \text{for all } n \in \mathbb{Z}.$$

Moreover, due to (2.24) $\sum_{n \in \mathbb{Z}} \|u_n\|^2_{L^2(\mathbb{R},\mathbb{C})}$ converges, and thus, we can chose $N \in \mathbb{N}$ such that

$$\sum_{|n| > N} \|u_n\|^2_{L^2(\mathbb{R},\mathbb{C})} \leq \frac{\varepsilon}{4}.$$

Then, again Parseval's theorem and the monotone convergence theorem yield

$$\begin{aligned}\left\| u - \sum_{n=-N}^{N} v_n \varphi_n \right\|_{L^2(S,\mathbb{C})}^2 &= \sum_{n=-N}^{N} \|u_n - v_n\|_{L^2(\mathbb{R},\mathbb{C})}^2 + \sum_{|n|>N} \|u_n\|_{L^2(\mathbb{R},\mathbb{C})}^2 \\ &\leq \sum_{n=-N}^{N} \frac{\varepsilon}{2^{|n|+2}} + \frac{\varepsilon}{4} = \frac{\varepsilon}{4} + 2 \cdot \frac{\varepsilon}{4} \left(\sum_{n=1}^{N} \frac{1}{2^n} \right) + \frac{\varepsilon}{4} \leq \varepsilon.\end{aligned}$$

□

Remark 2.6. (i) *Since, by Proposition 2.5 above, $\mathcal{D}$ is dense in $L^2(S,\mathbb{C})$, $\mathcal{F}_x$ can be extended to an isometric isomorphism $\mathcal{F}_x \colon L^2(S,\mathbb{C}) \to L^2(S,\mathbb{C})$, i.e., the equality in* (2.23) *holds true for all $u \in L^2(S,\mathbb{C})$. Moreover, for all $u, v \in L^2(S,\mathbb{C})$ we obtain*

$$\int_S \mathcal{F}_x[u](x,y) v(x,y) \, \mathrm{d}(x,y) = \int_S u(x,y) \mathcal{F}_x[v](x,y) \, \mathrm{d}(x,y) \quad \textit{for all } u, v \in L^2(S,\mathbb{C})$$

and

$$\overline{\mathcal{F}_x[u](x,y)} = \mathcal{F}_x^{-1}[\overline{u}](x,y)$$

(cf. (2.21) *and* (2.22)*)*

(ii) *Similar to the usual Fourier transform on $\mathbb{R}$ or $\mathbb{R}^d$, the new Fourier transform in x-direction defined above induces an isometric isomorphism*

$$\mathcal{F}_x \colon L^2(S) \to \Big\{ v \in L^2(S,\mathbb{C}) \colon v(-x,y) = \overline{v(x,y)} \textit{ for almost every } (x,y) \in S \Big\}.$$

Recall that $L^2(S)$ denotes the Lebesgue space of real-valued square integrable functions.

Moreover, in this thesis we will need a distributional version of the Fourier transform $\mathcal{F}_x$ defined above. Therefore, again analogous to the "usual" Fourier transform we consider

$$\mathcal{F}_x \colon \mathcal{D}' \to \mathcal{D}', \; (\mathcal{F}_x[f])[\varphi] := f[\mathcal{F}_x[\varphi]] \quad \text{for all } \varphi \in \mathcal{D}, \tag{2.25}$$

where $\mathcal{D}'$ denotes the topological dual space of $\mathcal{D}$. Again, we easily see that

$$\mathcal{F}_x^{-1} \colon \mathcal{D}' \to \mathcal{D}', \; (\mathcal{F}_x^{-1}[f])[\varphi] := f[\mathcal{F}_x^{-1}[\varphi]] \quad \text{for all } \varphi \in \mathcal{D}, \tag{2.26}$$

defines its inverse (cf. Lemma A.8) and thus, $\mathcal{F}_x \colon \mathcal{D}' \to \mathcal{D}'$ is bijective as well. For $u \in L^2(S,\mathbb{C})$ we use the usual interpretation as linear functionals on $\mathcal{D}$, i.e., we have

$$f_u[\varphi] := u[\varphi] := \int_S u(x,y)\varphi(x,y) \, \mathrm{d}(x,y) \quad \text{for all } \varphi \in \mathcal{D}.$$

Then, the equality (2.21) implies

$$\begin{aligned}(\mathcal{F}_x[f_u])[\varphi] = f_u[\mathcal{F}_x[\varphi]] &= \int_S u(x,y) \mathcal{F}_x[\varphi](x,y) \, \mathrm{d}(x,y) \\ &= \int_S \mathcal{F}_x[u](x,y) \varphi(x,y) \, \mathrm{d}(x,y) = f_{\mathcal{F}_x[u]}[\varphi] \quad \text{for all } \varphi \in \mathcal{D}.\end{aligned} \tag{2.27}$$

Moreover, we define derivatives of an element $f \in \mathcal{D}'$ in the following sense:

$$\left(\frac{\partial^{j+k} f}{\partial x^j \partial y^k}\right)[\varphi] := (-1)^{j+k} f\left[\frac{\partial^{j+k}\varphi}{\partial x^j \partial y^k}\right] \quad \text{for all } \varphi \in \mathcal{D}, \tag{2.28}$$

where $j, k \in \mathbb{N}_0$. In a similar way, for $f \in \mathcal{D}'$ and $x \in \mathbb{R}$ we set

$$(x^j f)[\varphi] := f[x^j \varphi] \quad \text{for all } \varphi \in \mathcal{D}. \tag{2.29}$$

Now, we are in a position to extend Lemma 2.4 to functions on the dual space $\mathcal{D}'$:

Lemma 2.7. *Let $j, k \in \mathbb{N}_0$. Then, for $f \in \mathcal{D}'$ the following assertions hold true:*

(i) $\mathcal{F}_x\left[\frac{\partial^{j+k} f}{\partial x^j \partial y^k}\right] = (\mathrm{i}x)^j \frac{\partial^k}{\partial y^k}\mathcal{F}_x[f]$ *(in $\mathcal{D}'$).*

(ii) *For $g := (-\mathrm{i}x)^j f$ we have $\mathcal{F}_x\left[\frac{\partial^k g}{\partial y^k}\right] = \frac{\partial^{j+k}}{\partial \xi^j \partial y^k}\mathcal{F}_x[f]$ (in $\mathcal{D}'$).*

Since the distributional Fourier transform is defined using the non-distributional one, the proof of Lemma 2.7 is just an application of Lemma 2.4 and will be postponed also to the Appendix (see Proof of Lemma 2.7 in Appendix A.2).

To the end of this Section, we prove a characterization of the (real-valued) Sobolev space $H^1(S)$ via the “new” Fourier transform $\mathcal{F}_x$ which will be very useful later on in Section 6.2.2. Therefore, as a first step, we show the following representation for the complex valued case.

Proposition 2.8. *The following representation formula holds true:*

$$H^1(S,\mathbb{C}) = \left\{u \in L^2(S,\mathbb{C})\colon \int_S (1+|\xi|^2)\, |\mathcal{F}_x[u](\xi,y)|^2 \,\mathrm{d}(\xi,y) < \infty,\ \frac{\partial u}{\partial y} \in L^2(S,\mathbb{C})\right\}.$$

Proof. First, we notice that

$$\begin{aligned}&\left\{u \in L^2(S,\mathbb{C})\colon \int_S (1+|\xi|^2)\, |\mathcal{F}_x[u](\xi,y)|^2 \,\mathrm{d}(\xi,y) < \infty,\ \frac{\partial u}{\partial y} \in L^2(S,\mathbb{C})\right\}\\ &\qquad = \left\{u \in L^2(S,\mathbb{C})\colon \int_S |\xi|^2\, |\mathcal{F}_x[u](\xi,y)|^2 \,\mathrm{d}(\xi,y) < \infty,\ \frac{\partial u}{\partial y} \in L^2(S,\mathbb{C})\right\}.\end{aligned}$$

Now, let $u \in L^2(S,\mathbb{C})$ be fixed. Then, by the bijectivity of the distributional Fourier transform and Lemma 2.7 (i), we obtain

$$\begin{aligned}\frac{\partial u}{\partial x} \in L^2(S,\mathbb{C}) \quad &\Leftrightarrow \quad \exists_{v\in L^2(S,\mathbb{C})}\colon \frac{\partial u}{\partial x} = v\\ &\Leftrightarrow \quad \exists_{v\in L^2(S,\mathbb{C})}\colon \frac{\partial}{\partial x} f_u = f_v\\ &\Leftrightarrow \quad \exists_{v\in L^2(S,\mathbb{C})}\colon \mathcal{F}_x\left[\frac{\partial}{\partial x} f_u\right] = \mathcal{F}_x[f_v]\\ &\Leftrightarrow \quad \exists_{v\in L^2(S,\mathbb{C})}\colon \mathrm{i}x\mathcal{F}_x[f_u] = \mathcal{F}_x[f_v]\end{aligned}$$

Then, (2.27) and the isometric property of $\mathcal{F}_x$ (cf. Remark 2.6 (i)) imply

$$\begin{aligned} \exists_{v\in L^2(S,\mathbb{C})}\colon \mathrm{i}x\mathcal{F}_x[f_u] = \mathcal{F}_x[f_v] \quad &\Leftrightarrow \quad \exists_{v\in L^2(S,\mathbb{C})}\colon \mathrm{i}x f_{\mathcal{F}_x[u]} = f_{\mathcal{F}_x[v]} \\ &\Leftrightarrow \quad \exists_{w\in L^2(S,\mathbb{C})}\colon x f_{\mathcal{F}_x[u]} = f_w, \end{aligned}$$

where we use the transformation $w := -\mathrm{i}\mathcal{F}_x[v]$ in the last step. Thus, we obtain

$$\begin{aligned} \exists_{w\in L^2(S,\mathbb{C})}\colon x f_{\mathcal{F}_x[u]} = f_w \quad &\Leftrightarrow \quad \exists_{w\in L^2(S,\mathbb{C})}\colon x\mathcal{F}_x[u] = w \\ &\Leftrightarrow \quad x\mathcal{F}_x[u] \in L^2(S,\mathbb{C}). \end{aligned}$$

Hence, combining all arguments above yields the assertion. □

Remark 2.9. *Considering only real valued functions, Proposition 2.8 implies the representation formula*

$$H^1(S) = \left\{ u \in L^2(S)\colon \int_S (1+|\xi|^2)\,|\mathcal{F}_x[u](\xi,y)|^2\,\mathrm{d}(\xi,y) < \infty,\ \frac{\partial u}{\partial y} \in L^2(S) \right\}$$

for the real valued Sobolev space. Note that the Fourier transform of u appearing in the integral might be complex-valued.

3 A Computer-assisted Proof for the Navier-Stokes Equations

In general, computer-assisted proofs in the field of ordinary or partial differential equations require a zero-finding formulation of the underlying problem. Therefore, we reformulate our problem as a zero-finding problem which will be discussed in the next Section. To guarantee a rigorous (analytical) proof of the existence of an exact solution a fixed-point argument is applied, for example Schauder's fixed-point Theorem in the case of bounded domains or Banach's fixed-point Theorem for unbounded domains (cf. [85, Section 2]). We use the current Chapter to describe the crucial steps needed for our computer-assisted existence proof for the Navier-Stokes equations (1.15).

The computer-assisted methods used are mainly based on techniques introduced by Plum (cf. [74, Section 6.1] and [85]). Starting from a numerical approximate solution of the Navier-Stokes equations (for the computation see Chapter 4), we compute a bound for its defect which will be described in Chapter 5. In addition to that, a norm bound for the inverse of the linearization at the approximate solution is needed. Therefore, bounds for the essential spectrum and the eigenvalues "close to" zero are required. The methods applied, for example the well-known Rayleigh Ritz method and a corollary of the Temple-Lehmann Theorem to obtain enclosures of the crucial eigenvalues of the linearization below the essential spectrum, are presented in Chapter 6.

With these data in hand, we use a fixed-point argument (see Theorem 3.4) to prove existence of an exact solution "nearby" the approximate one. In addition to the pure existence result, the methods in use also provide an enclosure of the exact solution measured in a suitable Sobolev norm.

Before reformulating our Navier-Stokes equations (1.15) as a zero-finding problem, we will use the techniques described in the previous Chapter to identify our problem with an equation in $H(\Omega)'$. By assumption, we have $V \in H^2(\Omega, \mathbb{R}^2) \cap C^1(\overline{\Omega}, \mathbb{R}^2)$ which, together with the definition of the right-hand side (cf. (1.14)), implies $g \in L^2(\Omega, \mathbb{R}^2)$. Thus, definition (2.13) shows that g indeed defines a bounded linear functional on $H(\Omega)$ which will be denoted by $g \in H(\Omega)'$ in the further course again. It becomes clear from the context whether $g \in L^2(\Omega, \mathbb{R}^2)$ or $g \in H(\Omega)'$ is the right identification.

Furthermore, applying (2.13) and the definition of the weak Laplacian (cf. (2.11)) our weak formulation (1.15) can be identified with an equation in $H(\Omega)'$, i.e., we consider the following equivalent formulation of the problem:

$$\text{Find } u \in H(\Omega) \text{ such that,} \quad -\Delta u + Re\,[(u \cdot \nabla)u + (u \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u] = g.$$

Again, we note that by construction the divergence-free part of our Navier-Stokes equations is modeled in the space $H(\Omega)$ (cf. Section 1.3).

3.1 Reformulation as a Zero-Finding Problem

To rewrite the weak formulation of the Navier-Stokes equations (1.15) as a zero-finding problem, we have a closer look at the different terms appearing in our weak formulation and define some auxiliary operators.

Therefore, we start with the non-linear term in our weak formulation (1.15) which in the following will be represented by the form

$$\mathrm{B}\colon H(\Omega)\times H(\Omega)\to H(\Omega)',\ \mathrm{B}(u,v)[\varphi]:=\mathit{Re}\int_\Omega[(u\cdot\nabla)v]\cdot\varphi\,\mathrm{d}(x,y)\quad\text{for all }\varphi\in H(\Omega). \tag{3.1}$$

Since the inner product is bilinear and the integral is linear directly from the definition, we conclude that B is indeed a bilinear form. Moreover, applying Lemma A.9 (i), we obtain

$$\|\mathrm{B}(u,v)\|_{H(\Omega)'}\le {C_4}^2\mathit{Re}\|u\|_{H_0^1(\Omega,\mathbb{R}^2)}\|v\|_{H_0^1(\Omega,\mathbb{R}^2)}\quad\text{for all }u,v\in H(\Omega)$$

which shows that $B(u,v)$ is a bounded linear functional on $H(\Omega)$ for all $u,v\in H(\Omega)$, i.e., $\mathrm{B}(u,v)\in H(\Omega)'$ for all $u,v\in H(\Omega)$ implying that B is well-defined a posteriori. Using matrix vector multiplication (cf. (A.6)), we can reformulate the definition of B and obtain

$$\mathrm{B}(u,v)[\varphi]=\mathit{Re}\int_\Omega\varphi^T(\nabla v)u\,\mathrm{d}(x,y)\quad\text{for all }u,v,\varphi\in H(\Omega).$$

Next, we are going to treat the linear part of our weak formulation similarly. Therefore, we first observe that the Poiseuille flow U (see (1.10)) and therefore Γ (see (1.12)) are not square integrable on Ω and thus not contained in our solution space $H(\Omega)$. Hence, a suitable space for the Poiseuille flow and Γ have to be found. Since U and V introduced in Chapter 1 are bounded, the right space is

$$W(\Omega):=\left\{u\in W^{1,\infty}(\Omega,\mathbb{R}^2)\colon\ \operatorname{div}u=0\right\}$$

which implies $U\in W(\Omega)$, as well as $\Gamma\in W(\Omega)$. Now, for arbitrary $w\in W(\Omega)$ we can define

$$\mathrm{B}_w\colon H(\Omega)\to L^2(\Omega,\mathbb{R}^2),\ \mathrm{B}_w\,u:=\mathit{Re}\,[(u\cdot\nabla)w+(w\cdot\nabla)u]\,. \tag{3.2}$$

Since the norms $\|w\|_{L^\infty(\Omega,\mathbb{R}^2)}$ and $\|\nabla w\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})}$ are finite for any $w\in W(\Omega)$ we directly obtain $\mathrm{B}_w\,u\in L^2(\Omega,\mathbb{R}^2)$ for all $u\in H(\Omega)$, i.e., B_w is well-defined. Using the linearity of the derivative and the bilinearity of the inner product we directly obtain that B_w is a linear operator.

Similar as above, we obtain the following alternative representation of B_w using matrix vector multiplication:

$$\mathrm{B}_w\,u=\mathit{Re}\,[(\nabla w)u+(\nabla u)w]\quad\text{for all }u\in H(\Omega), \tag{3.3}$$

i.e., $(\mathrm{B}_w\,u)[\varphi]=\mathit{Re}\int_\Omega\varphi^T\,[(\nabla w)u+(\nabla u)w]\,\mathrm{d}(x,y)$ for all $u,\varphi\in H(\Omega)$.

We note that by definition (2.13) $\mathrm{B}_w\,u\in L^2(\Omega,\mathbb{R}^2)$ actually defines a bounded linear functional on $H(\Omega)$ for all $u\in H(\Omega)$, i.e., as before the "symbol" $\mathrm{B}_w\,u$ can be interpreted as an element of $H(\Omega)'$. Again, from the context it becomes clear whether $\mathrm{B}_w\,u$ has to be read as square integrable function or needs to be understood as a bounded linear functional on $H(\Omega)$. Furthermore, from the theoretical setting in Section 2.2 we get the estimate $\|\mathrm{B}_w u\|_{H(\Omega)'} \leq C_2 \|\mathrm{B}_w u\|_{L^2(\Omega,\mathbb{R}^2)}$ holding true for all $u \in H(\Omega)$.

Applying Lemma A.10 (ii) we obtain

$$(\mathrm{B}_w u)[\varphi] = Re \int_\Omega \left(-w^T(\nabla\varphi)u + \varphi^T(\nabla u)w\right) \,\mathrm{d}(x,y) \quad \text{for all } u,\varphi \in H(\Omega)$$

which together with Lemma A.9 (iii) implies the following alternative estimate

$$\|\mathrm{B}_w u\|_{H(\Omega)'} \leq 2C_2 Re \|w\|_{L^\infty(\Omega,\mathbb{R}^2)} \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \tag{3.4}$$

for the norm of $\mathrm{B}_w u$ for any $u \in H(\Omega)$.

The linearity of the integral and the bilinearity of the inner product together with direct calculations show the following rules for transforming terms containing the operator B_w (see (3.2)) and the bilinear form B (see (3.1)):

Proposition 3.1. *The following assertions hold true:*

(i) $\mathrm{B}_v + \mathrm{B}_w = \mathrm{B}_{v+w}$ *for all* $v, w \in W(\Omega)$.

(ii) $\mathrm{B}_{-w} = -\mathrm{B}_w$ *for all* $w \in W(\Omega)$.

(iii) $\mathrm{B}_w u = \mathrm{B}(u,w) + \mathrm{B}(w,u)$ *for all* $w \in H(\Omega) \cap W(\Omega)$ *and* $u \in H(\Omega)$.

Since the proof of Proposition 3.1 consists only of simple calculations we postpone it to the Appendix (see Proof of Proposition 3.1 in Appendix A.3).

Finally, using the bounded operators B and B_w defined in (3.1) and (3.2) as well as the functional given by the right-hand side (cf. (1.14) and (2.13)), we are in a position to reformulate (1.15) as a zero-finding problem using the following operator

$$\mathrm{F}\colon H(\Omega) \to H(\Omega)', \ \mathrm{F}\,u := -\Delta u + \mathrm{B}(u,u) + \mathrm{B}_\Gamma u - g. \tag{3.5}$$

Directly from the definitions (3.1), (3.2) and (1.14) it follows that $u \in H(\Omega)$ solves our weak formulation of the Navier-Stokes equations (1.15) if and only if $\mathrm{F}\,u = 0$.

Since B is bilinear and B_w is linear for all $w \in W(\Omega)$ the zero-finding operator F is Fréchet differentiable with derivative

$$\mathrm{F}'\,\tilde{u}\colon H(\Omega) \to H(\Omega)', \ (\mathrm{F}'\,\tilde{u})u = -\Delta u + \mathrm{B}(u,\tilde{u}) + \mathrm{B}(\tilde{u},u) + \mathrm{B}_\Gamma u$$

at any point $\tilde{u} \in H(\Omega)$. If we consider this Fréchet derivative evaluated at a point $\tilde{u} \in H(\Omega) \cap W(\Omega)$ (which will be the case in the following Sections), Proposition 3.1 (iii) and (i) yield the equivalent formulation

$$(\mathrm{F}'\,\tilde{u})u = -\Delta u + \mathrm{B}_{\tilde{u}+\Gamma} u \quad \text{for all } u \in H(\Omega). \tag{3.6}$$

Remark 3.2. *We note that all integrals which are of the form as in the definitions of the operators* B *and* B_w *respectively are well-defined due to Lemma A.9.*

3.2 Existence and Enclosure Theorem

One crucial component for a successful computer-assisted proof is the computation of an accurate approximate solution to the problem under consideration. To obtain the approximate solution in principle every numerical algorithm can be used if it provides an approximation which belongs exactly to the domain of the zero-finding operator F, i.e., in our applications we require a numerical algorithm providing an exactly divergence-free solution in $H(\Omega)$. In all our examples we use divergence-free finite element methods to compute an approximate solution to our problem (1.15), or $\mathrm{F}\,u = 0$ respectively. For details about the computation procedure we refer the reader to Chapter 4.

In the further course of this Chapter, we suppose that we have an approximate solution $\tilde{\omega} \in H(\Omega) \cap W(\Omega)$ to $\mathrm{F}\,u = 0$ in hand explicitly. Moreover, we assume that $\tilde{\omega}$ is of the following form

$$\tilde{\omega} = \begin{cases} \tilde{\omega}_0, & \text{in } \Omega_0, \\ 0, & \text{in } \Omega \setminus \Omega_0, \end{cases} \tag{3.7}$$

for some suitable (usually bounded) computational domain $\Omega_0 \subseteq \Omega$ (cf. Figure 3.1) and

$$\tilde{\omega}_0 \in H(\Omega_0) := \left\{ u \in H_0^1(\Omega_0, \mathbb{R}^2) \colon \operatorname{div} u = 0 \right\}.$$

At this stage we want to emphasize that the Dirichlet boundary condition implies continuity of the normal component and thus, by construction $\tilde{\omega}$ is divergence-free on our entire domain Ω, i.e., the crucial assumption $\tilde{\omega} \in H(\Omega)$ is satisfied for approximate solutions of the form (3.7).

Moreover, we note that the structure assumed in (3.7) in most of the applications of computer-assisted proofs is not a restriction on the numerical algorithm used to compute the approximate solution, since most of the common methods yield a compactly supported approximate solution $\tilde{\omega} \in H(\Omega) \cap W(\Omega)$ anyway. However, the numerical algorithm has to provide an approximation which needs to be exactly divergence-free. Again, in Chapter 4 we describe one possibility how such an approximation procedure can be realized.

If we have computed some approximate solution $\tilde{\omega}$ to problem (1.15),

$$\omega := \tilde{\omega} - V \tag{3.8}$$

is an approximation for the velocity function in problem (1.11) which satisfies non-zero boundary conditions at the boundary of the obstacle. Furthermore, reversing the first transformation in Section 1.2 as well yields the approximation $U + \omega = \Gamma + \tilde{\omega}$ to the velocity function in the original Navier-Stokes equations (1.7).

Having an approximate solution $\tilde{\omega}$ in hand, we denote the linearization of F at $\tilde{\omega}$ by $\mathrm{L}_{U+\omega}$. Since $\tilde{\omega} \in H(\Omega) \cap W(\Omega)$, the Fréchet derivative formulated in (3.6) yields

$$\mathrm{L}_{U+\omega} \colon H(\Omega) \to H(\Omega)', \; \mathrm{L}_{U+\omega}\, u := \mathrm{F}'(\tilde{\omega})u = -\Delta u + \mathrm{B}_{\tilde{\omega}+\Gamma}\, u = -\Delta u + \mathrm{B}_{U+\omega}\, u, \tag{3.9}$$

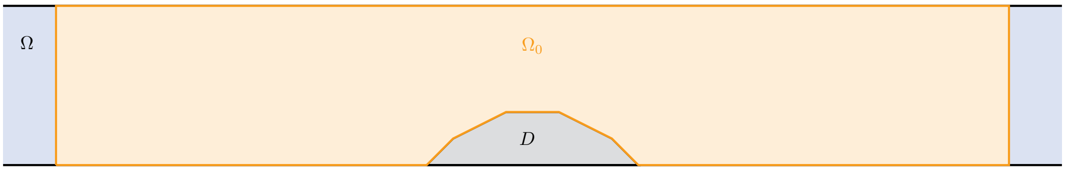


Figure 3.1: Example domain with obstacle and computational domain Ω_0

where we used (3.8) and the definition of Γ (see (1.12)) to calculate $\tilde{\omega} + \Gamma = U + \omega$ which justifies the index $U + \omega$ for the linearization operator $\mathrm{L}_{U+\omega}$ a posteriori.

Similar to the definition of the operator B_w for $w \in W(\Omega)$ we extend the definition of the linearization above (see (3.9)) to an operator L_w for arbitrary functions $w \in W(\Omega)$, i.e., we define the operator

$$\mathrm{L}_w \colon H(\Omega) \to H(\Omega)', \, \mathrm{L}_w \, u \coloneqq -\Delta u + \mathrm{B}_w \, u \tag{3.10}$$

which coincides with the linearization of F at the approximate solution $\tilde{\omega}$ for $w = U + \omega$.

Beside an accurate approximate solution $\tilde{\omega}$ the computer-assisted techniques used in this thesis require the following crucial assumptions (cf. [74, Section 6.1]):

(A1) Suppose a bound $\delta \geq 0$ for the defect (residual) of $\tilde{\omega}$ has been computed, i.e., the following estimate holds true

$$\|\mathrm{F}\, \tilde{\omega}\|_{H(\Omega)'} \leq \delta.$$

(A2) Assume a constant $K > 0$ is in hand such that

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq K \|\mathrm{L}_{U+\omega}\, u\|_{H(\Omega)'} = K \|\Phi^{-1}\, \mathrm{L}_{U+\omega}\, u\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega)$$

with $\mathrm{L}_{U+\omega}$ defined in (3.9). Recall that the last equality is a direct consequence of the definition of the isometric isomorphism Φ (cf. (2.14)).

We note that K satisfying (A2) is actually a norm bound for the inverse of $\mathrm{L}_{U+\omega}$. Moreover, assumption (A2) directly implies that $\mathrm{L}_{U+\omega}$ is one-to-one.

In the proof of the Existence Theorem 3.4 we need that $\mathrm{L}_{U+\omega}$ is not only one-to-one but additionally onto (see proof of Theorem 3.4 or [85, Section 2]). To prove the surjectivity of $\mathrm{L}_{U+\omega}$ the authors of most of the recent works using the same computer-assisted techniques exploited the fact that the operator $\Phi^{-1}\, \mathrm{L}_{U+\omega}$ is symmetric, and hence self-adjoint, which is not the case in our situation (see beginning of Chapter 6).

Nevertheless, similar as in recent works, we can use (A2) to show that the range of $\Phi^{-1}\, \mathrm{L}_{U+\omega}$ is closed.

To prove this assertion let $(u_n)_{n\in\mathbb{N}}$ be a sequence in $H(\Omega)$ and $w \in H(\Omega)$ such that $\Phi^{-1}\, \mathrm{L}_{U+\omega}\, u_n \to w$ as $n \to \infty$ in $H(\Omega)$. Thus, $(\Phi^{-1}\, \mathrm{L}_{U+\omega}\, u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $H(\Omega)$. Applying (A2) and the linearity of $\Phi^{-1}\, \mathrm{L}_{U+\omega}$ we obtain

$$\|u_n - u_m\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq K \|\Phi^{-1}\, \mathrm{L}_{U+\omega}\, u_n - \Phi^{-1}\, \mathrm{L}_{U+\omega}\, u_m\|_{H_0^1(\Omega,\mathbb{R}^2)} \to 0 \quad \text{as } n, m \to \infty.$$

Hence, $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $H(\Omega)$ and since $H(\Omega)$ is a Hilbert space (see Section 2.1) there exists some $u \in H(\Omega)$ such that $u_n \to u$ as $n \to \infty$ in $H(\Omega)$. Furthermore, using the fact that $\mathrm{L}_{U+\omega}$ is bounded we obtain

$$\|\Phi^{-1}\, \mathrm{L}_{U+\omega}\, u_n - \Phi^{-1}\, \mathrm{L}_{U+\omega}\, u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \|\mathrm{L}_{U+\omega}\|_{\mathcal{B}} \|u_n - u\|_{H_0^1(\Omega,\mathbb{R}^2)} \to 0 \quad \text{as } n \to \infty$$

and thus $(\Phi^{-1}\, \mathrm{L}_{U+\omega}\, u_n)_{n\in\mathbb{N}}$ converges to $\Phi^{-1}\, \mathrm{L}_{U+\omega}\, u$ as $n \to \infty$ in $H(\Omega)$. Hence, we showed $w = \Phi^{-1}\, \mathrm{L}_{U+\omega}\, u \in \mathrm{rg}(\Phi^{-1}\, \mathrm{L}_{U+\omega})$, i.e., the range of $\Phi^{-1}\, \mathrm{L}_{U+\omega}$ is closed.

Next, since $\operatorname{rg}(\Phi^{-1}\,\mathrm{L}_{U+\omega})$ is closed and Φ is bijective, the surjectivity of $\mathrm{L}_{U+\omega}$ follows if we can show that the range of $\Phi^{-1}\,\mathrm{L}_{U+\omega}$ is dense in $H(\Omega)$ (and thus coincides with the entire space $H(\Omega)$).

Again, we note that the density of the range would be a direct consequence if the operator $\Phi^{-1}\,\mathrm{L}_{U+\omega}$ would be self-adjoint (see for instance [89]). In our considerations we have a closer look at the adjoint operator $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\colon H(\Omega) \to H(\Omega)$ to overcome the lack of self-adjointness (cf. [74, Section 9.4.1.2]). Exploiting the equality

$$\operatorname{rg}(\Phi^{-1}\,\mathrm{L}_{U+\omega})^{\perp} = \ker((\Phi^{-1}\,\mathrm{L}_{U+\omega})^*),$$

it suffices to prove that $\ker((\Phi^{-1}\,\mathrm{L}_{U+\omega})^*) = \{0\}$, or since $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*$ is linear to guarantee that $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*$ is one-to-one.

Therefore, in addition to (A1) and (A2), we make the following assumption for the adjoint operator (cf. [74, Section 9.4.1.2]):

(A3) Let a constant $K^* > 0$ be in hand such that

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq K^* \|(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* u\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega).$$

Then, if (A3) is satisfied, the same arguments as already mentioned for the operator $\Phi^{-1}\,\mathrm{L}_{U+\omega}$ above imply that the adjoint operator $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*$ is one-to-one.

Finally, combining all arguments above, we proved the following Proposition:

Proposition 3.3. *Let constants $K > 0$ and $K^* > 0$ be in hand such that the assumptions (A2) and (A3) are satisfied. Then the linearization $\mathrm{L}_{U+\omega}$ of F at $\tilde{\omega}$ is bijective.*

For the computation of the constants K and K^*, a substantial use of computer-assisted methods is needed. A manner of computing such constants δ, K and K^* will be addressed in Chapter 5 as well as in Sections 6.1 and 6.2. For the further course of this Chapter, we assume that we have already computed the desired constants satisfying assumptions (A1), (A2) and (A3) respectively. Then, we are in a position to formulate the existence and enclosure theorem (cf. [85, Theorem 1 (p. 25)]) for our Navier-Stokes equations (1.15).

Theorem 3.4. *Let $\tilde{\omega} \in H(\Omega) \cap W(\Omega)$ be an approximate solution of* (1.15) *and constants $\delta \geq 0$ and $K, K^* > 0$ be computed satisfying the assumptions (A1), (A2) and (A3) respectively. If*

$$4K^2 {C_4}^2 Re\, \delta < 1, \tag{3.11}$$

then there exists a locally unique solution $u^* \in H(\Omega)$ of (1.15) *satisfying the error enclosure*

$$\|u^* - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \frac{2K\delta}{1 + \sqrt{1 - 4K^2 {C_4}^2 Re\, \delta}}.$$

Proof. The proof is an application of Banach's Fixed-point theorem and follows the lines of Plum presented for instance in [85, Proof of Theorem 1 (p. 25)]. To improve the readability of the proof in the following, we just write L instead of $\mathrm{L}_{U+\omega}$. In a first step, we equivalently rewrite the zero finding problem $\mathrm{F}\,u = 0$ into a fixed-point equation. Therefore, introducing the error $v := u - \tilde{\omega}$ we obtain

$$\mathrm{L}\,v = -\,\mathrm{F}\,\tilde{\omega} - (\mathrm{F}(\tilde{\omega} + v) - \mathrm{F}\,\tilde{\omega} - \mathrm{L}\,v)$$

or, due to the bijectivity of L, which is provided by Proposition 3.3 using our assumptions (A2) and (A3), we obtain the equivalent fixed-point formulation

$$v = -\mathrm{L}^{-1}(\mathrm{F}\,\tilde{\omega} + (\mathrm{F}(\tilde{\omega} + v) - \mathrm{F}\,\tilde{\omega} - \mathrm{L}\,v)) =: \mathrm{T}\,v,$$

where the right-hand side defines a fixed-point operator $\mathrm{T}\colon H(\Omega) \to H(\Omega)$ which will be considered in the further course.

Applying the definition of F (see (3.5)) and L (see (3.9)) respectively and using the linearity of B_Γ we obtain

$$\begin{aligned}
&\mathrm{F}(\tilde{\omega} + v) - \mathrm{F}(\tilde{\omega} + w) - \mathrm{L}(v - w)\\
&\quad = -\Delta(\tilde{\omega} + v) + \mathrm{B}(\tilde{\omega} + v, \tilde{\omega} + v) + \mathrm{B}_\Gamma(\tilde{\omega} + v) - g\\
&\qquad - (-\Delta(\tilde{\omega} + w) + \mathrm{B}(\tilde{\omega} + w, \tilde{\omega} + w) + \mathrm{B}_\Gamma(\tilde{\omega} + w) - g)\\
&\qquad - (-\Delta(v - w) + \mathrm{B}_{U+\omega}(v - w))\\
&\quad = \mathrm{B}(\tilde{\omega} + v, \tilde{\omega} + v) - \mathrm{B}(\tilde{\omega} + w, \tilde{\omega} + w) + \mathrm{B}_\Gamma(v - w) - \mathrm{B}_{U+\omega}(v - w)
\end{aligned}$$

for all $v, w \in H(\Omega)$. Moreover, Proposition 3.1 (ii) and (i) imply

$$\mathrm{B}_\Gamma(v - w) - \mathrm{B}_{U+\omega}(v - w) = -\,\mathrm{B}_{-\Gamma}(v - w) - \mathrm{B}_{\Gamma+\tilde{\omega}}(v - w) = -\,\mathrm{B}_{\tilde{\omega}}(v - w)$$

for all $v, w \in H(\Omega)$. Hence, using the mean value theorem, the bilinearity of B as well as Proposition 3.1 (ii) and (iii), we get

$$\begin{aligned}
&\mathrm{F}(\tilde{\omega} + v) - \mathrm{F}(\tilde{\omega} + w) - \mathrm{L}(v - w)\\
&\quad = \mathrm{B}(\tilde{\omega} + v, \tilde{\omega} + v) - \mathrm{B}(\tilde{\omega} + w, \tilde{\omega} + w) - \mathrm{B}_{\tilde{\omega}}(v - w)\\
&\quad = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{B}(\tilde{\omega} + tv + (1-t)w, \tilde{\omega} + tv + (1-t)w) - t\,\mathrm{B}_{\tilde{\omega}}(v - w)\right]\,\mathrm{d}t\\
&\quad = \int_0^1 \left[\mathrm{B}(v - w, \tilde{\omega} + tv + (1-t)w) + \mathrm{B}(\tilde{\omega} + tv + (1-t)w, v - w) - \mathrm{B}_{\tilde{\omega}}(v - w)\right]\,\mathrm{d}t\\
&\quad = \int_0^1 \left[\mathrm{B}(v - w, tv + (1-t)w) + \mathrm{B}(tv + (1-t)w, v - w)\right]\,\mathrm{d}t.
\end{aligned}$$

Taking the dual norm and applying Lemma A.9 (i) we obtain

$$\begin{aligned}
&\| \mathrm{F}(\tilde{\omega}+v) - \mathrm{F}(\tilde{\omega}+w) - \mathrm{L}(v-w)\|_{H(\Omega)'} \\
&\quad = \left\| \int_0^1 \left[\mathrm{B}(v-w, tv+(1-t)w) + \mathrm{B}(tv+(1-t)w, v-w)\right] \mathrm{d}t \right\|_{H(\Omega)'} \\
&\quad \le \sup_{\substack{\varphi\in H(\Omega)\\ \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}=1}} \int_0^1 \left[|\mathrm{B}(v-w, tv+(1-t)w)[\varphi]| + |\mathrm{B}(tv+(1-t)w, v-w)[\varphi]|\right] \mathrm{d}t \\
&\quad \le 2{C_4}^2 Re \sup_{\substack{\varphi\in H(\Omega)\\ \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}=1}} \int_0^1 \|v-w\|_{H_0^1(\Omega,\mathbb{R}^2)} \|tv+(1-t)w\|_{H_0^1(\Omega,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)} \,\mathrm{d}t \\
&\quad \le 2{C_4}^2 Re \|v-w\|_{H_0^1(\Omega,\mathbb{R}^2)} \int_0^1 \left[t\|v\|_{H_0^1(\Omega,\mathbb{R}^2)} + (1-t)\|w\|_{H_0^1(\Omega,\mathbb{R}^2)}\right] \mathrm{d}t \\
&\quad = {C_4}^2 Re \left(\|v\|_{H_0^1(\Omega,\mathbb{R}^2)} + \|w\|_{H_0^1(\Omega,\mathbb{R}^2)}\right) \|v-w\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } v,w \in H(\Omega).
\end{aligned} \tag{3.12}$$

Now, we define

$$\alpha := \frac{2K\delta}{1+\sqrt{1-4K^2{C_4}^2 Re\,\delta}} = \frac{1}{2K{C_4}^2 Re}\left(1-\sqrt{1-4K^2{C_4}^2 Re\,\delta}\right)$$

and consider

$$\mathcal{V} := \{v \in H(\Omega) \colon \|v\|_{H_0^1(\Omega,\mathbb{R}^2)} \le \alpha\}. \tag{3.13}$$

Assumption (A2) implies

$$\begin{aligned}
\|\mathrm{T}\, v\|_{H_0^1(\Omega,\mathbb{R}^2)} &= \|{-\mathrm{L}^{-1}} \left(\mathrm{F}\,\tilde{\omega} + (\mathrm{F}(\tilde{\omega}+v) - \mathrm{F}\,\tilde{\omega} - \mathrm{L}\, v)\right)\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\le K\|\mathrm{F}\,\tilde{\omega} + (\mathrm{F}(\tilde{\omega}+v) - \mathrm{F}\,\tilde{\omega} - \mathrm{L}\, v)\|_{H(\Omega)'} \\
&\le K\left(\|\mathrm{F}\,\tilde{\omega}\|_{H(\Omega)'} + \|\mathrm{F}(\tilde{\omega}+v) - \mathrm{F}\,\tilde{\omega} - \mathrm{L}\, v\|_{H(\Omega)'}\right)
\end{aligned}$$

for all $v \in H(\Omega)$. Using (3.12) with $w=0$ we obtain

$$\|\mathrm{T}\, v\|_{H_0^1(\Omega,\mathbb{R}^2)} \le K\left(\|\mathrm{F}\,\tilde{\omega}\|_{H(\Omega)'} + {C_4}^2 Re \|v\|^2_{H_0^1(\Omega,\mathbb{R}^2)}\right) \quad \text{for all } v \in H(\Omega)$$

and thus, applying assumptions (A1) and (3.13), as well as the definition of α we compute

$$\begin{aligned}
\|\mathrm{T}\, v\|_{H_0^1(\Omega,\mathbb{R}^2)} &\le K\left(\delta + {C_4}^2 Re \|v\|^2_{H_0^1(\Omega,\mathbb{R}^2)}\right) \le K\left(\delta + {C_4}^2 Re\, \alpha^2\right) \\
&= K\left(\delta + \frac{1}{4K^2{C_4}^2 Re}\left(1-\sqrt{1-4K^2{C_4}^2 Re\,\delta}\right)^2\right) \\
&= \frac{1}{2K{C_4}^2 Re}\left(1-\sqrt{1-4K^2{C_4}^2 Re\,\delta}\right) = \alpha \quad \text{for all } v \in \mathcal{V}.
\end{aligned} \tag{3.14}$$

Thus, $\mathrm{T}\, v \in \mathcal{V}$ for all $v \in \mathcal{V}$, i.e., T maps $\mathcal{V}$ into itself.

In almost the same manner (A2), (3.12) and (3.13) imply

$$\begin{aligned}
\|\mathrm{T}\, v - \mathrm{T}\, w\|_{H_0^1(\Omega,\mathbb{R}^2)} &= \|{-\mathrm{L}^{-1}} \left(\mathrm{F}(\tilde{\omega}+v) - \mathrm{F}(\tilde{\omega}+w) - \mathrm{L}(v-w)\right)\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\le K\|\mathrm{F}(\tilde{\omega}+v) - \mathrm{F}(\tilde{\omega}+w) - \mathrm{L}(v-w)\|_{H(\Omega)'} \\
&\le {C_4}^2 Re\, K \left(\|v\|_{H_0^1(\Omega,\mathbb{R}^2)} + \|w\|_{H_0^1(\Omega,\mathbb{R}^2)}\right) \|v-w\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\le 2{C_4}^2 Re\, K\, \alpha \|v-w\|_{H_0^1(\Omega,\mathbb{R}^2)}
\end{aligned}$$

for all $v, w \in \mathcal{V}$. Moreover, using the definition of α and assumption (3.11) we calculate

$$2{C_4}^2 Re\, K\, \alpha = 1 - \sqrt{1 - 4K^2{C_4}^2 Re\, \delta} < 1$$

which proves that T is a contraction on $\mathcal{V}$.

Thus, Banach's Fixed-point theorem yields the existence of a locally unique fixed-point $v^* \in \mathcal{V}$ of T. Hence, by $u^* := v^* + \tilde{\omega}$ we obtain a locally unique solution of $\mathrm{F}\, u = 0$ satisfying the error estimate $\|u^* - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} = \|v^*\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \alpha$ which finishes the proof. □

Corollary 3.5. *If Theorem 3.4 is successful and thus, provides the existence of an exact solution u^*, the embeddings* (2.12) *directly imply the following enclosure results for the solution:*

$$\|u^* - \tilde{\omega}\|_{L^p(\Omega,\mathbb{R}^2)} \leq C_p \|u^* - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \frac{2K\delta C_p}{1 + \sqrt{1 - 4K^2{C_4}^2 Re\, \delta}} \quad \text{for all } p \in [2, \infty).$$

Remark 3.6. (i) *The constant K^* has no influence on the value of the defect bound provided by Theorem 3.4. Hence, in contrast to K we do not require a "moderate" constant K^* satisfying assumption (A3), i.e., any constant $K^* > 0$ (in hand) is suitable for the application of Theorem 3.4.*

(ii) *Having proved the existence of a solution u^* of the Navier-Stokes equations* (1.15) *(together with the results in Chapter 7) and using the transformations in Chapter 1, we obtain the existence of a solution $\Gamma + u^*$ to the original problem* (1.7).

(iii) *Assumption* (3.11) *is a direct demand on the accuracy of the approximate solution used in Theorem 3.4 via the smallness of the defect bound δ, which transfers the crucial work to the computer.*

Extending the Region of Local Uniqueness

In Theorem 3.4 we are interested in an error bound for the exact solution (measured in the $H(\Omega)$-norm) which is as small as possible. However, this results in a "very small" region of local uniqueness since the radius of the ball $\mathcal{V}$, on which the fixed-point operator T is a self-mapping and a contraction, considered in the proof of Theorem 3.4 is chosen "minimal".

But if the application of Theorem 3.4 is successful, i.e., if we are able to validate the crucial inequality $4K^2{C_4}^2 Re\, \delta < 1$, we try to "slightly enlarge" the region of local uniqueness. Therefore, we have a closer look at the crucial inequality $\delta \leq \frac{\alpha}{K} - G(\alpha)$ appearing in the abstract setting for computer-assisted proofs (cf. [85, Theorem 1]), where the function G is defined by $G(t) := {C_4}^2 Re\, t^2$ for all $t \in [0, \infty)$ in the Navier-Stokes case (cf. proof of Theorem 3.4). Then, we consider the function

$$h\colon [0, \infty) \to \mathbb{R},\ h(t) := \frac{t}{K} - G(t) = \frac{t}{K} - {C_4}^2 Re\, t^2 \tag{3.15}$$

and by direct calculations we get the following value for its maximum (cf. Figure 3.2):

$$\delta_{\max} := \max\{h(t)\colon t \in [0, \infty)\} = h\left(\frac{1}{2K{C_4}^2 Re}\right) = \frac{1}{4K^2{C_4}^2 Re}. \tag{3.16}$$

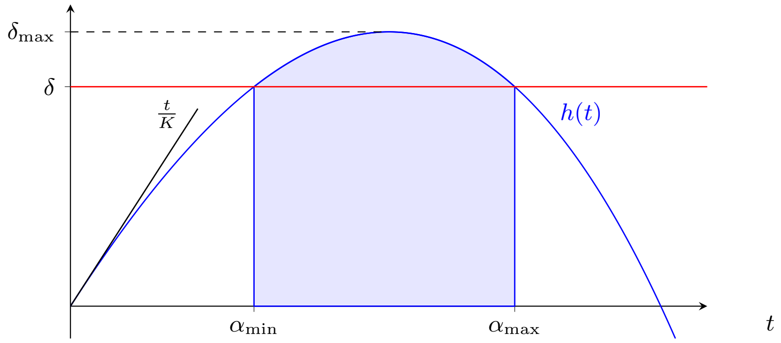


Figure 3.2: Possible range for the error bound α

To compute the range of possible values $[\alpha_{\min}, \alpha_{\max}]$ satisfying the crucial abstract inequality $\delta \le \frac{\alpha}{K} - G(\alpha)$, we have to solve the equation $h(t) = \delta$ (cf. Figure 3.2). We note that this equations is solvable since $4K^2{C_4}^2 Re\,\delta < 1$ holds true (since we supposed that our proof is successful) and thus, $\delta \le \delta_{\max}$ is satisfied (cf. (3.16)). Directly from the definition of h, we obtain

$$h(t) = \delta \quad \Leftrightarrow \quad t \in \left\{ \frac{1}{2K{C_4}^2 Re} \left(1 \pm \sqrt{1 - 4K^2{C_4}^2 Re\delta} \right) \right\}$$

implying

$$\alpha_{\min} = \frac{1}{2K{C_4}^2 Re} \left(1 - \sqrt{1 - 4K^2{C_4}^2 Re\delta} \right) = \frac{2K\delta}{1 + \sqrt{1 - 4K^2{C_4}^2 Re\,\delta}},$$
$$\alpha_{\max} = \frac{1}{2K{C_4}^2 Re} \left(1 + \sqrt{1 - 4K^2{C_4}^2 Re\delta} \right) = \frac{2K\delta}{1 - \sqrt{1 - 4K^2{C_4}^2 Re\,\delta}}.$$

Having a closer look at Theorem 3.4 we see that $\alpha_{\min}$ actually coincides with the error bound therein, i.e., in our proof of Theorem 3.4 for the set $\mathcal{V}$ we actually use the minimal radius of uniqueness with respect to the constants δ and K.

Moreover, since h is concave (note that $h''(t) = -2{C_4}^2 Re < 0$ for all $t \in [0, \infty)$) we obtain

$$\delta < \frac{\alpha}{K} - {C_4}^2 Re\,\alpha^2 \quad \text{for all } \alpha \in (\alpha_{\min}, \alpha_{\max}). \tag{3.17}$$

In view of the uniqueness result presented in [74, Theorem 6.2] we can prove the following Theorem for the radius of uniqueness in the context of our Navier-Stokes equations.

Theorem 3.7. *Let the assumptions of Theorem 3.4 be satisfied, i.e., Theorem 3.4 provides the existence of a solution $u^* \in H(\Omega)$ of (1.15) with*

$$\|u^* - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \le \frac{2K\delta}{1 + \sqrt{1 - 4K^2{C_4}^2 Re\,\delta}} = \alpha_{min}.$$

Then u^* is a locally unique solution of (1.15) in

$$\{u \in H(\Omega) \colon \|u - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \le \alpha_0\} \quad \textit{for all } \alpha_0 \in (\alpha_{min}, \alpha_{max}).$$

Proof. Again, we follow the lines of Plum in [74, Proof of Theorem 6.2]. Therefore, let $\alpha_0 \in (\alpha_{\min}, \alpha_{\max})$ and u_0 be a second solution of our Navier-Stokes equations (1.15) with $\|u_0 - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \alpha_0$. Then, since u_0 is a solution, by construction $v_0 := u_0 - \tilde{\omega}$ is a fixed-point of the operator T introduced in the proof of Theorem 3.4. Analogously to the first inequality in (3.14) we calculate

$$\|v_0\|_{H_0^1(\Omega,\mathbb{R}^2)} = \|Tv_0\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq K\left(\delta + {C_4}^2 Re\,\|v_0\|^2_{H_0^1(\Omega,\mathbb{R}^2)}\right)$$

which directly yields

$$h(\|v_0\|_{H_0^1(\Omega,\mathbb{R}^2)}) = \frac{\|v_0\|_{H_0^1(\Omega,\mathbb{R}^2)}}{K} - {C_4}^2 Re\,\|v_0\|^2_{H_0^1(\Omega,\mathbb{R}^2)} \leq \delta, \tag{3.18}$$

where h is defined as in (3.15).

Assuming $\|v_0\|_{H_0^1(\Omega,\mathbb{R}^2)} > \alpha_{\min}$, (3.17) would imply $h(\|v_0\|_{H_0^1(\Omega,\mathbb{R}^2)}) > \delta$ (recall that $\|v_0\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \alpha_0 < \alpha_{\max}$) which is a contradiction to (3.18). Hence, we conclude that $\|v_0\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \alpha_{\min}$ which shows $v_0 \in \mathcal{V}$ with $\mathcal{V}$ defined in (3.13). Since the fixed-point $v^* \in \mathcal{V}$ provided by Banach's Fixed-point theorem in the proof of Theorem 3.4 is locally unique in $\mathcal{V}$ we conclude $v_0 = v^*$ which directly implies $u_0 = v_0 + \tilde{\omega} = v^* + \tilde{\omega} = u^*$ and the assertion follows. □

3.3 Interval Arithmetic

Since we are interested in an analytic proof for the Navier-Stokes equations, our crucial assumption (3.11) in Theorem 3.4 has to be checked rigorously. Moreover, for the computation of the constants δ, K and K^* as well as for the validation of inequalities with the computer, interval arithmetic calculations are required in order to take rounding errors into account.

We start this Section with a brief introduction to interval arithmetic on $\mathbb{R}$ which is based on ideas of Moore in [65]. For a more detailed description we refer the reader to the literature like the book of Alefeld and Herzberger [3].

In the following, we denote the set of real intervals with $[\mathbb{R}]$. For two real intervals $[\underline{a}, \overline{a}], [\underline{b}, \overline{b}] \in [\mathbb{R}]$ we define the four basic arithmetic operations $\oplus, \ominus, \odot, \oplus\!\!\!\!\div$ element wise, i.e., for $* \in \{+, -, \cdot, \div\}$ we set

$$[\underline{a}, \overline{a}] \circledast [\underline{b}, \overline{b}] := \left\{a * b \colon a \in [\underline{a}, \overline{a}], b \in [\underline{b}, \overline{b}]\right\}, \tag{3.19}$$

where we assume $0 \notin [\underline{b}, \overline{b}]$ if $* = \div$. Since $* \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $* \in \{+, -, \cdot, \div\}$ defines a continuous function on the compact set $[\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}] \subseteq \mathbb{R}^2$ the resulting enclosure on the right-hand side of (3.19) is again a real interval, i.e., it is contained in $[\mathbb{R}]$. Moreover, we get the following computation formulas (cf. [3, p. 2]):

- $[\underline{a}, \overline{a}] \oplus [\underline{b}, \overline{b}] = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]$,
- $[\underline{a}, \overline{a}] \ominus [\underline{b}, \overline{b}] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}]$,
- $[\underline{a}, \overline{a}] \odot [\underline{b}, \overline{b}] = [\min\left\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\right\}, \max\left\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\right\}]$,
- $[\underline{a}, \overline{a}] \oplus\!\!\!\!\div [\underline{b}, \overline{b}] = [\underline{a}, \overline{a}] \odot \left[\frac{1}{\overline{b}}, \frac{1}{\underline{b}}\right]$.

We note that interval arithmetic computations are very sensitive with respect to effects of overestimation. Even using simple calculations, for instance multiplications, this effect is visible. To get an impression of the occurring difficulties we consider the interval $X := [-1, 1]$. Computing the enclosure of $X^2 = \{x^2 : x \in X\}$ with the "naive" operation $\odot$ defined above yields the resulting enclosure interval $X \odot X = [-1, 1]$ which is way too large since the square of two real numbers is non-negative. Taking this fact into account, the resulting enclosing interval can be reduced significantly and we obtain the sharper enclosure $[0, 1]$. This small example shows that even in simple applications of interval arithmetic, one has to be careful using naive implementations.

In addition to the algebraic operations presented above, interval arithmetic versions of the standard functions like $\sqrt{\cdot}, \exp, \sin, \cos, \tan, \dots$ are required. Moreover, strategies to solve linear systems and non-linear equations as well as matrix eigenvalue problems rigorously are of interest. We do not want to go into further details here and refer the reader to the literature again.

For our existence theorem we use the computer for the computation of an approximate solution and constants which are crucial for the success of our analytical proof. Moreover, the assumption of Theorem 3.4, i.e., the crucial inequality (3.11), has to be validated on the computer rigorously. Since also today's computers can only handle finitely many (exact) floating-point numbers, in each computation rounding errors occur. To capture these errors, the interval arithmetic ideas above are applied to the set of floating-point numbers $\mathbb{F} \subseteq \mathbb{R}$ instead of the entire space $\mathbb{R}$, where in each operation additional rounding steps have to be performed. Therefore, the upper and lower bounds of the resulting interval are rounded upwards and downwards respectively to finally obtain a rigorous enclosure that is representable on the computer. Note that the IEEE 574 standard for floating-point arithmetic provides all necessary rounding modes which are needed to perform the interval arithmetic operations mentioned above.

Remark 3.8. *Since a computer can represent only finitely many floating-point numbers exactly, we can validate equalities just for a small amount of reals, i.e., in general we cannot prove an equality on the computer rigorously. However, inequalities of two floating-point numbers can easily be checked with the computer. Thus, we can validate the assumptions in our existence theorem rigorously with the computer if the occurring constants are computed with interval arithmetic operations and thus, are rigorously enclosed in intervals with floating-point bounds.*

For interval arithmetic computations on the computer there exist several different libraries. We want to mention only a few of them like the MATLAB-toolbox INTLAB developed by Rump (see [90]) which provides several interval arithmetic algorithms of high accuracy to treat linear systems, matrix eigenvalue problems and other applications.

Since in our applications our programs are written in C++ we use the C-XSC library (see [53] and [43]) which provides the basic interval arithmetic operations and a various amount of standard functions for interval arithmetic computations. Moreover, a large package of sample problems and algorithms are included in C-XSC.

Note that the latest version of C-XSX was released in 2014 and thus only supports the C++14 standard and cannot be used with newer versions of C++. Nevertheless, in our applications we could use the library making small adaptions and introducing a wrapper class for the interval arithmetic data types (cf. [110]). Moreover, we want to mention the MPFI library (based on the MPFR library [21]) by Nathalie Revol and Fabrice Rouillier which supports the latest C++ standard (see [87]).

Using the real interval arithmetic presented above one can introduce complex intervals which in the following will be denoted by $[\mathbb{C}] := [\mathbb{R}] \times [\mathbb{R}]$. Note that already for the complex multiplication and division, additional ideas are needed to avoid huge overestimation in the results. We do not want to go into further details here, nevertheless, we mention that the library C-XSC introduced above also provides an implementation of complex interval arithmetic of high accuracy and several complex standard functions which still is an advantage compared to the MPFI library (although the latest version of C-XSC is somehow obsolete).

4 Computation of an Approximate Solution

As mentioned in the previous Section, our computer-assisted proof heavily depends on the computation of an accurate approximate solution $\tilde{\omega}$, or ω respectively. In this Chapter we present one strategy to obtain such an approximate solution with "sufficiently small" defect (cf. Chapter 5). Before going into further details about the computation we shortly recall the structure of the approximate solution (cf. (3.7)). Thus, for our numerical computations we first fix a computational domain $\Omega_0 \subseteq \Omega$ (cf. Figure 4.1) and compute an approximate solution $\tilde{\omega}_0 \in H(\Omega_0) = \left\{u \in H_0^1(\Omega_0, \mathbb{R}^2) \colon \operatorname{div} u = 0\right\}$ using finite element methods. This approximation procedure using divergence-free finite elements will be discussed later. Then, we obtain our desired approximate solution $\tilde{\omega} \in H(\Omega)$ simply by extending $\tilde{\omega}_0$ by zero outside of computational domain Ω_0, i.e., we set

$$\tilde{\omega} = \begin{cases} \tilde{\omega}_0, & \text{in } \Omega_0, \\ 0, & \text{in } \Omega \setminus \Omega_0. \end{cases}$$

Again, we want to mention that, using the fact $\operatorname{div} \tilde{\omega}_0 = 0$, by construction our approximate solution $\tilde{\omega}$ is divergence-free on the entire domain Ω since the Dirichlet boundary condition implies continuity of the normal component.

We note that all algorithms presented in this Chapter can be realized with usual numerical means, i.e., no interval arithmetic computations are needed in the algorithms below. All rounding and discretization errors occurring during the approximation process are captured within the fixed-point argument used in Theorem 3.4. Thus, the computational time is not increased comparing to "usual" numerical approximation processes.

Since we transformed the original version of the Navier-Stokes equations using the function $V \in H^2(\Omega, \mathbb{R}^2) \cap C^1(\overline{\Omega}, \mathbb{R}^2)$ (cf. Chapter 1) we need to fix a concrete function V for the computation of our desired approximate solution. Thus, in the following we first present a possible choice for V. Afterwards, we explain how to use divergence-free finite elements to compute an approximate solution of our Navier-Stokes equations (1.15).

4.1 Computation of V

Recall that V is introduced to model the boundary condition on the boundary of the obstacle ∂D correctly (cf. Section 1.2) which results in the fact that V needs to coincide

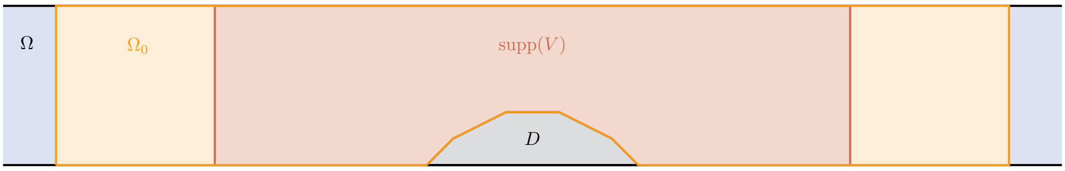


Figure 4.1: Computational domain Ω_0 and supp(V)

with the Poiseuille flow U on ∂D, whereas on the remaining boundary our function V needs to vanish. Moreover, by assumption, V has to be divergence-free with compact support contained within $[-d_0, d_0] \times [0,1]$ for some suitable width $d_0 > d_1$ with the constant d_1 introduced in Chapter 1 to describe the obstacle.

Since the function V is chosen somehow "arbitrary", it is recommended to choose the computational domain Ω_0 large enough such that $\operatorname{supp}(V)$ is contained in Ω_0, i.e., such that $[-d_0, d_0] \times [0,1] \subseteq \overline{\Omega_0}$, to compensate the effects (in the interior of Ω) originating from the "arbitrary" function V by our finite element solution part (cf. Figure 4.1).

To construct the desired function $V \in H^2(\Omega, \mathbb{R}^2) \cap C^1(\overline{\Omega}, \mathbb{R}^2)$ satisfying the assumptions above, we first fix a piecewise polynomial function $\eta \in C^2(\mathbb{R})$. Therefore, we make the ansatz

$$\eta(z) = \begin{cases} 0, & z \in (-\infty, 0), \\ \sum_{i=0}^{5} \alpha_i z^i, & z \in [0,1], \\ 1, & z \in (1, \infty), \end{cases}$$

which leads to the matching conditions

$$\eta(0) = 0, \quad \eta(1) = 1 \quad \text{and} \quad \eta'(0) = \eta''(0) = 0 = \eta'(1) = \eta''(1)$$

at the interfaces to guarantee the desired regularity of V. Using these matching conditions, our ansatz leads to a linear system for the coefficients $\alpha_0, \ldots, \alpha_5 \in \mathbb{R}$ which can easily be solved by direct calculations. Hence, we obtain

$$\eta(z) = \begin{cases} 0, & z \in (-\infty, 0), \\ z^3(6z^2 - 15z + 10), & z \in [0,1], \\ 1, & z \in (1, \infty). \end{cases}$$

Starting from the function η, we are in a position to define a function ϕ which will limit the support of V in x-direction to the interval $[-d_0, d_0]$. Thus, we set

$$\phi\colon \mathbb{R} \to \mathbb{R},\ \phi(x) := \begin{cases} \eta\Big(\frac{d_0+x}{d_0-d_1}\Big), & x \in (-\infty, -d_1], \\ 1, & x \in (-d_1, d_1), \\ \eta\Big(\frac{d_0-x}{d_0-d_1}\Big), & x \in [d_1, \infty). \end{cases}$$

Here, d_1 is the constant introduced in Chapter 1 to describe the expansion of the obstacle in x-direction.

Since we have $\phi(-d_1) = \eta(1) = 1$ and $\phi(d_1) = \eta(1) = 1$ together with the fact that $\eta \in C^2(\mathbb{R})$, it is easy to see that ϕ is also of class $C^2(\mathbb{R})$. Additionally, due to the fact that $\eta(z) = 0$ for all $z \le 0$ and $\eta(1) = 1$, we have

$$\phi(x) = 0 \quad \text{for all } x \in \mathbb{R} \setminus (-d_0, d_0) \quad \text{and} \quad \phi(x) = 1 \quad \text{for all } x \in [-d_1, d_1], \tag{4.1}$$

i.e., the support of ϕ is contained in $[-d_0, d_0]$ (cf. Figure 4.2).

In the following, we have to distinguish the different types of obstacles described in Chapter 1. First, we consider the case of obstacles located at the boundary of the strip. Later on, we discuss the strategy how to deal with obstacles detached from the boundary.

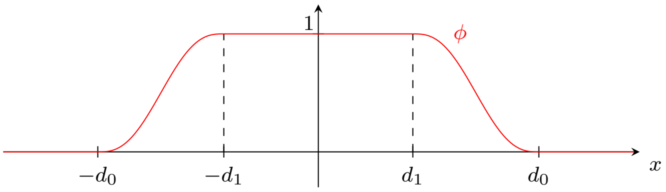


Figure 4.2: Graph of function ϕ

Obstacles Located at the Boundary of S

In this setting, we assume that the obstacle D is contained in the (disconnected) set $[-d_1, d_1] \times ([0, d_2] \cup [d_3, 1])$. We note that in the case of a "single" obstacle at one side of the boundary (without loss of generality we assume that the obstacle is located at the bottom of the strip) we (formally) set $d_3 = 1$ in the calculations below. In this case, the obstacle is actually contained in the single set $[-d_1, d_1] \times [0, d_2]$.

First, we are interested in a function ψ which coincides with $-U_1$ on the set $[0, d_2] \cup [d_3, 1]$ (cf. (1.9)). Therefore, we use the function η again to define an auxiliary function θ as follows

$$\theta\colon [0,1] \to \mathbb{R},\ \theta(y) := \eta\Big(\frac{y-d_2}{d_3-d_2}\Big). \tag{4.2}$$

Obviously, due to the regularity of η our auxiliary function θ is of class C^2 on the interval $[0, 1]$. With θ in hand together with the first component of the Poiseuille flow U_1 (see (1.10)), we define

$$\begin{aligned} \psi\colon [0,1] \to \mathbb{R},\ \psi(y) &:= -\int_0^y U_1(t)\,\mathrm{d}t + \left(\int_0^1 U_1(t)\,\mathrm{d}t\right)\theta(y) \\ &= -\,y^2\left(\frac{1}{2} - \frac{1}{3}y\right) + \frac{1}{6}\theta(y). \end{aligned} \tag{4.3}$$

Due to $\theta \in C^2([0, 1])$, we directly obtain $\psi \in C^2([0, 1])$ and, by definition of θ, we calculate

$$\psi'(y) = -U_1(y) + \frac{1}{6}\theta'(y) \quad \text{for all } y \in [0,1]. \tag{4.4}$$

Moreover, since for all $y \in [0, d_2]$ we calculate $\frac{y-d_2}{d_3-d_2} \leq 0$ and for all $y \in [d_3, 1]$ we obtain $\frac{y-d_2}{d_3-d_2} \geq 1$ (recall that $d_2 < d_3$). Hence, the definition of θ and the properties of η directly yield $\theta'(y) = 0$ for all $y \in [0, d_2] \cup [d_3, 1]$ which implies

$$\psi'(y) = -U_1(y) \quad \text{for all } y \in [0,d_2] \cup [d_3,1] \tag{4.5}$$

(cf. Figure 4.3).

Having both functions ψ and ϕ in hand, we are in a position to define the desired function

$$V(x,y) := \begin{pmatrix} -\phi(x)\psi'(y) \\ \phi'(x)\psi(y) \end{pmatrix} \quad \text{for all } (x,y) \in S \supseteq \Omega. \tag{4.6}$$

In the further course, we check that V defined above satisfies the required assumptions.

Since U_1 is smooth and η is piecewise polynomial, and thus piecewise smooth as well, by construction of V, we directly conclude $V \in H^2(\Omega, \mathbb{R}^2) \cap C^1(\overline{\Omega}, \mathbb{R}^2)$. Furthermore, we observe

$$\operatorname{div} V(x,y) = \frac{\partial(-\phi(x)\psi'(y))}{\partial x} + \frac{\partial(\phi'(x)\psi(y))}{\partial y} = -\phi'(x)\psi'(y) + \phi'(x)\psi'(y) = 0$$

for all $(x,y) \in S \supseteq \Omega$, i.e., V is divergence-free by construction. Moreover, since we have $\frac{d_0+x}{d_0-d_1} \leq 0$ for all $x \in (-\infty, -d_0]$ and $\frac{d_0-x}{d_0-d_1} \leq 0$ for all $x \in [d_0, \infty)$ (recall that $d_0 > d_1$), the definition of ϕ and η respectively, imply $\phi'(x) = 0$ for all $x \in \mathbb{R} \setminus (-d_0, d_0)$. Together with (4.1) we obtain $\operatorname{supp}(V) \subseteq [-d_0, d_0] \times [0,1]$.

In addition to that, (4.1) and (4.5) imply

$$-\phi(x)\psi'(y) = U_1(y) \quad \text{for all } (x,y) \in [-d_1, d_1] \times ([0, d_2] \cup [d_3, 1]) \tag{4.7}$$

and, again by (4.1) together with the definition of ϕ and the identity $\eta'(1) = 0$ we conclude $\phi'(x) = 0$ for all $x \in [-d_1, d_1]$. Hence, we obtain

$$\phi'(x)\psi(y) = 0 \quad \text{for all } (x,y) \in [-d_1, d_1] \times ([0, d_2] \cup [d_3, 1]). \tag{4.8}$$

Combining both identities above and restricting the domains to the parts contained in Ω, we get $V(x,y) = U(x,y)$ for all $(x,y) \in ([-d_1, d_1] \times ([0, d_2] \cup [d_3, 1])) \cap \Omega$.

To show that V satisfies the desired boundary conditions on ∂D, we first recall that we are in the case where the obstacle D is contained in the set $[-d_1, d_1] \times ([0, d_2] \cup [d_3, 1])$ (which in particular holds for the boundary ∂D). Hence, we obtain

$$V(x,y) = U(x,y) \quad \text{for all } (x,y) \in \partial D.$$

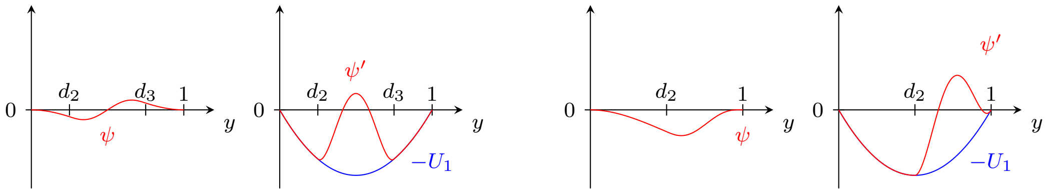
Finally, we have to show that V vanishes on the remaining boundary $\partial\Omega \setminus \partial D$. Using the definition of θ (see (4.2)) we conclude $\theta(0) = 0$ and $\theta(1) = 1$ respectively, hence, inserting this into the definition of V (see (4.3)), we obtain

$$\begin{aligned} \psi(0) &= -\int_0^0 U_1(t)\,\mathrm{d}t + \left(\int_0^1 U_1(t)\,\mathrm{d}t\right)\theta(0) = 0, \\ \psi(1) &= -\int_0^1 U_1(t)\,\mathrm{d}t + \left(\int_0^1 U_1(t)\,\mathrm{d}t\right)\theta(1) = 0. \end{aligned} \tag{4.9}$$

Furthermore, due to (4.5) and the boundary values of the Poiseuille flow we calculate

$$\psi'(0) = -U_1(0) = 0 \quad \text{and} \quad \psi'(1) = -U_1(1) = 0.$$

Altogether, using the definition of V we obtain $V(x,0) = V(x,1) = 0$ for all $x \in \mathbb{R}$ which finally proves the desired boundary condition for V.



(a) Obstacle at both sides of the boundary (b) Obstacle only at the bottom

Figure 4.3: Graph of functions ψ and ψ' if D is located at ∂S

Remark 4.1. *In our applications the accuracy of our approximate solution and thus the success of our proof heavily depends on the difference $d_3 - d_2$. Hence, it is recommendable to chose $d_3 = 1$ whenever it is possible. Additionally, it is recommendable to choose the constant d_2 as small as possible as well to obtain more accurate approximate solutions.*

Obstacles Detached from the Boundary of S

In this case, we suppose that the obstacle D is contained in the set $[-d_1, d_1] \times [d_2, d_3]$. Instead of the auxiliary function θ in this setting we consider the function $\bar{\theta}$ defined by

$$\bar{\theta}\colon [0,1] \to \mathbb{R},\ \bar{\theta}(y) := \begin{cases} \frac{1}{2}\eta\left(\frac{y}{d_2}\right), & y \in [0, d_2), \\ \frac{1}{2}, & y \in [d_2, d_3], \\ \frac{1}{2} + \frac{1}{2}\eta\left(\frac{y-d_3}{1-d_3}\right), & y \in (d_3, 1]. \end{cases}$$

Furthermore, using definition (4.3) from the previous case (with θ replaced by $\bar{\theta}$) we define the function $\psi\colon [0,1] \to \mathbb{R}$. Again, ψ is of class $C^2(\mathbb{R})$ and mutatis mutandis to (4.4) we obtain $\psi'(y) = -U_1(y) + \frac{1}{6}\bar{\theta}'(y)$ for all $y \in [0,1]$. Using the identities $\eta'(0) = \eta'(1) = 0$ the definition of $\bar{\theta}$ implies $\bar{\theta}(y) = 0$ for all $y \in [d_2, d_3]$. Hence, the definition of ψ yields

$$\psi'(y) = -U_1(y) \quad \text{for all } y \in [d_2, d_3]$$

(cf. Figure 4.4).

Similar to the previous case, defining V via the formula (4.6) (with $\bar{\psi}$ instead of ψ) again yields the desired solenoidal function $V \in H^2(\Omega, \mathbb{R}^2) \cap C^1(\overline{\Omega}, \mathbb{R}^2)$. The same arguments as before show that the support of V is indeed contained in $[-d_1, d_1] \times [0,1]$. Thus, again exploiting the structure of ϕ we conclude

$$V(x,y) = U(x,y) \quad \text{for all } (x,y) \in ([-d_1, d_1] \times [d_2, d_3]) \cap \Omega$$

(cf. (4.7) and (4.8) respectively). Since we are in the case $D \subseteq [-d_1, d_1] \times [d_2, d_3]$ this identity in particular holds true on the boundary of the obstacle ∂D.

Additionally, similar to the previous case we calculate $\psi(0) = \psi(1) = 0$ (cf. (4.9)) and due to $\bar{\theta}'(0) = \bar{\theta}'(1) = 0$ we conclude $\psi'(0) = \psi'(1) = 0$ which together with the definition of V implies $V(x,0) = V(x,1) = 0$ for all $x \in \mathbb{R}$.

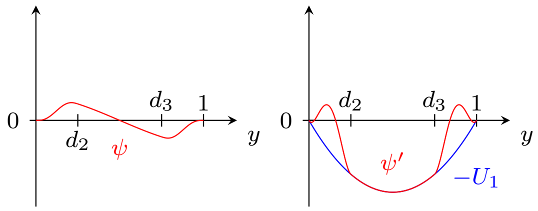


Figure 4.4: Graph of functions ψ and ψ' if D is detached from ∂S

4.2 A Finite Element Solution to $\mathrm{F}\,\boldsymbol{u} = 0$

A Suitable Divergence-Free Finite Element

With the auxiliary function V defined in the previous Section in hand, in the following we will have a closer look at the computation of the approximate solution $\tilde{\omega} \in H(\Omega)$ with $\mathrm{F}\,\tilde{\omega} \approx 0$. We note that one of the crucial assumptions of Theorem 3.4 is the fact that the approximate solution $\tilde{\omega}$ is an element of our solution space $H(\Omega)$, i.e., the numerical algorithms in use must guarantee an approximate solution which is exactly divergence-free. As already mentioned, we want to use finite element methods to compute an approximate solution, however, in this procedure some difficulties appear.

Before presenting details about the divergence-free element in use, we shortly recall the ideas of finite element methods. Therefore, we suppose that the boundary of the computational domain Ω_0 is polygonal, i.e., piecewise linear. In all our examples, we consider domains Ω which have a polygonal boundary as well. Nevertheless, also domains Ω with a "smoother" obstacle can be treated with our approach. Again, we can choose a polygonal computational domain $\Omega_0 \subseteq \Omega$ and extend the approximate solution by zero as described in (3.7). However, one has to be careful when calculating the defect bound et cetera since V is non-zero in $([-d_1, d_1] \times [0,1]) \setminus \Omega_0$ (cf. red parts in Figure 4.5), i.e., it is not sufficient anymore to compute integrals only on the computational domain.

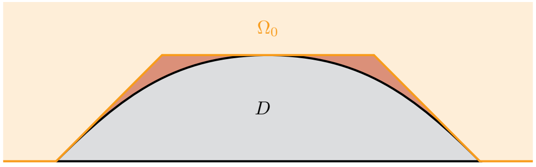


Figure 4.5: Computational domain Ω_0 and "smooth" obstacle

In the case of a "smooth" obstacle isoparametric finite elements come to mind. Since we require a divergence-free approximate solution for our computer-assisted proof, we would require isoparametric finite elements that provide a divergence-free approximate solution by construction. Since the implementation of such isoparametric elements is challenging we only consider polygonal computational domains in the further course of this thesis to avoid this difficulties.

Remark 4.2. *In many applications one can choose the computational domain Ω_0 such that there exists a radius $R > 0$ with $\Omega_0 = S_R \cap \Omega$ where $S_R := (-R, R) \times \mathbb{R}$. However, for instance in the case of a smooth obstacle this is not possible because in our approximation procedure the computational domain is "only" polygonal (cf. Figure 4.5).*

Since the computational domain is supposed to be polygonal, we can fix a triangulation of Ω_0 into N triangles (cf. [12, Definition 5.1]), i.e., there are (closed) triangles $\mathcal{T}_1, \ldots, \mathcal{T}_N$ such that

- $\overline{\Omega_0} = \bigcup_{i=1}^{N} \mathcal{T}_i$.
- If $i, j \in \{1, \ldots, N\}$ are such that $\mathcal{T}_i \cap \mathcal{T}_j = \{z\}$, then z is a corner of $\mathcal{T}_i$ and $\mathcal{T}_j$.

- If $i, j \in \{1, \dots, N\}$, $i \neq j$ are such that the intersection $\mathcal{T}_i \cap \mathcal{T}_j$ contains more than a single point, then $\mathcal{T}_i \cap \mathcal{T}_j$ is an edge of $\mathcal{T}_i$ and $\mathcal{T}_j$.

The collection of the triangles defined above, i.e., the finite element mesh, will be denoted by $\mathcal{M} := \{\mathcal{T}_i \colon i = 1, \dots, N\}$.

One of the crucial ideas of finite element methods consists of local shape functions defined on each cell. To minimize the computational effort, the local shape functions are defined on a reference cell $\hat{\mathcal{T}}$ first, and then in a second step local basis functions on a cell $\mathcal{T}$ are defined via the transformation $\Phi_{\mathcal{T}} \colon \hat{\mathcal{T}} \to \mathcal{T}$. Since there exist various finite elements and since the actual implementation differs from case to case, we only present the ideas about the construction and implementation of the finite elements used in our examples to compute a divergence-free approximation for the velocity. Moreover, in Section 9.4 we shortly describe the implementation of Raviart-Thomas finite elements of higher order for triangles. For a more general overview about finite element methods and the construction of local shape functions, especially for the basic Lagrangian finite elements which will also be needed in Section 6.2.1, we refer the reader to common finite element books for instance by Brenner and Scott [13], Boffi et al. [10] as well as Braess [12].

Since we are interested in divergence-free approximate solutions, the common mixed finite elements like Raviart-Thomas or Taylor-Hood elements cannot be applied because they only yield approximate solutions which are divergence-free with respect to a finite dimensional space of test functions, i.e., when testing with a finite dimensional subspace of $L^2(\Omega)$, but not exactly divergence-free which is not sufficient in our applications (cf. Theorem 3.4). Therefore, we need other elements that provide an exactly divergence-free approximation already by construction.

A possible choice that comes in mind is the Scott-Vogelius finite element (cf. [17] and [37]) which yields divergence-free approximations. Another possibility for our computations could be the Powell-Sabin finite element presented for instance in [118]. Nevertheless, we do not use any of the finite elements mentioned above since the implementation of those elements requires non-standard mesh refinements which are not part of the standard implementation of the finite element software package M++ developed by Wieners et al. (see [113]). For more details about the software package we refer the reader to Section 9.4.

In this thesis we use another possibility to obtain an exactly divergence-free approximation using the well-known Argyris element which yields finite element solutions that are globally C^1-functions, i.e., also the first derivatives of the finite element solutions are not only continuous on each cell, but globally continuous on the entire computational domain Ω_0. Therefore, the finite dimensional subspace $\mathcal{V}^{\mathcal{T}}$ of local shape functions is spanned by polynomials up to degree 5, i.e., $\mathcal{V}^{\mathcal{T}} = \mathbb{P}^5(\mathcal{T})$. Since we have $\dim \mathbb{P}^5(\mathcal{T}) = 21$, to uniquely determine an element of $\mathbb{P}^5(\mathcal{T})$, we need to consider 21 degrees of freedom which are presented in Figure 4.6. Here • denotes the evaluation at a corner of the triangle, the inner circle denotes the evaluation of the first derivative at the corresponding corner and the outer circle denotes the evaluation of the second derivatives at the corner. Moreover, ■ denotes the evaluation of the normal derivative at the midpoint of each of the three faces of $\mathcal{T}$ (cf. Example 3.2.10 in [13]).

In the following, we shortly address the construction of the local shape functions corresponding to the Argyris element on a fixed cell $\mathcal{T}$. For a more detailed description we refer the reader to Section 9.4 where the implementation of the Argyris element is presented. First, let $(\hat{x}_0, \hat{y}_0) := (0, 0)$, $(\hat{x}_1, \hat{y}_1) := (1, 0)$, $(\hat{x}_2, \hat{y}_2) := (0, 1)$ and define the reference

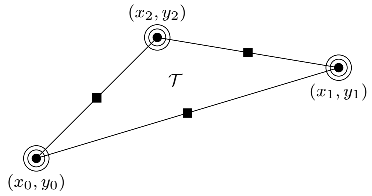


Figure 4.6: Degrees of freedom for the Argyris element

triangle $\hat{\mathcal{T}} := \operatorname{conv}\{(\hat{x}_0, \hat{y}_0), (\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2)\}$. Furthermore, by $\Phi_{\mathcal{T}} \colon \hat{\mathcal{T}} \to \mathcal{T}$ we denote the (bijective) affine linear transformation from the reference cell to $\mathcal{T}$ (cf. Figure 4.7).

In the further course, we suppose that we have constructed local shape functions $\hat{\zeta}_1, \ldots, \hat{\zeta}_{21}$ corresponding to the degrees of freedom described above on the reference triangle (cf. Figure 4.6). In contrast to the definition of local shape function in the Lagrangian case, for the Argyris element we cannot simply use our transformation $\Phi_{\mathcal{T}}$ to define the local shape functions on a cell $\mathcal{T}$ one by one. In Section 9.4 we show that we actually need a linear combination of all 21 transformed shape functions to obtain the correct local basis functions, i.e., there exists a matrix $C^{\mathcal{T}} \in \mathbb{R}^{21 \times 21}$ such that the local shape functions

$$\zeta_i^{\mathcal{T}} : \mathcal{T} \to \mathbb{R},\ \zeta_i^{\mathcal{T}}(x, y) := \sum_{k=1}^{21} C_{k,i}^{\mathcal{T}}\, \hat{\zeta}_k(\Phi_{\mathcal{T}}^{-1}(x, y)) \quad \text{for all } i = 1, \ldots, 21 \tag{4.10}$$

form a dual basis corresponding to the degrees of freedom on $\mathcal{T}$ introduced above (cf. Figure 4.6) and thus, we obtain $\mathcal{V}^{\mathcal{T}} = \operatorname{span}\left\{\zeta_1^{\mathcal{T}}, \ldots, \zeta_{21}^{\mathcal{T}}\right\}$. Furthermore, we are in a position to evaluate the local shape functions on $\mathcal{T}$ by evaluating the reference shape functions (and form a suitable linear combination). Using the chain rule, we calculate similar transformations for the gradients and the Hessian matrices (cf. Section 9.4.1).

Using the local basis functions defined above, we can define new local vector-valued shape functions

$$\xi_i^{\mathcal{T}} : \mathcal{T} \to \mathbb{R}^2,\ \xi_i^{\mathcal{T}}(x, y) := \begin{pmatrix} -\frac{\partial \zeta_i^{\mathcal{T}}}{\partial y}(x, y) \\ \frac{\partial \zeta_i^{\mathcal{T}}}{\partial x}(x, y) \end{pmatrix} \quad \text{for all } i = 1, \ldots, 21.$$

Note that the same arguments as above imply that each component of these new basis functions can be evaluated using the local gradients on the reference element $\hat{\mathcal{T}}$ together with the corresponding transformation (cf. Section 9.4.1).

Hence, we directly get

$$\operatorname{div} \xi_i^{\mathcal{T}}(x, y) = -\frac{\partial^2 \zeta_i^{\mathcal{T}}}{\partial x \partial y}(x, y) + \frac{\partial^2 \zeta_i^{\mathcal{T}}}{\partial x \partial y}(x, y) = 0 \quad \text{for all } (x, y) \in \mathcal{T}, i = 1, \ldots, 21,$$

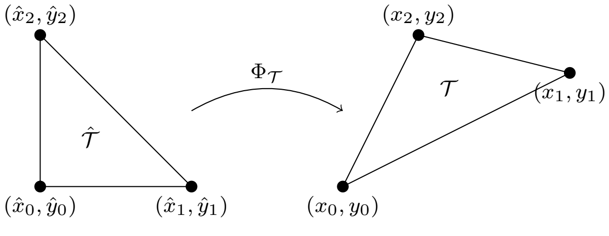


Figure 4.7: Triangle and reference triangle with corresponding transformation

i.e., the new local shape functions are exactly divergence-free by construction which was one of our main goals. For more details about the properties of these new basis functions we refer the reader to [88, Section 3.1].

Furthermore, since $\zeta_1^{\mathcal{T}}, \ldots, \zeta_{21}^{\mathcal{T}}$ are polynomials up to degree 5, by definition, $\xi_1^{\mathcal{T}}, \ldots, \xi_{21}^{\mathcal{T}}$ are polynomials up to degree 4 which will be important to determine the correct quadrature rule needed for instance in the computation of the defect bound (cf. Chapter 5).

Having defined local divergence-free shape functions on each $\mathcal{T} \in \mathcal{M}$ and since $\zeta_1, \ldots, \zeta_{21}$ are global C^1-functions (implying the continuity of the first derivatives), we are in a position to introduce the following finite dimensional subspace in $H(\Omega_0)$:

$$H_{\mathcal{M}}(\Omega_0) := \left\{ v \in C(\overline{\Omega_0}, \mathbb{R}^2) \colon v\big|_{\partial\Omega_0} = 0,\, v\big|_{\mathcal{T}} \in \operatorname{span}\left\{\xi_1^{\mathcal{T}}, \ldots, \xi_{21}^{\mathcal{T}}\right\} \text{ for all } \mathcal{T} \in \mathcal{M} \right\}.$$

We want to emphasize that by construction of our mesh $\mathcal{M}$ for any fixed $(x,y) \in \overline{\Omega_0}$ we find a corresponding triangle $\mathcal{T}$ such that $(x,y) \in \mathcal{T}$. Hence, using the definition of $H_{\mathcal{M}}(\Omega_0)$ for any $v \in H_{\mathcal{M}}(\Omega_0)$ we obtain $v\big|_{\mathcal{T}} = \sum_{i=1}^{21} v_i^{\mathcal{T}} \xi_i^{\mathcal{T}}$. Thus, we compute

$$\operatorname{div} v(x,y) = \sum_{i=1}^{21} v_i^{\mathcal{T}} \operatorname{div} \xi_i^{\mathcal{T}}(x,y) = 0$$

which shows that any function in $H_{\mathcal{M}}(\Omega_0)$ is exactly divergence-free on our complete computational domain.

We close this Section with a short remark about difficulties might appear in our numerical approximation procedure caused by the special choice of our polygonal computational domain Ω_0.

Remark 4.3. *Since our domain Ω is a perturbed strip we cannot avoid reentrant corners in our computational domain Ω_0 (cf. Figure 4.1 and Figure 4.5). It is well known, that reentrant corners have negative effects on the numerical approximation process. One possibility to treat these difficulties are corner singular functions which have been used successfully in the context of computer-assisted proofs by Dagmar Rütters in [89]. For general information about the advantage and usage of corner singular function we refer the reader to [78]. Nevertheless, we cannot exploit the advantage of corner singular functions to increase the accuracy of our approximate solution since we would have to construct exactly divergence-free versions of the corner singular functions (Recall that the approximate solution has to be divergence-free exactly). To overcome the difficulty of reentrant corners anyway, we perform an additional mesh refinement at the reentrant corners which in our examples still leads to an approximate solution with sufficient accuracy.*

Newton's Method

To obtain a solution to our non-linear equation $\mathrm{F}\,u = 0$ we use Newton's method, i.e., in our numerical approximation procedure we perform finitely many Newton steps to compute an approximate solution of high accuracy. To get a better understanding, we shortly recall Newton's method in Banach spaces X and Y applied to a continuously Fréchet-differentiable operator $\mathrm{F}\colon X \to Y$. Starting from $\tilde{\omega}^{(0)} \in X$ the sequence $(w^{(n)})_{n\in\mathbb{N}}$ defined by $w^{(0)} := \tilde{\omega}^{(0)}$, $w^{(n+1)} := w^{(n)} + v^{(n)}$, where $v^{(n)} \in X$ solves

$$(\mathrm{F}'(w^{(n)}))[v^{(n)}] = -\,\mathrm{F}(w^{(n)}), \tag{4.11}$$

converges to a solution of $\mathrm{F}\,u = 0$ if $\|\tilde{\omega}^{(0)}\|$ is "sufficiently small" (cf. [8]).

Since solving equation (4.11) exactly is a quite difficult task, we only compute a sequence of approximate solutions $\tilde{\omega}^{(1)}, \tilde{\omega}^{(2)}, \ldots$. Therefore, in each step we solve (4.11) only approximately, i.e., we compute an approximate solution $\tilde{v}^{(n)} \in X$ to (4.11) (with $w^{(n)}$ replaced by $\tilde{\omega}^{(n)}$) and set $\tilde{\omega}^{(n+1)} := \tilde{\omega}^{(n)} + \tilde{v}^{(n)}$. We stop the iteration process if we have reached the situation such that the error $\|\tilde{\omega}^{(n_0)} - \tilde{\omega}^{(n_0-1)}\| = \|\tilde{v}^{(n_0-1)}\|$ in step $n_0 \in \mathbb{N}$ is smaller than some prescribed tolerance. Then, $\tilde{\omega}^{(n_0)}$ is our desired approximate solution of $\mathrm{F}\,u = 0$.

5 Computation of the Defect Bound

The purpose of this Chapter is the computation of the desired defect bound satisfying (A1), i.e., in the following we present a procedure to compute a constant δ such that

$$\|\mathrm{F}\,\tilde{\omega}\|_{H(\Omega)'} = \|\mathrm{F}(\omega + V)\|_{H(\Omega)'} \leq \delta.$$

Rewriting the operator F (see (3.5)) with the definitions of the bilinear operator B (see (3.1)) as well as of the operator B_Γ (see (3.2)), we obtain

$$\begin{aligned} \mathrm{F}\,\tilde{\omega} &= -\Delta\tilde{\omega} + \mathrm{B}(\tilde{\omega},\tilde{\omega}) + \mathrm{B}_\Gamma\,\tilde{\omega} - g \\ &= -\operatorname{div}\nabla\tilde{\omega} + \mathit{Re}\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g. \end{aligned} \tag{5.1}$$

As suggested in the book by Nakao, Plum and Watanabe (see [74, Section 7.2]), for deriving the desired defect bound δ, we first compute an approximation $\tilde{\rho} \in H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2})$ to the derivative of the approximate solution $\tilde{\omega}$, i.e., we require $\tilde{\rho} \approx \nabla\tilde{\omega}$ and $\operatorname{div}\tilde{\rho} \approx \Delta\tilde{\omega}$ in a suitable sense which will be clarified in the further course.

At the end of this Section, we present the procedure used in our applications to obtain the desired approximation $\tilde{\rho} \in H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2})$ satisfying the assertions above. Nevertheless, for the moment, we will suppose that such an approximation $\tilde{\rho}$ is already computed. Thus, using (5.1) we can rewrite $\mathrm{F}\,\tilde{\omega}$ again and obtain

$$\begin{aligned} \mathrm{F}\,\tilde{\omega} &= \operatorname{div}\tilde{\rho} - \operatorname{div}\nabla\tilde{\omega} - \operatorname{div}\tilde{\rho} + \mathit{Re}\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g \\ &= \operatorname{div}(\tilde{\rho} - \nabla\tilde{\omega}) - \operatorname{div}\tilde{\rho} + \mathit{Re}\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g. \end{aligned} \tag{5.2}$$

Hence, by the triangle inequality we calculate

$$\begin{aligned} \|\mathrm{F}\,\tilde{\omega}\|_{H(\Omega)'} \leq{}& \|\operatorname{div}(\tilde{\rho} - \nabla\tilde{\omega})\|_{H(\Omega)'} \\ &+ \|-\operatorname{div}\tilde{\rho} + \mathit{Re}\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g\|_{H(\Omega)'}. \end{aligned} \tag{5.3}$$

Additionally, by construction of $\tilde{\rho}$ we have $\tilde{\rho} - \nabla\tilde{\omega} \in L^2(\Omega, \mathbb{R}^{2\times 2})$ which together with (2.10) yields

$$\|\operatorname{div}(\tilde{\rho} - \nabla\tilde{\omega})\|_{H(\Omega)'} \leq \|\tilde{\rho} - \nabla\tilde{\omega}\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}.$$

Moreover, keep in mind that $\|\tilde{\rho} - \nabla\tilde{\omega}\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}$ becomes “small” since we construct $\tilde{\rho}$ such that $\tilde{\rho} \approx \nabla\tilde{\omega}$ (cf. approximation procedure for $\tilde{\rho}$ at the end of this Section).

To treat the second norm in (5.3), we apply Cauchy-Schwarz' inequality and use the fact $\tilde{\rho} \in H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2})$ as well as the embedding result in (2.12) to get

$$\begin{aligned}
&\|-\operatorname{div}\tilde{\rho} + Re\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g\|_{H(\Omega)'} \\
&= \sup_{\substack{\varphi\in H(\Omega)\\ \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}=1}} \left| \int_\Omega (-\operatorname{div}\tilde{\rho} + Re\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g)\cdot\varphi\,\mathrm{d}(x,y)\right| \\
&\le \sup_{\substack{\varphi\in H(\Omega)\\ \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}=1}} \|-\operatorname{div}\tilde{\rho} + Re\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g\|_{L^2(\Omega,\mathbb{R}^2)}\|\varphi\|_{L^2(\Omega,\mathbb{R}^2)} \\
&\le C_2\|-\operatorname{div}\tilde{\rho} + Re\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g\|_{L^2(\Omega,\mathbb{R}^2)}. \qquad (5.4)
\end{aligned}$$

Thus, similar to [74, Section 7.2] the computation of the desired defect bound δ (which needs the evaluation of an $H(\Omega)'$-norm) can be reduced to the verified evaluation of the two $L^2(\Omega,\mathbb{R}^2)$-norms

$$\|\tilde{\rho} - \nabla\tilde{\omega}\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \quad \text{and} \quad \|-\operatorname{div}\tilde{\rho} + Re\,[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega}] - g\|_{L^2(\Omega,\mathbb{R}^2)}.$$

The computation of the two norms above is comparably easy in contrast to the computation of the $H(\Omega)'$-norm since only integrals are involved. Nevertheless, the verified evaluation of the second integral turned out to be a more challenging task than one might expect in advance. This difficulty appears because of the relatively high polynomial degree of the function V contained in the integrand (cf. definition of g in (1.14) and definition of V in Section 4.1).

To get a better understanding of this difficulty, we have to study the definition of V in detail (cf. (4.6)). By construction, each component of V is piecewise polynomial with maximal degree 9 (if we consider a suitable subdomain of Ω). Similarly, we see that the components of ∇V are polynomial with degree at most 8. Hence, the term $(V\cdot\nabla)V$ (appearing in the definition of g) is polynomial with maximal degree 17. Thus, to rigorously evaluate a $L^2(\Omega,\mathbb{R}^2)$-norm containing this term, a verified quadrature rule which is exact up to at least degree 34 is necessary or, alternatively, we could use a quadrature rule of lower order with verified remainder term bound. Both possibilities turned out to be not efficient in our examples because on the one hand the computation of quadrature rules of higher order is very challenging since for that purpose large non-linear systems have to be solved rigorously (cf. Section 9.3). On the other hand the computational effort for the computation of the remainder term bounds would become too large. Therefore, in the following, we present a third alternative to overcome this difficulty and show how to evaluate $L^2(\Omega,\mathbb{R}^2)$-norms containing the term $(V\cdot\nabla)V$ rigorously using a verified quadrature rule of "lower order".

First, in addition to the approximation $\tilde{\rho}$, we approximate V in the divergence-free space $\{u \in H^1(\Omega,\mathbb{R}^2) \colon \operatorname{div} u = 0\}$ by $\tilde{V}$ (recall that V does not vanish at the boundary of the obstacle) which has lower maximal polynomial degree. Recall that V is non-zero at the boundary of the obstacle and hence, an approximation in $H(\Omega)$ is not possible. Nevertheless, for the approximation procedure we can use the same divergence-free finite element based on the Argyris element as in the approximation process of $\tilde{\omega}$ but now without zero boundary condition at ∂D. Since the maximal polynomial degree of the divergence-free finite element basis functions is 4 (cf. Section 4.2) the term $(\tilde{V}\cdot\nabla)\tilde{V}$ is now of polynomial degree 7. Thus, a verified quadrature rule which is exact up to polynomial degree 14 is sufficient to evaluate $L^2(\Omega,\mathbb{R}^2)$-norms containing this term. The same approach shows that also $L^2(\Omega,\mathbb{R}^2)$-norms with the terms $(\tilde{\omega}\cdot\nabla)\tilde{\omega}$, $(\tilde{\omega}\cdot\nabla)\tilde{V}$ and $(\tilde{V}\cdot\nabla)\tilde{\omega}$ can be treated with such a quadrature rule. Note that all terms containing the Poiseuille flow U are of lower maximal polynomial degree and thus, a quadrature rule which is exact up to degree 14 suffices to handle $L^2(\Omega,\mathbb{R}^2)$-norms with these terms anyway.

For the moment, we assume that an approximation $\tilde{V}\in\left\{u\in H^1(\Omega,\mathbb{R}^2)\colon \operatorname{div}u=0\right\}$ is in hand and postpone the description of the approximation procedure to the end of this Section. Together with the approximation $\tilde{\rho}$ and using the definition of g (see (1.14)) we calculate

$$\begin{aligned}\mathrm{F}\,\tilde{\omega} &= \operatorname{div}(\tilde{\rho}-\nabla\tilde{\omega})-\operatorname{div}\tilde{\rho}+\Delta V-f+\mathit{Re}\big[(\tilde{\omega}\cdot\nabla)\tilde{\omega}+(\tilde{\omega}\cdot\nabla)U+(U\cdot\nabla)\tilde{\omega}\\ &\qquad-((U+\tilde{\omega})\cdot\nabla)V-(V\cdot\nabla)(U+\tilde{\omega})+(V\cdot\nabla)V\big]\\ &=\operatorname{div}(\tilde{\rho}-\nabla\tilde{\omega})-\operatorname{div}\tilde{\rho}+\Delta V-f+\mathit{Re}\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega}+(\tilde{\omega}\cdot\nabla)U+(U\cdot\nabla)\tilde{\omega}\\ &\qquad-((U+\tilde{\omega})\cdot\nabla)\tilde{V}-(\tilde{V}\cdot\nabla)(U+\tilde{\omega})+(\tilde{V}\cdot\nabla)\tilde{V}\\ &\qquad+((U+\tilde{\omega})\cdot\nabla)(\tilde{V}-V)+((\tilde{V}-V)\cdot\nabla)(U+\tilde{\omega})+(V\cdot\nabla)V-(\tilde{V}\cdot\nabla)\tilde{V}\Big].\end{aligned}$$

Moreover, we have $(V\cdot\nabla)V-(\tilde{V}\cdot\nabla)\tilde{V}=((V-\tilde{V})\cdot\nabla)\tilde{V}+(V\cdot\nabla)(V-\tilde{V})$ which together with the calculations above yields

$$\begin{aligned}\mathrm{F}\,\tilde{\omega} &= \operatorname{div}(\tilde{\rho}-\nabla\tilde{\omega})-\operatorname{div}\tilde{\rho}+\Delta V-f+\mathit{Re}\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega}+(\tilde{\omega}\cdot\nabla)U+(U\cdot\nabla)\tilde{\omega}\\ &\qquad-((U+\tilde{\omega})\cdot\nabla)\tilde{V}-(\tilde{V}\cdot\nabla)(U+\tilde{\omega})+(\tilde{V}\cdot\nabla)\tilde{V}\\ &\qquad+((U+\tilde{\omega}-V)\cdot\nabla)(\tilde{V}-V)+((\tilde{V}-V)\cdot\nabla)(U+\tilde{\omega}-\tilde{V})\Big].\end{aligned}$$

Recall that by the definition of Γ (see (1.12)) and ω (see (3.8)), we obtain the equality $U+\tilde{\omega}-V=U+\omega$. If we define

$$\tilde{\Gamma}:=U-\tilde{V}\qquad\text{and}\qquad \tilde{g}:=f-\Delta V+\mathit{Re}[(\tilde{V}\cdot\nabla)\tilde{\Gamma}+(U\cdot\nabla)\tilde{V}] \tag{5.5}$$

we can retrieve the original structure (cf. (5.2)) as follows

$$\begin{aligned}\mathrm{F}\,\tilde{\omega} &= \operatorname{div}(\tilde{\rho}-\nabla\tilde{\omega})-\operatorname{div}\tilde{\rho}+\mathit{Re}\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega}+(\tilde{\omega}\cdot\nabla)\tilde{\Gamma}+(\tilde{\Gamma}\cdot\nabla)\tilde{\omega}\Big]-\tilde{g}\\ &\qquad+\mathit{Re}\Big[((U+\omega)\cdot\nabla)(\tilde{V}-V)+((\tilde{V}-V)\cdot\nabla)(U+\tilde{\omega}-\tilde{V})\Big],\end{aligned}$$

where the last term is an additional error term which is expected to be "small" if the approximation $\tilde{V}$ is sufficiently good.

We note that it is not necessary to replace ΔV (in $\tilde{g}$) since its maximal polynomial degree is 7 (cf. definition of V) and thus, small enough to be square-integrated exactly with a quadrature rule which is exact up to order 14.

Using the triangle inequality, we obtain the following estimate for the $H(\Omega)'$-norm

$$\begin{aligned}\|\mathrm{F}\,\tilde{\omega}\|_{H(\Omega)'} &\le \|\operatorname{div}(\tilde{\rho}-\nabla\tilde{\omega})\|_{H(\Omega)'}\\ &\qquad+\left\|-\operatorname{div}\tilde{\rho}+\mathit{Re}\left[(\tilde{\omega}\cdot\nabla)\tilde{\omega}+(\tilde{\omega}\cdot\nabla)\tilde{\Gamma}+(\tilde{\Gamma}\cdot\nabla)\tilde{\omega}\right]-\tilde{g}\right\|_{H(\Omega)'}\\ &\qquad+\mathit{Re}\left(\|((U+\omega)\cdot\nabla)(\tilde{V}-V)\|_{H(\Omega)'}+\|((\tilde{V}-V)\cdot\nabla)(U+\tilde{\omega}-\tilde{V})\|_{H(\Omega)'}\right).\end{aligned}$$

The same calculations as above (cf. (5.4)) imply $\|\operatorname{div}(\tilde{\rho} - \nabla\tilde{\omega})\|_{H(\Omega)'} \leq \|\tilde{\rho} - \nabla\tilde{\omega}\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}$ and

$$\begin{aligned}&\Big\| -\operatorname{div}\tilde{\rho} + Re\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\tilde{\Gamma} + (\tilde{\Gamma}\cdot\nabla)\tilde{\omega}\Big] - \tilde{g}\Big\|_{H(\Omega)'}\\ &\quad\leq C_2\Big\| -\operatorname{div}\tilde{\rho} + Re\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\tilde{\Gamma} + (\tilde{\Gamma}\cdot\nabla)\tilde{\omega}\Big] - \tilde{g}\Big\|_{L^2(\Omega,\mathbb{R}^2)}.\end{aligned}$$

The arguments above show that both $L^2(\Omega,\mathbb{R}^2)$-norms can be rigorously evaluated using a quadrature rule which integrates polynomials exactly at least up to degree 14.

Finally, we are left with the verified computation of the additional $H(\Omega)'$-norms. Applying Lemma A.9 (iii) (note that $\tilde{V} - V \in H^1(\Omega,\mathbb{R}^2)$) we obtain

$$\begin{aligned}&\int_\Omega \Big[((U+\omega)\cdot\nabla)(\tilde{V} - V)\Big]\cdot\varphi\,\mathrm{d}(x,y)\\ &\quad\leq C_2\|\nabla(\tilde{V} - V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})}\|U+\omega\|_{L^\infty(\Omega_0,\mathbb{R}^2)}\|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}\end{aligned}$$

for all $\varphi \in H(\Omega)$ which directly yields

$$\|((U+\omega)\cdot\nabla)(\tilde{V} - V)\|_{H(\Omega)'} \leq C_2\|\nabla(\tilde{V} - V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})}\|U+\omega\|_{L^\infty(\Omega_0,\mathbb{R}^2)}.$$

In almost the same manner, using Lemma A.9 (ii) we compute

$$\|((\tilde{V} - V)\cdot\nabla)(U+\tilde{\omega} - \tilde{V})\|_{H(\Omega)'} \leq C_2\|\tilde{V} - V\|_{L^2(\Omega_0,\mathbb{R}^2)}\|\nabla(U+\tilde{\omega} - \tilde{V})\|_{L^\infty(\Omega_0,\mathbb{R}^{2\times 2})}.$$

Hence, to evaluate the norms $\|\nabla(\tilde{V} - V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})}$ and $\|\tilde{V} - V\|_{L^2(\Omega_0,\mathbb{R}^2)}$, we need a quadrature rule that integrates polynomials exactly up to degree 18 (recall that V has maximal polynomial degree 9). Furthermore, two uniform norms (or at least upper bounds of those) have to be computed, i.e., it suffices to compute upper bounds to the L^∞-norms which might result in a slightly larger but computable defect bound δ. Nevertheless, we expect the defect between V and $\tilde{V}$ to be "small" if the approximation $\tilde{V}$ is accurate and thus, the additional terms (in contrast to the first approach presented at the beginning of this Chapter) will become "small" too. In the following Section we present one possible procedure to calculate upper bounds for the uniform norms.

Having computed all norms mentioned above (or at least upper bounds), the previous calculations yield the following estimate on the defect

$$\begin{aligned}\|\mathrm{F}\,\tilde{\omega}\|_{H(\Omega)'} &\leq \|\tilde{\rho} - \nabla\tilde{\omega}\|_{L^2(\Omega,\mathbb{R}^{2\times 2})}\\ &\quad + C_2\Big(\Big\| -\operatorname{div}\tilde{\rho} + Re\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\tilde{\Gamma} + (\tilde{\Gamma}\cdot\nabla)\tilde{\omega}\Big] - \tilde{g}\Big\|_{L^2(\Omega,\mathbb{R}^2)}\\ &\qquad + Re\Big[\|\nabla(\tilde{V} - V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})}\|U+\omega\|_{L^\infty(\Omega_0,\mathbb{R}^2)}\\ &\qquad\quad + \|\tilde{V} - V\|_{L^2(\Omega_0,\mathbb{R}^2)}\|\nabla(U+\tilde{\omega} - \tilde{V})\|_{L^\infty(\Omega_0,\mathbb{R}^{2\times 2})}\Big]\Big)\\ &=: \delta,\end{aligned} \tag{5.6}$$

i.e., the right-hand side of (5.6) defines our desired defect bound.

Remark 5.1. *We note that in contrast to the computation of the approximations (like $\tilde{\omega}$, $\tilde{\rho}$ and $\tilde{V}$) for the computation of the norms interval arithmetic operations are required. Especially, for the quadrature rules in use all quadrature points and their corresponding weights have to be computed rigorously. For more details about the latter fact we refer the reader to Section 9.3.*

5.1 Computation of the L^∞-Norms

In the following, we describe a strategy to obtain upper bounds for the uniform norms needed in definition (5.6). In contrast to the verified computation of L^2-norms (or other L^p-norms with $p < \infty$) the evaluation of L^∞-norms is a quite difficult task. In this situation we have to calculate (or estimate) the range of a function which in a sense requires the verified evaluation of a function over a “non-discrete” set.

Since we are interested in estimating the uniform norms on our computational domain Ω_0 we can reduce the computation to each of the (triangular) cells of our finite element mesh $\mathcal{M}$ (see Section 4.2), i.e., we exploit the fact

$$\|w\|_{L^\infty(\Omega_0)} = \max_{\mathcal{T}\in\mathcal{M}} \|w\|_{L^\infty(\mathcal{T})} = \max_{\mathcal{T}\in\mathcal{M}} \max_{(x,y)\in\mathcal{T}} |w(x,y)| \quad \text{for all } w \in L^\infty(\Omega_0).$$

Having a closer look a the definition of the defect bound δ (cf. (5.6)) we see that it suffices to consider only polynomial functions $w \in L^\infty(\Omega_0)$ in the further course. To estimate the norm $\|w\|_{L^\infty(\mathcal{T})}$ on a fixed triangle $\mathcal{T}$ we use the fact that $\mathcal{T}$ is contained in our finite element mesh $\mathcal{M}$ and thus, there exists the corresponding (bijective) transformation $\Phi_\mathcal{T}$ for the reference triangle first introduced in Section 4.2 (cf. Figure 4.7). Hence, we obtain the identity

$$\|w\|_{L^\infty(\mathcal{T})} = \max_{(x,y)\in\mathcal{T}} |w(x,y)| = \max_{(\hat{x},\hat{y})\in\hat{\mathcal{T}}} |w(\Phi_\mathcal{T}(\hat{x},\hat{y}))| = \|w \circ \Phi_\mathcal{T}\|_{L^\infty(\hat{\mathcal{T}})}$$

which allows us to compute the uniform norm of w on a cell $\mathcal{T}$ (contained in our finite element mesh) by computing the uniform norm of the transformed function $w \circ \Phi_\mathcal{T}$ on the reference cell $\hat{\mathcal{T}}$.

Finally, to compute the desired norm of $w \circ \Phi_\mathcal{T}$ on the reference cell $\hat{\mathcal{T}}$ we actually enclose the range of $w \circ \Phi_\mathcal{T}$, i.e., we compute an enclosure interval $R_w^\mathcal{T}$ such that $w(\Phi_\mathcal{T}(\hat{x},\hat{y})) \in R_w^\mathcal{T}$ for all $(\hat{x},\hat{y}) \in \hat{\mathcal{T}}$. With such an enclosure interval in hand, together with the identity above, we calculate

$$\|w\|_{L^\infty(\mathcal{T})} = \|w \circ \Phi_\mathcal{T}\|_{L^\infty(\hat{\mathcal{T}})} = \max_{(\hat{x},\hat{y})\in\hat{\mathcal{T}}} |w(\Phi_\mathcal{T}(\hat{x},\hat{y}))| \leq \max\left\{|\min R_w^\mathcal{T}|, |\max R_w^\mathcal{T}|\right\}.$$

Before going into further details about the estimate on the norms needed in (5.6), we shortly present a strategy to compute the desired enclosing interval $R_w^\mathcal{T}$ on an abstract level. Therefore, we are going to exploit the expansion of $w \circ \Phi_\mathcal{T}$ using Bernstein polynomials. In [45] and [46], Hungerbühler and Garloff describe techniques to enclose the range of polynomials on a triangle using a basis of Bernstein polynomials of appropriate degree.

We do not want to go into the details about the theory of Bernstein polynomials, however, we shortly recall the definition of such polynomials on the reference triangle $\hat{\mathcal{T}}$. Therefore, let $k \in \mathbb{N}$ and $I^{(k)} := \{(i,j) \colon i,j = 0,\ldots,k,\ i+j \leq k\}$. Then for $(i,j) \in I^{(k)}$ the Bernstein polynomials of degree k are defined as

$$p_{i,j}^{(k)}(\hat{x},\hat{y}) := \binom{k}{i}\binom{k-i}{j}\hat{x}^i\hat{y}^j(1-\hat{x}-\hat{y})^{k-i-j} \quad \text{for all } (\hat{x},\hat{y}) \in \hat{\mathcal{T}}.$$

For the reader's convenience we sort the Bernstein polynomials defined above in a list with one index instead of the multi-indices "i, j" which results in the set $\left\{p_1^{(k)}, \ldots, p_{M^{(k)}}^{(k)}\right\}$ of Bernstein polynomials with $M^{(k)}$ denoting the number of polynomials.

Now, let p be an arbitrary polynomial of degree k on $\hat{\mathcal{T}}$ given in Bernstein expansion, i.e.,

$$p(\hat{x}, \hat{y}) = \sum_{j=1}^{M^{(k)}} b_j^{(k)} p_j^{(k)}(\hat{x}, \hat{y}) \quad \text{for all } (\hat{x}, \hat{y}) \in \hat{\mathcal{T}}$$

with Bernstein coefficients $b_j^{(k)} \in \mathbb{R}$. Then, [45, Theorem 1] (after resorting as above) yields

$$\min_{j=1,\ldots,M^{(k)}} b_j^{(k)} \leq p(\hat{x}, \hat{y}) \leq \max_{j=1,\ldots,M^{(k)}} b_j^{(k)} \quad \text{for all } (\hat{x}, \hat{y}) \in \hat{\mathcal{T}} \tag{5.7}$$

which provides the desired enclosing interval for the polynomial p. To exploit the enclosing result (5.7), in the further course, we are going to rewrite our function $w \circ \Phi_{\mathcal{T}}$ in terms of Bernstein polynomials.

Bernstein Coefficients of a Finite Element Solution on $\mathcal{T}$

First, we describe a strategy to compute the desired Bernstein coefficients in the case where our function w is actually a finite element solution computed with our divergence-free finite elements based on the Argyris element (cf. Section 4.2), i.e., w is of the form $w\big|_{\mathcal{T}} = \sum_{i=1}^{21} w_i^{\mathcal{T}} \xi_i^{\mathcal{T}}$ for suitable coefficients $w_1^{\mathcal{T}}, \ldots, w_{21}^{\mathcal{T}} \in \mathbb{R}$. Recall that the local finite element basis functions $\xi_1^{\mathcal{T}}, \ldots, \xi_{21}^{\mathcal{T}}$ are defined using the gradients of the Argyris shape functions $\zeta_1^{\mathcal{T}}, \ldots, \zeta_{21}^{\mathcal{T}}$. Since we are interested in enclosing the range of the components of $w\big|_{\mathcal{T}}$, we exemplary show how to enclose the second component $w_2\big|_{\mathcal{T}} = \sum_{i=1}^{21} w_i^{\mathcal{T}} \frac{\partial \zeta_i^{\mathcal{T}}}{\partial x}$, i.e., we want to compute an interval $R_{w,2}^{\mathcal{T}}$ such that $w_2\big|_{\mathcal{T}}(x, y) \in R_{w,2}^{\mathcal{T}}$ for all $(x, y) \in \mathcal{T}$. An interval $R_{w,1}^{\mathcal{T}}$ for the first component can be computed mutatis mutandis.

Using the definition of the local basis functions (see (4.10)) together with the chain rule we exploit the fact that the derivatives $\frac{\partial \zeta_1^{\mathcal{T}}}{\partial x}, \ldots, \frac{\partial \zeta_{21}^{\mathcal{T}}}{\partial x}$ can be written in terms of the derivatives of the reference functions $\frac{\partial \hat{\zeta}_k}{\partial \hat{x}}$ and $\frac{\partial \hat{\zeta}_k}{\partial \hat{y}}$ for $k = 1, \ldots, 21$ (cf. (9.4)). Since the Argyris functions are of maximal polynomial degree 5 we directly obtain that $\frac{\partial \hat{\zeta}_k}{\partial \hat{x}}$ and $\frac{\partial \hat{\zeta}_k}{\partial \hat{y}}$ have polynomial degree at most 4 for all $i = 1, \ldots, 21$ and thus, we can expand them in the Bernstein basis $\left\{p_1^{(4)}, \ldots, p_{15}^{(4)}\right\}$ on $\hat{\mathcal{T}}$ presented above (note that $\dim \mathbb{P}^4(\hat{\mathcal{T}}) = 15$). Then, using the representation formula presented in (9.4) together with the bijectivity of the transformation $\Phi_{\mathcal{T}}$, we obtain

$$w_2(\Phi_{\mathcal{T}}(\hat{x}, \hat{y})) = \sum_{i=1}^{21} w_i^{\mathcal{T}} \frac{\partial \zeta_i^{\mathcal{T}}}{\partial x}(\Phi_{\mathcal{T}}(\hat{x}, \hat{y})) = \sum_{j=1}^{15} \hat{b}_j^{(4,w_2)} p_j^{(4)}(\hat{x}, \hat{y}) \quad \text{for all } (\hat{x}, \hat{y}) \in \hat{\mathcal{T}} \tag{5.8}$$

(cf. Appendix A.5) with coefficients $\hat{b}_1^{(4,w_2)}, \ldots, \hat{b}_{15}^{(4,w_2)} \in \mathbb{R}$ (depending on $w_1^{\mathcal{T}}, \ldots, w_{21}^{\mathcal{T}}$). We do not want to compute the coefficients at this stage and refer the reader to the calculations in Appendix A.5. Hence, if we are interested in computing the uniform norm of a finite element solution (restricted to the triangle $\mathcal{T}$), (5.7) implies that

$$R_w^{\mathcal{T}} := \left[\min_{j=1,\dots,15} \hat{b}_j^{(4,w_2)}, \max_{j=1,\dots,15} \hat{b}_j^{(4,w_2)} \right]$$

is an appropriate enclosure interval. Finally, the coefficients $\hat{b}_1^{(4,w_1)}, \dots, \hat{b}_{15}^{(4,w_1)} \in \mathbb{R}$ for the first component of w can be computed using the same strategy.

Furthermore, applying the same techniques to the Hessian matrices of the Argyris shape functions (cf. (9.5)), we obtain Bernstein coefficients for the components of the derivative ∇w (cf. Appendix A.5). We note that in this case it suffices to use Bernstein polynomials of degree at most 3 since only derivatives of the shape functions $\frac{\partial \hat{\zeta}_k}{\partial \hat{x}}$ and $\frac{\partial \hat{\zeta}_k}{\partial \hat{y}}$ for $k = 1, \dots, 21$ are involved.

Enclosing the Range of $U + \omega$ on $\mathcal{T}$

In the enclosing procedure for $U + \omega = \Gamma + \tilde{\omega}$ we exploit the structure of Γ on suitable subregions of our finite element mesh. Therefore, depending on the different obstacle types introduced in Chapter 1 and Section 4.1 respectively, we divide our computational domain into subregions (R_1) to (R_4) presented in Figure 5.1. Since we consider each of the subregions in details later, we postpone a concrete definition of them for a while. Furthermore, we make additional assumptions on the cells contained in our finite element mesh, i.e., we assume that

$$([-d_1, d_1] \times \{d_2, d_3\}) \cap \left(\bigcup_{\mathcal{T} \in \mathcal{M}} \operatorname{int}(\mathcal{T}) \right) = \emptyset \tag{5.9}$$

and

$$\{(x, y) \colon x \in \{\pm d_0, \pm d_1\}, \, y \in [0, 1]\} \cap \left(\bigcup_{\mathcal{T} \in \mathcal{M}} \operatorname{int}(\mathcal{T}) \right) = \emptyset \tag{5.10}$$

hold true for $\mathcal{M}$. In all our applications it turned out that those additional assumptions are not a strong restriction when choosing a mesh for the computation of the approximate solution. In the following, we consider each of the subregions (R_1) to (R_4) separately, state an explicit definition of them depending on the problem type and show how to enclose the range of $U + \omega$ therein.

1. If the cell $\mathcal{T}$ under investigation is contained in the subregion (R_1), i.e., it holds $\mathcal{T} \subseteq \overline{\Omega_0} \setminus ((-d_0, d_0) \times [0, 1])$, we exploit the fact that $\operatorname{supp}(V) \subseteq [-d_0, d_0] \times [0, 1]$ (cf. Section 4.1) which directly implies $\Gamma = U$ in (R_1) and thus, $U + \omega = U + \tilde{\omega}$. Hence, we compute the Bernstein coefficients $\hat{b}_1^{(4,U_1)}, \dots, \hat{b}_{15}^{(4,U_1)} \in \mathbb{R}$ (corresponding to the Bernstein basis with degree at most 4) of $U_1 \circ \Phi_{\mathcal{T}}$ by purely analytical means. Combining those coefficients with the coefficients of the finite element solution $\tilde{\omega}$ (cf. (5.8)) we obtain the identity

$$((U + \tilde{\omega})(\Phi_{\mathcal{T}}(\hat{x}, \hat{y})))_1 = \sum_{j=1}^{15} \left(\hat{b}_j^{(4,\tilde{\omega}_1)} + \hat{b}_j^{(4,U_1)} \right) p_j^{(4)}(\hat{x}, \hat{y}) \quad \text{for all } (\hat{x}, \hat{y}) \in \hat{\mathcal{T}}$$

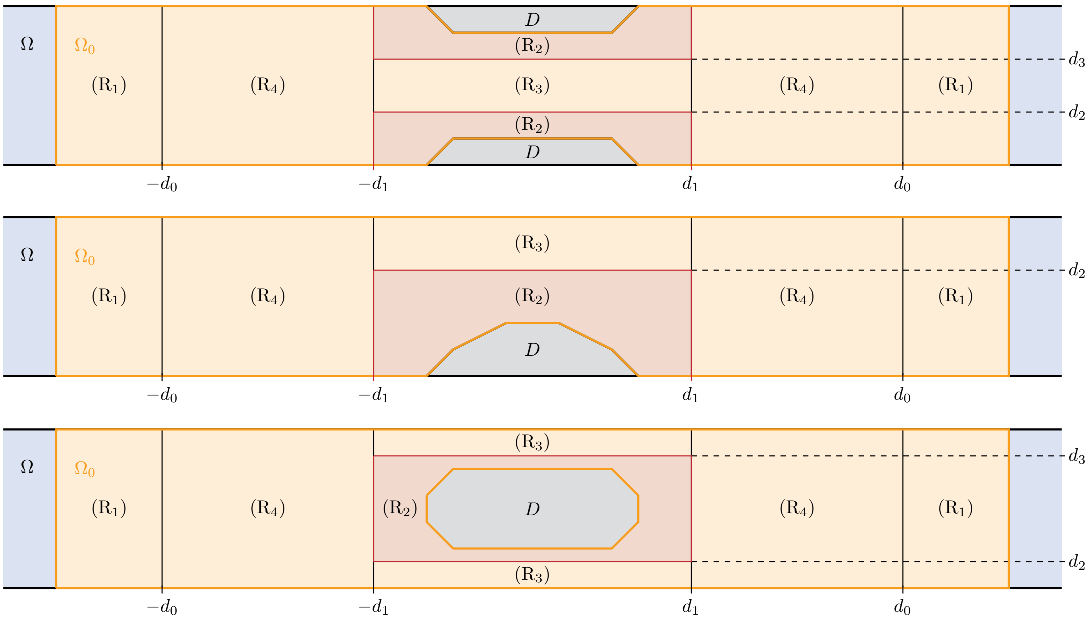


Figure 5.1: Different subregions for computing the L^∞-norm

for the first component. Hence, using the enclosure result mentioned above (cf. (5.7)) in subregion (R$_1$), for the first component we obtain the enclosure interval

$$R^{\mathcal{T}}_{(U+\omega)_1} := \left[\min_{j=1,\dots,15}\left(\hat{b}_j^{(4,\tilde{\omega}_1)} + \hat{b}_j^{(4,U_1)}\right), \max_{j=1,\dots,15}\left(\hat{b}_j^{(4,\tilde{\omega}_1)} + \hat{b}_j^{(4,U_1)}\right)\right].$$

Since the second component of U vanishes, the enclosing interval for the finite element solution $\tilde{\omega}$ is already the desired one, i.e., we define

$$R^{\mathcal{T}}_{(U+\omega)_2} := \left[\min_{j=1,\dots,15} \hat{b}_j^{(4,\tilde{\omega}_2)}, \max_{j=1,\dots,15} \hat{b}_j^{(4,\tilde{\omega}_2)}\right].$$

2. Subregion (R$_2$) is supposed to be containing all parts of the obstacle. Hence, in the case of an obstacle detached from the boundary ∂S, (R$_2$) is given by the single part $[-d_1, d_1] \times [d_2, d_3]$. If the obstacle is located at both sides of the boundary of the strip we define the subregion to be the set $[-d_1, d_1] \times ([0, d_2] \cup [d_3, 1])$. Moreover, considering an obstacle located only at one side of the boundary (R$_2$) reduces to the single part $[-d_1, d_1] \times [0, d_2]$. Using the definition of V (see Section 4.1) we obtain $V = U$ (cf. (4.7)) and thus, $\Gamma = 0$ on (R$_2$) which directly leads to the enclosure intervals

$$R^{\mathcal{T}}_{(U+\omega)_i} := \left[\min_{j=1,\dots,15} \hat{b}_j^{(4,\tilde{\omega}_i)}, \max_{j=1,\dots,15} \hat{b}_j^{(4,\tilde{\omega}_i)}\right] \quad \text{for } i = 1, 2$$

for cells contained in subregion (R$_2$).

3. Considering subregion (R$_3$) (which consists of the complement of the interior of subregion (R$_2$) with respect to $[-d_1, d_1] \times [0, 1]$, cf. Figure 5.1) we use the results of Section 4.1, especially the definition of V as well as the fact that $\phi \equiv 1$ and $\phi' \equiv 0$

on (R_3) to compute

$$\Gamma(x,y) = \begin{pmatrix} U_1(y) + \psi'(y) \\ 0 \end{pmatrix} \qquad \text{for all } (x,y) \in [-d_0, d_0] \times [d_2, 1],$$

which directly implies

$$R^{\mathcal{T}}_{(U+\omega)_2} := \left[\min_{j=1,\ldots,15} \hat{b}_j^{(4,\tilde{\omega}_2)},\ \max_{j=1,\ldots,15} \hat{b}_j^{(4,\tilde{\omega}_2)} \right].$$

It remains to compute an enclosure interval for the first component of $U+\omega$. Therefore, using the definition of ψ (cf. (4.3) and (4.4)) we calculate $\Gamma(x,y) = \frac{1}{6}\theta'(y)$ for all $(x,y) \in \mathcal{T}$ (in the case of obstacle located at the boundary), or $\Gamma(x,y) = \frac{1}{6}\bar{\theta}'(y)$ for all $(x,y) \in \mathcal{T}$ (in the case of obstacles detached from the boundary) respectively. Since both cases coincide in the further course, we restrict our explanations to the (first) case where the obstacle is located at the boundary.

Now, since θ' (cf. (4.2)) has polynomial degree of at most 4 we compute corresponding Bernstein coefficients $\hat{b}_1^{(4,\theta')}, \ldots, \hat{b}_{15}^{(4,\theta')} \in \mathbb{R}$ by analytical means. Hence, the enclosure result (5.7) yields the enclosing interval

$$R^{\mathcal{T}}_{(U+\omega)_1} := \left[\min_{j=1,\ldots,15} \left(\hat{b}_j^{(4,\tilde{\omega}_1)} + \frac{1}{6}\hat{b}_j^{(4,\theta')} \right),\ \max_{j=1,\ldots,15} \left(\hat{b}_j^{(4,\tilde{\omega}_1)} + \frac{1}{6}\hat{b}_j^{(4,\theta')} \right) \right].$$

4. Finally, in subregion (R_4), which is given by the set $([-d_0,-d_1] \cup [d_1,d_0]) \times [0,1]$, we have to consider the complete definition of Γ (cf. (4.6)):

$$\Gamma(x,y) = \begin{pmatrix} U_1(y) + \phi(x)\psi'(y) \\ -\phi'(x)\psi(y) \end{pmatrix} \qquad \text{for all } (x,y) \in ([-d_0,-d_1] \cup [d_1,d_0]) \times [0,1].$$

We note that by construction of Γ each component consists of polynomial functions of degree 9. Therefore, it is no longer sufficient to consider Bernstein polynomials of degree at most 4. For subregion (R_4) we consider the Bernstein basis $\left\{p_1^{(9)}, \ldots, p_{55}^{(9)}\right\}$ on $\hat{\mathcal{T}}$.

A similar computation as before yields the Bernstein coefficients $\hat{b}_1^{(9,\tilde{\omega}_i)}, \ldots, \hat{b}_{55}^{(9,\tilde{\omega}_i)}$ for $i = 1,2$ corresponding to our finite element solution $\tilde{\omega}$. Moreover, by analytic calculations we obtain the Bernstein coefficients $\hat{b}_1^{(9,\Gamma_1)}, \ldots, \hat{b}_{55}^{(9,\Gamma_i)}$ for the components of Γ_i (where $i = 1,2$) which finally results in the enclosure intervals

$$R^{\mathcal{T}}_{(U+\omega)_i} := \left[\min_{j=1,\ldots,55} \left(\hat{b}_j^{(9,\tilde{\omega}_i)} + \hat{b}_j^{(9,\Gamma_i)} \right),\ \max_{j=1,\ldots,55} \left(\hat{b}_j^{(9,\tilde{\omega}_i)} + \hat{b}_j^{(9,\Gamma_i)} \right) \right] \qquad \text{for } i = 1,2$$

for cells contained in subregion (R_4).

Remark 5.2. *To enclose the range of the function $U+\omega$ one can think of a naive implementation via dividing the complete domain into small squares and use interval arithmetic evaluation of $U+\omega$ on each square. However, since in this approach the structure of the contained functions cannot be exploited, the computational effort (and thus the computation time) heavily increases compared to the strategy presented above.*

Enclosing the Range of $\nabla(U + \tilde\omega - \tilde V)$ on $\mathcal{T}$

Finally, we are left with the computation of the uniform norm of $\nabla(U+\tilde\omega-\tilde V)$ on a given (triangular) cell $\mathcal{T}$. Since $\tilde\omega - \tilde V$ is an Argyris finite element function we apply the techniques mentioned before to calculate the corresponding Bernstein coefficients $\hat b_1^{(3,(\nabla(\tilde\omega-\tilde V))_{i,j})}, \dots, \hat b_{15}^{(3,(\nabla(\tilde\omega-\tilde V))_{i,j})} \in \mathbb{R}$ for $i,j = 1,2$. Again, the Bernstein coefficients $\hat b_1^{(3,(\nabla U)_{1,2})}, \dots, \hat b_{15}^{(3,(\nabla U)_{1,2})_{i,j})} \in \mathbb{R}$ for the derivative of the Poiseuille flow $(\nabla U)_{1,2}$ can be obtained by purely analytical calculations. Finally, we obtain the enclosing intervals

$$R^{\mathcal{T}}_{(\nabla(U+\tilde\omega-\tilde V))_{i,j}} := \begin{cases} \Bigg[\min\limits_{k=1,\dots,10}\left(\hat b_k^{(3,(\nabla(\tilde\omega-\tilde V))_{1,2})} + \hat b_k^{(3,(\nabla U)_{1,2})}\right), \\ \qquad \max\limits_{k=1,\dots,10}\left(\hat b_k^{(3,(\nabla(\tilde\omega-\tilde V))_{1,2})} + \hat b_k^{(3,(\nabla U)_{1,2})}\right)\Bigg], & (i,j)=(1,2), \\ \left[\min\limits_{k=1,\dots,10} \hat b_k^{(3,(\nabla(\tilde\omega-\tilde V))_{i,j})}, \max\limits_{k=1,\dots,10} \hat b_k^{(3,(\nabla(\tilde\omega-\tilde V))_{i,j})}\right], & \text{otherwise} \end{cases}$$

independent of the different subregions introduced in the previous Subsection.

Remark 5.3. *Since all functions appearing within the norm are (exactly) divergence-free, we actually do not need to to compute both of the intervals* $R^{\mathcal{T}}_{(\nabla(U+\tilde\omega-\tilde V))_{1,1}}$ *and* $R^{\mathcal{T}}_{(\nabla(U+\tilde\omega-\tilde V))_{2,2}}$ *(nor the corresponding Bernstein coefficients). In our applications we exploit the solenoidal structure, i.e.,* $\frac{\partial(U+\tilde\omega-\tilde V)_1}{\partial x} = -\frac{\partial(U+\tilde\omega-\tilde V)_2}{\partial y}$ *holds true on our computational domain* Ω_0*, and use the formula*

$$\begin{aligned}\|\nabla(U+\tilde\omega-\tilde V)\|^2_{L^\infty(\Omega_0,\mathbb{R}^{2\times2})} = 2\|(\nabla(U+\tilde\omega-\tilde V))_{1,1}\|^2_{L^\infty(\Omega_0)} + \|(\nabla(U+\tilde\omega-\tilde V))_{1,2}\|^2_{L^\infty(\Omega_0)} \\ + \|(\nabla(U+\tilde\omega-\tilde V))_{2,1}\|^2_{L^\infty(\Omega_0)}\end{aligned}$$

to compute the desired norm on the entire computational domain.

5.2 Approximations for the Computation of the Defect Bound

We are left with the computation of the approximations $\tilde\rho$ and $\tilde V$ needed in the definition of our defect bound δ. Since the computation of $\tilde V$ is independent of $\tilde\rho$ but not the other way around it makes sense to compute $\tilde V$ in advance.

Computation of $\tilde V$

As already mentioned in the beginning of this Chapter for the approximation procedure we can use the same divergence-free finite element as for the computation of the approximation $\tilde\omega$ (without zero-boundary condition at the obstacle). Since we are interested in a "small" defect bound δ it is advisable to approximate V using a minimization method. Having a closer look at the definition of δ (cf. (5.6)) we see that it makes sense to minimize the following expression

$$\|\nabla(\tilde V - V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times2})}\|U+\omega\|_{L^\infty(\Omega_0,\mathbb{R}^2)} + \|\tilde V - V\|_{L^2(\Omega_0,\mathbb{R}^2)}\|\nabla(U+\tilde\omega-\tilde V)\|_{L^\infty(\Omega_0,\mathbb{R}^{2\times2})}.$$

We are interested in a preferably simple approximation procedure for $\tilde{V}$ which simultaneously results in a small additional error term in the formula for the defect bound δ.

First, to define the functional for the minimization process we replace $\tilde{V}$ by the original function V in the second uniform norm, i.e., instead of $\|\nabla(U+\tilde{\omega}-\tilde{V})\|_{L^\infty(\Omega_0,\mathbb{R}^2)}$ we consider $\|\nabla(U+\tilde{\omega}-V)\|_{L^\infty(\Omega_0,\mathbb{R}^2)} = \|\nabla(U+\omega)\|_{L^\infty(\Omega_0,\mathbb{R}^2)}$ (which is now independent of $\tilde{V}$). We might expect that this change does not have a large effect on the accuracy of $\tilde{V}$ since for a "good" approximation $\tilde{V}$ to V we have $\tilde{V} \approx V$. We note that in the evaluation of the right-hand side of definition (5.6) we cannot make this replacement, i.e., for the computation of the defect bound δ we have to consider the norms initially stated in (5.6).

Hence, we consider the functional $J\colon \big\{u \in H^1(\Omega,\mathbb{R}^2)\colon\ \operatorname{div} u = 0\big\} \to \mathbb{R}$ given by

$$\begin{aligned} J(\tilde{V}) :=& \frac{1}{2}\|\nabla(\tilde{V}-V)\|^2_{L^2(\Omega_0,\mathbb{R}^{2\times2})}\|U+\omega\|^2_{L^\infty(\Omega_0,\mathbb{R}^2)} \\ &+ \frac{1}{2}\|\tilde{V}-V\|^2_{L^2(\Omega_0,\mathbb{R}^2)}\|\nabla(U+\omega)\|^2_{L^\infty(\Omega_0,\mathbb{R}^{2\times2})}, \end{aligned}$$

which leads to a "simple" linear algebraic system which only needs to be solved approximately, i.e., the computational effort for the minimization process is relatively small.

Again, we note that both uniform norms in the definition of J are now independent of $\tilde{V}$ and thus, can be computed in advance. During the approximation process, i.e., during the minimization of the functional J, the norms only play the role of additional weights in front of the two terms $\|\nabla(\tilde{V}-V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times2})}$ and $\|\tilde{V}-V\|_{L^2(\Omega_0,\mathbb{R}^2)}$.

Computation of $\tilde{\rho}$

Having computed the approximation $\tilde{V}$, we are in a position to start the approximation procedure for $\tilde{\rho}$. Recall that in the beginning of the Chapter we defined $\tilde{\Gamma} = U - \tilde{V}$.

To compute the desired approximation $\tilde{\rho}$, we follow the lines in [74, Section 7.2] and minimize the functional $J\colon H(\operatorname{div},\Omega,\mathbb{R}^{2\times2}) \to \mathbb{R}$ defined by

$$\begin{aligned} J(\tilde{\rho}) :=& \frac{1}{2}\|\tilde{\rho}-\nabla\tilde{\omega}\|^2_{L^2(\Omega,\mathbb{R}^{2\times2})} \\ &+ \frac{1}{2}{C_2}^2\left\| -\operatorname{div}\tilde{\rho} + Re\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\tilde{\Gamma} + (\tilde{\Gamma}\cdot\nabla)\tilde{\omega}\Big] - \tilde{g}\right\|^2_{L^2(\Omega,\mathbb{R}^2)}. \end{aligned}$$

Again, we apply finite element methods to compute the desired approximation $\tilde{\rho}$. Since we have the additional assumption $\tilde{\rho} \in H(\operatorname{div},\Omega,\mathbb{R}^{2\times2})$ we have to use finite elements that provide solutions in $H(\operatorname{div},\Omega,\mathbb{R}^{2\times2})$ exactly. Therefore, we could use vector-valued Lagrangian elements which yield solutions in $H^1(\Omega,\mathbb{R}^{2\times2})$ but in our applications it turned out that these elements are not suitable for the approximation procedure since in this situation a finite element space using Lagrangian finite elements requires a relatively large number of finite element basis functions which increases the computational effort for this ansatz space. An alternative is given by the Raviart Thomas elements which provide a solution in $H(\operatorname{div},\Omega,\mathbb{R}^{2\times2})$ a priori. Since the lowest order Raviart Thomas element results in a relatively rough error bound, we have to use higher order versions which were not part of the finite element software package M++ before. In Section 9.4.2 we describe a procedure how to implement higher order Raviart Thomas elements on triangles which are now integrated in the software package M++.

In particular, in our applications we solve the (finite dimensional) discrete problem

$$\begin{aligned}&\text{Find } \tilde{\rho}_h \in RT_2 \text{ such that}\\ &{C_2}^2\langle -\operatorname{div}\tilde{\rho}_h + Re\Big[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\tilde{\Gamma} + (\tilde{\Gamma}\cdot\nabla)\tilde{\omega}\Big] - \tilde{g}, \operatorname{div}\varphi_h\rangle_{L^2(\Omega,\mathbb{R}^2)}\\ &\qquad = \langle \tilde{\rho}_h - \nabla\tilde{\omega}, \varphi_h\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} \quad \text{for all } \varphi_h \in RT_2\end{aligned}$$

to obtain the desired approximation in $H(\operatorname{div},\Omega,\mathbb{R}^{2\times 2})$.

Remark 5.4. *Due to the definition of the involved norms it is easy to see that the computation of the two rows of $\tilde{\rho}$ are independent of each other, i.e., each row can be computed separately. In our applications we exploit this fact to reduce the computational effort which shows up in a lower computational time on the one hand and also in a significant reduction of memory allocation on the other hand.*

6 Computation of the Norm Bounds

We present two approaches to compute the desired norm bounds K, and K^* respectively, needed in assumptions (A2), and (A3) respectively. In Section 6.1 we describe an approach based on ideas of Wieners (cf. [112]), where the success of this approach is directly linked to the Reynolds number, i.e., the approach is expected to fail if the Reynolds number is chosen too "large".

The second approach presented in Section 6.2 uses bounds for the essential spectrum and enclosures for the eigenvalues of $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{L}_{U+\omega}$ and $\Phi^{-1}\,\mathrm{L}_{U+\omega}(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*$ "close" to zero to obtain the desired norm bounds K, and K^* respectively. For the latter, the Rayleigh-Ritz Method, the Temple-Lehmann Method together with its Goerisch extension and a homotopy method introduced by Plum (cf. e.g. [14, Section 4.2]) play crucial roles. More detailed explanations about the abstract eigenvalue methods to obtain the desired norm bounds and its application in our examples will be given in Section 6.2.1.

Before describing both approaches in detail, we will have a closer look at the adjoint operator $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*$ which appears in assumption (A3).

Using the properties of the isometric isomorphism Φ and applying Lemma A.10 (ii) we calculate

$$\begin{aligned}
&\langle(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*u,\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\
&\quad= \langle\Phi^{-1}\,\mathrm{L}_{U+\omega}\,\varphi,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} = (\mathrm{L}_{U+\omega}\,\varphi)[u] = (-\Delta\varphi+\mathrm{B}_{U+\omega}\,\varphi)[u]\\
&\quad= \int_\Omega (\nabla\varphi\bullet\nabla u + Re\,[(\varphi\cdot\nabla)(U+\omega)+((U+\omega)\cdot\nabla)\varphi]\cdot u)\;\mathrm{d}(x,y)\\
&\quad= \int_\Omega \left(\nabla\varphi\bullet\nabla u + Re\left[u^T(\nabla(U+\omega))\varphi+u^T(\nabla\varphi)(U+\omega)\right]\right)\mathrm{d}(x,y) \qquad (6.1)\\
&\quad= \int_\Omega \left(\nabla\varphi\bullet\nabla u + Re\left[u^T(\nabla(U+\omega))\varphi-\varphi^T(\nabla u)(U+\omega)\right]\right)\mathrm{d}(x,y)\\
&\quad= \int_\Omega \left(\nabla u\bullet\nabla\varphi + Re\,\varphi^T\left[(\nabla(U+\omega))^T u-(\nabla u)(U+\omega)\right]\right)\mathrm{d}(x,y)
\end{aligned}$$

for all $u,\varphi\in H(\Omega)$. Thus, analogously to the definition of the operator B_w for $w\in W(\Omega)$ (see (3.3)) we define

$$\hat{\mathrm{B}}_w\colon H(\Omega)\to L^2(\Omega,\mathbb{R}^2),\ \hat{\mathrm{B}}_w\,u := Re\left[(\nabla w)^T u-(\nabla u)w\right] \qquad (6.2)$$

for $w\in W(\Omega)$, i.e., $(\hat{\mathrm{B}}_w\,u)[\varphi] = Re\int_\Omega\varphi^T\left[(\nabla w)^Tu-(\nabla u)w\right]\,\mathrm{d}(x,y)$ for all $u,\varphi\in H(\Omega)$. Similar as in Section 3.1 we see that for any $w\in W(\Omega)$ the expression $\hat{\mathrm{B}}_w\,u\in L^2(\Omega,\mathbb{R}^2)$ defines an element in $H(\Omega)'$ satisfying $\|\hat{\mathrm{B}}_w\,u\|_{H(\Omega)'}\le C_2\|\hat{\mathrm{B}}_w\,u\|_{L^2(\Omega,\mathbb{R}^2)}$ for all $u\in H(\Omega)$.

Additionally, using the definition of $\hat{\mathrm{B}}_w$ above we set

$$\hat{\mathrm{L}}_w \colon H(\Omega) \to H(\Omega)', \ \hat{\mathrm{L}}_w\, u = -\Delta u + \hat{\mathrm{B}}_w\, u \tag{6.3}$$

for $w \in W(\Omega)$. If we chose $w = U + \omega$ the operator $\hat{\mathrm{L}}_{U+\omega}$ is the counter part of the linearization $\mathrm{L}_{U+\omega}$ in the dual formulation (cf. (3.9)).

Applying the definitions of the operators $\hat{\mathrm{L}}_{U+\omega}$ and $\hat{\mathrm{B}}_{U+\omega}$ as well as the calculations in (6.1) we obtain

$$\langle (\Phi^{-1}\,\mathrm{L}_{U+\omega})^* u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} = (\hat{\mathrm{L}}_{U+\omega}\, u)[\varphi] = \langle \Phi^{-1}\,\hat{\mathrm{L}}_{U+\omega}\, u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u, \varphi \in H(\Omega),$$

i.e., the following equality of operators holds true:

$$(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* = \Phi^{-1}\,\hat{\mathrm{L}}_{U+\omega}\,. \tag{6.4}$$

Hence, assumption (A3) now reads as

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \le K^* \|(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* u\|_{H_0^1(\Omega,\mathbb{R}^2)} = K^* \|\Phi^{-1}\,\hat{\mathrm{L}}_{U+\omega}\, u\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega).$$

Remark 6.1. *Comparing the definitions of* $\mathrm{L}_{U+\omega}$ *(see* (3.9)*) and* $\hat{\mathrm{L}}_{U+\omega}$ *(see* (6.3)*) we see that* $\Phi^{-1}\,\mathrm{L}_{U+\omega}$ *is not symmetric and thus, not self-adjoint which finally justifies the strategy introduced in Chapter 3.*

In analogy to Proposition 3.1 in Section 3.1 we obtain the following Proposition for $\hat{\mathrm{B}}_w$ (with $w \in W(\Omega)$). Again, its proof is postponed to the Appendix (see Proof of Proposition 6.2 in Appendix A.3).

Proposition 6.2. *The following assertions hold true:*

(i) $\hat{\mathrm{B}}_v + \hat{\mathrm{B}}_w = \hat{\mathrm{B}}_{v+w}$ *for all* $v, w \in W(\Omega)$.

(ii) $\hat{\mathrm{B}}_{-w} = -\,\hat{\mathrm{B}}_w$ *for all* $w \in W(\Omega)$.

In the following we describe both approaches for the computation of the norm bounds in detail.

6.1 First Approach

The approach to compute the desired norm bounds presented in this Section, first, was suggested by Wieners in [112]. Therefore, we set the constant σ used in the inner product on $H(\Omega)$ (see (2.5)) to zero. Note that the choice $\sigma = 0$ is possible since Poincaré's inequality holds for the strip S and, hence, for our domain $\Omega \subseteq S$ (see Lemma A.4 and Remark A.5).

Since we fixed $\sigma = 0$ the definition of the isometric isomorphism Φ (see (2.14) and (2.15)) together with the definition of the linearization $\mathrm{L}_{U+\omega}$ (see (3.9)) yields

$$\|u\|^2_{H_0^1(\Omega,\mathbb{R}^2)} = (-\Delta u)[u] = (\mathrm{L}_{U+\omega}\, u)[u] - (\mathrm{B}_{U+\omega}\, u)[u] \quad \text{for all } u \in H(\Omega).$$

Thus, applying Lemma A.10 (ii) and Lemma A.9 (iii) (cf. (3.4)) we conclude

$$\begin{aligned}\|u\|^2_{H^1_0(\Omega,\mathbb{R}^2)} &\leq \|\mathrm{L}_{U+\omega}\, u\|_{H(\Omega)'}\|u\|_{H^1_0(\Omega,\mathbb{R}^2)} + \|\mathrm{B}_{U+\omega}\, u\|_{H(\Omega)'}\|u\|_{H^1_0(\Omega,\mathbb{R}^2)}\\ &\leq \|\mathrm{L}_{U+\omega}\, u\|_{H(\Omega)'}\|u\|_{H^1_0(\Omega,\mathbb{R}^2)} + 2C_2 Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)}\|u\|^2_{H^1_0(\Omega,\mathbb{R}^2)}\end{aligned}$$

for all $u \in H(\Omega)$. Rearranging the terms in the inequality directly implies

$$\left(1 - 2C_2 Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)}\right)\|u\|_{H^1_0(\Omega,\mathbb{R}^2)} \leq \|\mathrm{L}_{U+\omega}\, u\|_{H(\Omega)'} \quad \text{for all } u \in H(\Omega).$$

Recall that assumption (A2) requires a constant $K \geq 0$ with $\|u\|_{H^1_0(\Omega,\mathbb{R}^2)} \leq K\|\mathrm{L}_{U+\omega}\, u\|_{H(\Omega)'}$ for all $u \in H(\Omega)$. Hence, if $2C_2 Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)} < 1$ is satisfied, assumption (A2) holds for all

$$K \geq \frac{1}{1 - 2C_2 Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)}}. \tag{6.5}$$

Since we are interested in a "small" norm bound K our desired bound can be chosen as the constant (or at least a "slightly increased" upper bound) given by the right-hand side of (6.5).

In almost the same manner, using the alternative representation of assumption (A3) with the operator $\hat{\mathrm{L}}_{U+\omega}$ (see (6.3)), together with Lemma A.9 (ii) and (iii) we obtain that assumption (A3) is satisfied for all

$$K^* \geq \frac{1}{1 - 2C_2 Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)}}$$

if $2C_2 Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)} < 1$ holds true again. Thus, in both cases we can chose the same constant for K on the one hand and for K^* on the other hand (cf. Remark 6.5 and [74, bottom of p. 335]).

Note that for the validation of the crucial inequality $2C_2 Re\|U+\omega\|_{L^\infty(\Omega)} < 1$ as well as for the definition of our norm bounds an upper bound for the norm $\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)}$ is sufficient. A possible procedure for the computation of such an upper bound is described in Section 5.1.

In contrast to the calculation in Section 5.1 we have to evaluate the supremum over the entire domain Ω but not only over the computational domain. However, these methods are applicable since only the Poiseuille flow U has support outside of our computational domain Ω_0 (see Chapter 4), i.e., the unbounded part can be treated by purely analytical calculations. Hence, we can split the norm as follows and use the techniques presented in Section 5.1:

$$\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)} = \max\left\{\|U+\omega\|_{L^\infty(\Omega_0,\mathbb{R}^2)}, \|U\|_{L^\infty(\Omega\setminus\Omega_0,\mathbb{R}^2)}\right\}.$$

Note that $\|U\|_{L^\infty(\Omega\setminus\Omega_0,\mathbb{R}^2)} = \frac{1}{4}$ can easily be computed analytically and thus, we obtain directly

$$\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)} \geq \|U\|_{L^\infty(\Omega\setminus\Omega_0,\mathbb{R}^2)} = \frac{1}{4}. \tag{6.6}$$

Remark 6.3. (i) *Using the lower bound for $\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)}$ in* (6.6) *we see that the crucial inequality $2C_2 Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)} < 1$ can only be satisfied if the Reynolds number is "sufficiently small". Hence, our first approach will fail in cases where the Reynolds number becomes "too large" (cf. [112]).*

(ii) *Due to the unboundedness of our domain Ω, in contrast to the paper by Wieners (cf. [112]), we have to deal with the L^∞-norm whereas in the original work a L^4-norm is used (recall that $U+\omega \notin L^4(\Omega,\mathbb{R}^2)$ in our considerations).*

6.2 Second Approach

The second approach uses spectral bounds to obtain the desired constants K satisfying (A2) and K^* with (A3) respectively. The crucial steps are based on ideas of Plum described for instance in [85, Section 4.3] and [74, Section 9.4]. To the best of the author's knowledge all former applications of the same computer-assisted techniques used for unbounded domains up to the present day (see e.g. [14], [63], [86], [117], ...) strongly exploit the self-adjointness of the operator $\Phi^{-1}\,\mathrm{L}_{U+\omega}$ and use a spectral decomposition argument to compute K. Note that in the self-adjoint case the computation of a constant K^* is not needed since the self-adjointness directly yields the surjectivity of $\Phi^{-1}\,\mathrm{L}_{U+\omega}$ which, in the end, was a crucial assumption for a successful proof of Theorem 3.4 (cf. Section 3.2 and proof of Theorem 1 in [14]).

At this stage we want to mention that in contrast to our approach, Nakao's method (cf. e.g. [68], [70], [72], [75], [76], ...) does not require the self-adjointness. However, this method is only applicable to bounded domains which is not the case in our considerations.

Due to the lack of self-adjointness in our application (cf. Remark 6.1), we need to find another approach to compute constants the K and K^* satisfying (A2) and (A3) respectively. The strategy presented in the further course first was introduced on an abstract level by Plum in [74, Section 9.4]. In this thesis we present the first application of this approach.

As suggested in [74, Section 9.4.1.2], we rewrite assumption (A2) as follows

$$\langle u, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \le K^2 \langle \Phi^{-1}\,\mathrm{L}_{U+\omega}\, u, \Phi^{-1}\,\mathrm{L}_{U+\omega}\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega). \tag{6.7}$$

Hence, to obtain the desired constant K we have to determine a (positive) lower bound for

$$\inf_{\substack{u\in H(\Omega)\\ u\neq 0}} \frac{\langle \Phi^{-1}\,\mathrm{L}_{U+\omega}\, u, \Phi^{-1}\,\mathrm{L}_{U+\omega}\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}}{\langle u, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}},$$

which leads to the computation of a (positive) lower bound $\underline{\sigma}$ for the spectral points of the following eigenvalue problem in weak formulation:

$$u \in H(\Omega),\ \langle \Phi^{-1}\,\mathrm{L}_{U+\omega}\, u, \Phi^{-1}\,\mathrm{L}_{U+\omega}\, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \lambda \langle u, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega). \tag{6.8}$$

For the computation of $\underline{\sigma}$ two separate tasks have to be considered. On the one hand we have to compute a (positive) lower for the essential spectrum of (6.8) (see Section 6.2.2) and on the other hand a (positive) lower bound for the isolated eigenvalues of (6.8) has to be found (see Section 6.2.1). For both of these tasks we make enormous use of the computer which will be described in the corresponding Sections.

Remark 6.4. *Note that the essential spectrum of* (6.8) *is defined via the associated self-adjoint operator* $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{L}_{U+\omega}$. *For more detailed information we refer the reader to Section 6.2.2 and [74, Section 10.2.1].*

For the moment we suppose that we have such a lower bound $\underline{\sigma} > 0$ in hand. Then, by (6.7), assumption (A2) is satisfied for all K such that

$$K \geq \frac{1}{\sqrt{\underline{\sigma}}},$$

i.e., it makes sense to chose K equal to $\frac{1}{\sqrt{\underline{\sigma}}}$ or at least as a tight upper bound which in general will be the case in applications since we have to take all rounding errors in the calculation into account (cf. description of interval arithmetic in Section 3.3).

In almost the same manner, we have to compute a lower bound $\underline{\sigma}^* > 0$ for the spectral points of the eigenvalue problem

$$u \in H(\Omega),\ \langle (\Phi^{-1} \mathrm{L}_{U+\omega})^* u, (\Phi^{-1} \mathrm{L}_{U+\omega})^* \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \lambda \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega), \tag{6.9}$$

and hence, the similar arguments as above show that (A3) holds true for all K^* satisfying

$$K^* \geq \frac{1}{\sqrt{\underline{\sigma}^*}}.$$

Using the alternative representation of the adjoint operator (see (6.4)), we can equivalently rewrite the second eigenvalue problem (6.9) and obtain the eigenvalue problem

$$u \in H(\Omega),\ \langle \Phi^{-1} \hat{\mathrm{L}}_{U+\omega}\, u, \Phi^{-1} \hat{\mathrm{L}}_{U+\omega}\, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \lambda \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega), \tag{6.10}$$

which is of the same form as the first eigenvalue problem (6.8).

Remark 6.5. *We note that the positive eigenvalues of the eigenvalue problems* (6.8) *and* (6.9) *coincide (cf. [74, bottom of p. 335]). However, one of these problems might have an eigenvalue* 0*. Thus, it is not sufficient to consider only one of these eigenvalue problems, but we have to compute a lower bound for the smallest eigenvalue for both problems. In the case when both smallest eigenvalues are proven to be positive we a posteriori conclude that K and K^* coincide.*

In the following we postpone the computation of the lower bound for the essential spectrum to Section 6.2.2 and describe the procedure to compute the lower bounds for the isolated eigenvalues of the eigenvalue problems (6.8) and (6.9) below the essential spectrum (if such eigenvalues exist) in detail beforehand.

Since we have the alternative representation of the second eigenvalue problem (see (6.10)) at several stages in the further course we can treat both problems simultaneously concerning the description of the method. Therefore, most of the following formulations contain two parts (which sometimes are very similar) corresponding either to eigenvalue problem (6.8) or to problem (6.9) (and (6.10) respectively). Note that, if the arguments in both parts are quite similar, we only describe the steps of the first part in detail and add some remarks for the second.

Before providing the details concerning the lower bounds for the isolated eigenvalues and the essential spectrum we shortly present a strategy which allows us to compute the defect bound δ and the norm bounds K and K^* respectively using different approximate solutions.

Transfer Norm Bounds for Different Approximate Solutions

In this short Subsection we assume that a constant K_c is already computed using the approximate solution $\omega_c = \tilde{\omega}_c - V \in H(\Omega)$, i.e., we consider the eigenvalue problem (6.8) with ω replaced by the new approximate solution ω_c. Hence, we suppose that we have a constant K_c in hand such that

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq K_c \|\mathrm{L}_{U+\omega_c} u\|_{H(\Omega)'} \quad \text{for all } u \in H(\Omega).$$

Here, the index c indicates that in our applications ω_c is an approximation computed on a coarse finite element mesh whereas the approximate solution ω (used for the computation of the defect bound δ) is still computed on a fine finite element mesh (which in general results in a more accurate approximate solution compared to that one on the coarse mesh) to obtain a small defect bound δ.

Since assumptions (A1) - (A3) needed in our existence and enclosure theorem have to be computed using the same approximate solution (cf. proof of Theorem 3.4), in the further course we present a strategy to compute the desired norm bound K using the “coarse” norm bound K_c.

Therefore, we first calculate

$$\begin{aligned}\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} &\leq K_c(\|(\mathrm{L}_{U+\omega_c} - \mathrm{L}_{U+\omega})u\|_{H(\Omega)'} + \|\mathrm{L}_{U+\omega} u\|_{H(\Omega)'}) \\ &\leq K_c(\|\mathrm{L}_{U+\omega_c} - \mathrm{L}_{U+\omega}\|_{\mathcal{B}} \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} + \|\mathrm{L}_{U+\omega} u\|_{H(\Omega)'}) \quad \text{for all } u \in H(\Omega).\end{aligned}$$

Now, using the definitions of $\mathrm{L}_{U+\omega_c}$ and $\mathrm{L}_{U+\omega}$ respectively (cf. (3.10)) together with Proposition 3.1 (i) and (ii) we obtain

$$(\mathrm{L}_{U+\omega_c} - \mathrm{L}_{U+\omega})u = -\Delta u + \mathrm{B}_{U+\omega_c} u + \Delta u - \mathrm{B}_{U+\omega} u = \mathrm{B}_{\tilde{\omega}_c - \tilde{\omega}} u \quad \text{for all } u \in H(\Omega).$$

Thus, the definition of $\mathrm{B}_{\tilde{\omega}_c - \tilde{\omega}}$ (cf. (3.2)) and Lemma A.9 (i) yield

$$\begin{aligned}|(\mathrm{B}_{\tilde{\omega}_c - \tilde{\omega}} u)[\varphi]| &= Re \left| \int_\Omega ([(u \cdot \nabla)(\tilde{\omega}_c - \tilde{\omega})] \cdot \varphi + [((\tilde{\omega}_c - \tilde{\omega}) \cdot \nabla)u] \cdot \varphi) \,\mathrm{d}(x,y) \right| \\ &= Re \left| \int_\Omega (-[(u \cdot \nabla)\varphi] \cdot (\tilde{\omega}_c - \tilde{\omega}) - [((\tilde{\omega}_c - \tilde{\omega}) \cdot \nabla)\varphi] \cdot u) \,\mathrm{d}(x,y) \right| \\ &\leq 2Re \|u\|_{L^4(\Omega,\mathbb{R}^2)} \|\nabla \varphi\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \|\tilde{\omega}_c - \tilde{\omega}\|_{L^4(\Omega,\mathbb{R}^2)} \\ &\leq 2ReC_4 \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)} \|\tilde{\omega}_c - \tilde{\omega}\|_{L^4(\Omega,\mathbb{R}^2)}\end{aligned}$$

for all $\varphi \in H(\Omega)$ which, together with the first estimate, implies

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq K_c(2ReC_4\|\tilde{\omega}_c - \tilde{\omega}\|_{L^4(\Omega,\mathbb{R}^2)} \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} + \|\mathrm{L}_{U+\omega} u\|_{H(\Omega)'}) \quad \text{for all } u \in H(\Omega),$$

i.e., if $2ReC_4K_c\|\tilde{\omega}_c - \tilde{\omega}\|_{L^4(\Omega,\mathbb{R}^2)} < 1$ we obtain the norm bound

$$K := \frac{K_c}{1 - 2ReC_4K_c\|\tilde{\omega}_c - \tilde{\omega}\|_{L^4(\Omega,\mathbb{R}^2)}} \tag{6.11}$$

satisfying (A2).

Remark 6.6. *If both functions $\tilde{\omega}$ and $\tilde{\omega}_c$ are approximate solutions to the same (and by numerical evidence hopefully existing) exact solution we might expect the difference $\|\tilde{\omega}_c - \tilde{\omega}\|_{L^4(\Omega,\mathbb{R}^2)}$ to be "small", i.e., the new norm bound K is not much "larger" than K_c (cf.* (6.11)*).*

In the further course, for the reader's convenience we omit the index c again and present the computation of the desired lower eigenvalue bounds.

6.2.1 Bound for the Isolated Eigenvalues

Before having a closer look at our procedure to obtain the desired lower eigenvalue bounds for our Navier-Stokes equations, we first investigate the computation of eigenvalue bounds in an abstract setting. Therefore, let H denote a separable (complex) Hilbert space endowed with the inner product $\langle \cdot\,, \cdot \rangle$. Moreover, we suppose that $M\colon H \times H \to \mathbb{C}$ denotes a bounded, positive definite hermitian sesquilinear form on H and consider the following abstract eigenvalue problem in weak formulation

$$u \in H, \quad M(u, \varphi) = \lambda \langle u, \varphi \rangle \quad \text{for all } \varphi \in H. \tag{6.12}$$

The methods described in the further course follow the lines in [74, Section 10.2]. Note that the form of the eigenvalue problem (6.12) considered in this thesis slightly differs from the eigenvalue problem in [74, (10.45)]. Nevertheless, since we suppose M to be positive definite and hermitian, we can use M as an inner product on H to obtain an eigenvalue problem fitting into the setting considered in [74, Section 10.2] (see also [74, Remark 10.20 (a)]).

Furthermore, recall that problem (6.12) is equivalent to an eigenvalue problem for a self-adjoint operator $R\colon H \to H$. Using this equivalence, we define the essential spectrum of (6.12) via the essential spectrum of the operator R (cf. [74, Section 10.2.1]). In the further course, let $\sigma_0 \in \mathbb{R} \cup \{\infty\}$ denote the infimum of the essential spectrum of problem (6.12) or R respectively. Additionally, in the following we assume for the moment that $\sigma_0 > 0$.

Note that, in general applications the infimum of the essential spectrum might not be computable explicitly. In this case, we might replace σ_0 by a "slightly decreased", but computable, lower bound for the essential spectrum.

Now, to obtain upper bounds for the eigenvalues of problem (6.12) below the essential spectrum, the well known Rayleigh-Ritz method comes in mind (see e.g. [108, Theorem 7.2] or [109, Chapter 2]).

Theorem 6.7 (Rayleigh-Ritz)**.** *Let $n \in \mathbb{N}$ and $v_1, \dots, v_n \in H$ denote linearly independent trial functions. Moreover, define the matrices*

$$A_0 := (M(v_i, v_j))_{i,j=1,\dots,n} \quad \text{and} \quad A_1 := (\langle v_i, v_j \rangle)_{i,j=1,\dots,n} \tag{6.13}$$

and let $\hat{\lambda}_1 \leq \hat{\lambda}_2 \leq \cdots \leq \hat{\lambda}_n$ denote the eigenvalues of the matrix eigenvalue problem

$$A_0 x = \hat{\lambda} A_1 x. \tag{6.14}$$

Then, if $\hat{\lambda}_n < \sigma_0$, there are at least n eigenvalues of problem (6.12) *below σ_0 and, for the n smallest of these, denoted by $\lambda_1 \leq \cdots \leq \lambda_n$ and counted by multiplicity,*

$$\lambda_i \leq \hat{\lambda}_i \quad \textit{for all } i = 1, \ldots, n$$

holds true.

Note that the accuracy of the upper bounds provided by Theorem 6.7 strongly depends on the choice of the trial functions $v_1, \ldots, v_n$. To obtain "good" upper bounds it makes sense to use approximate eigenfunctions corresponding to the first n eigenvalues of problem (6.12) as trial functions. To compute such approximate eigenfunctions, the Rayleigh-Ritz method, as an approximation method, can be applied to a much higher number of (simple) trial functions, for instance finite element basis functions (see Section 9.5).

At a first glance this procedure seems to be inefficient, since we have to perform two separate eigenvalue computations. However, in our applications we need verified eigenvalue bounds which requires interval arithmetic calculations to enclose the eigenvalues of the matrix eigenvalue problem (6.14) (cf. Section 9.5.1). Since those interval arithmetic eigenvalue enclosures are connected to a large computational effort, we are interested in a small number of trial functions in the subsequent verification step. In contrast to that, the computation of approximate eigenfunctions starting with a huge number of trial functions needs no interval arithmetic computations at all. In our applications we use an LOBPCG method which is an iterative eigensolver implemented in the Finite Element Software M++ (cf. Section 9.5).

The next step is about the computation of lower eigenvalue bounds which in contrast to the previous explanations is a quite difficult task. To obtain the desired lower eigenvalue bounds we apply a method first introduced by Lehmann (see [58]) and Maehly (see [61]) which later was improved by Goerisch (see [9]). We want to use the following (real valued) version which can be proved using [119, Theorem 2.4] and [14, Theorem 3].

Theorem 6.8. *Let X denote a vector space, $b(\cdot,\cdot)$ some positive semi-definite symmetric bilinear form and $T\colon H \to X$ a linear operator satisfying*

$$b(T\psi, T\varphi) = M(\psi, \varphi) \quad \textit{for all } \psi, \varphi \in H. \tag{6.15}$$

Moreover, let $v_1, \ldots, v_n \in H$ be linearly independent and $w_1, \ldots, w_n$ satisfy

$$b(w_j, T\varphi) = \langle v_j, \varphi \rangle \quad \textit{for all } \varphi \in H,\ j = 1, \ldots, n. \tag{6.16}$$

In addition to A_0 and A_1 (defined in (6.13)*), set the matrix*

$$A_2 := \left(b(w_i, w_j)\right)_{i,j=1,\ldots,n}.$$

Furthermore, let some $\rho \in (0, \sigma_0]$ be chosen such that there are at most finitely many eigenvalues of (6.12) *below ρ, and such that*

$$[v \in \operatorname{span}\{v_1, \ldots, v_n\} \textit{ and } M(v, \varphi) = \rho\langle v, \varphi\rangle \textit{ for all } \varphi \in H] \ \Rightarrow\ v = 0. \tag{6.17}$$

Let $\tau_1 \leq \cdots \leq \tau_k < 0$ denote the negative eigenvalues (counted by multiplicity) of the matrix eigenvalue problem

$$(A_0 - \rho A_1)x = \tau(A_0 - 2\rho A_1 + \rho^2 A_2)x \tag{6.18}$$

(here, the matrix on the right-hand side is positive definite). Then, there are at least k eigenvalues of problem (6.12) *below ρ, and for the k largest of these, denoted by $\kappa_k^\rho \leq \kappa_{k-1}^\rho \leq \cdots \leq \kappa_1^\rho (< \rho)$ and counted by multiplicity, satisfy*

$$\kappa_j^\rho \geq \rho - \frac{\rho}{1-\tau_j} \quad \textit{for all } j = 1, \dots, k.$$

Similar to the Rayleigh-Ritz method, for Theorem 6.8 we use approximate eigenfunctions $v_1, \dots, v_n \in H$ as trial functions again. Moreover, let $\hat{\lambda}_1 \leq \cdots \leq \hat{\lambda}_n$ denote the corresponding Rayleigh-Ritz bounds for the n smallest eigenvalues of problem (6.12) provided by Theorem 6.7 using $v_1, \dots, v_n$ as trial functions. Here, the number of trial functions n is chosen such that $\hat{\lambda}_n$ is less than σ_0. If we can chose $n \geq 1$, the Rayleigh-Ritz method implies that there are at least n eigenvalues below σ_0 satisfying $\lambda_j \leq \hat{\lambda}_j$ for all $j = 1, \dots, n$. Moreover, we assume that we have some constant $\rho \in (0, \sigma_0)$ in hand such that there are at most finitely many eigenvalues of (6.12) below ρ and such that ρ satisfies

$$\hat{\lambda}_n < \rho \leq \lambda_{n+1} < \sigma_0 \tag{6.19}$$

if an $(n+1)$st eigenvalue of (6.12) below the essential spectrum exists. Otherwise, if no such eigenvalue exists, ρ needs to satisfy $\hat{\lambda}_n < \rho < \sigma_0$.

Note that, by the choice of $v_1, \dots, v_n$, the left inequality $\hat{\lambda}_n < \rho$ from above implies assumption (6.17) in Theorem 6.8. To prove this assertion let $v = \sum_{i=1}^n \beta_i v_i \in \operatorname{span}\{v_1, \dots, v_n\}$ with some $\beta \in \mathbb{R}^n$, $\beta \neq 0$ and $M(v, \varphi) = \rho \langle v, \varphi \rangle_H$ for all $\varphi \in H$. Hence, we get

$$(A_0\beta)_j = \sum_{i=1}^n \beta_i M(v_j, v_i) = M(v_j, v) = \rho \langle v_j, v \rangle_H = \rho \sum_{i=1}^n \beta_i \langle v_j, v_i \rangle_H = \rho (A_1 \beta)_j$$

for all $j = 1, \dots, n$, i.e., $A_0\beta = \rho A_1 \beta$, implying $\rho \in \{\hat{\lambda}_1, \dots, \hat{\lambda}_n\}$ (cf. Theorem 6.7) which contradicts $\hat{\lambda}_n < \rho$ (recall $\hat{\lambda}_1 \leq \cdots \leq \hat{\lambda}_n$).

Furthermore, using the assumptions above we see that the matrix $A_0 - \rho A_1$ in (6.18) is negative definite (cf. [14, p. 75]) which implies that all n eigenvalues of (6.18) are strictly negative and thus, Theorem 6.8 yields lower bounds for the n largest eigenvalues of (6.12) below ρ. However, by $\rho \leq \lambda_{n+1}$ these are also the n smallest eigenvalues of (6.12). Thus, combining these lower bounds with the upper bounds provided by the Rayleigh-Ritz method we obtain two-sided eigenvalue enclosures for the n smallest eigenvalues of (6.12).

Therefore, the successful computation of the desired two-sided eigenvalue enclosures, or especially the lower bounds, to the n smallest eigenvalues of problem (6.12) strongly depends on the a-priori information (6.19) which requires an explicit lower bound ρ for the $(n+1)$st eigenvalue of problem (6.12) to be in hand. Thus, at first glance these methods for the computation of lower eigenvalue bounds seem to be inapplicable in a general setting. Fortunately, the a-priori lower bound ρ for the $(n+1)$st eigenvalue does not necessarily need to be a very accurate bound since Theorem 6.8, applied with a rather rough bound ρ, yields very precise lower bounds anyway.

In the further course, we describe a procedure to obtain the desired a-priori information, i.e., the rough lower bound ρ, using a homotopy method (cf. [14]). Therefore, we primarily apply Theorem 6.8 with $n = 1$ which yields the following Corollary:

Corollary 6.9. *Let X, b, T as in Theorem 6.8. Moreover, let $v \in H \setminus \{0\}$ and $w \in X$ such that (6.16) holds (with v, w instead of v_j, w_j). Moreover, let $\rho \in (0, \sigma_0]$ be chosen such that there are at most finitely many eigenvalues of (6.12) below ρ, and*

$$\frac{M(v,v)}{\langle v,v\rangle} < \rho. \tag{6.20}$$

Then, there is an eigenvalue λ of problem (6.12) satisfying

$$\frac{\rho\langle v,v\rangle - M(v,v)}{\rho b(w,w) - \langle v,v\rangle} \leq \lambda < \rho. \tag{6.21}$$

Note that assumption (6.20) coincides with the first inequality in (6.19) for $n = 1$. Thus, the same arguments as before show that condition (6.20) implies assumption (6.17) from Theorem 6.8.

Finally, we are left with the computation of a rough lower bound ρ for the $(n + 1)$st eigenvalue. In many applications, such a lower can be obtained using a homotopy method together with a base problem for which the first eigenvalues are explicitly known or at least can be enclosed "easily". In the following, we describe the application of a homotopy method first introduced by Plum in [14] (see also [74, Section 10.2.4]) which reduces the computational effort compared to older homotopy versions since in each homotopy step only matrix eigenvalue problems of very small "size" (in our applications 1 or 2) need to be solved using interval arithmetic computations.

Starting from a "base problem" (cf. [74, Section 10.2.4]), we first of all assume that a bounded, positive definite hermitian sesquilinear form $M_0\colon H_0 \times H_0 \to \mathbb{C}$ defined on a second separable (complex) Hilbert space $(H_0, \langle\,\cdot\,,\,\cdot\,\rangle_0)$ is at hand such that

$$H_0 \supseteq H \quad \text{and} \quad \frac{M_0(u,u)}{\langle u,u\rangle_0} \leq \frac{M(u,u)}{\langle u,u\rangle} \quad \text{for all } u \in H \setminus \{0\}\,. \tag{6.22}$$

Moreover, we suppose that a constant $\rho_0 \in \mathbb{R}$ is known such that the following base problem

$$u \in H_0, \quad M_0(u,\varphi) = \lambda^{(0)}\langle u,\varphi\rangle_0 \quad \text{for all } \varphi \in H_0 \tag{6.23}$$

has exactly n_0 eigenvalues $\lambda_1^{(0)} \leq \cdots \leq \lambda_{n_0}^{(0)}$ counted by multiplicity in $(0, \rho_0)$ and $\rho_0 \leq \sigma_0^{(0)}$, where $\sigma_0^{(0)}$ denotes the infimum of the essential spectrum of (6.23) (which is again defined via the essential spectrum of the corresponding self-adjoint operator $R^{(0)}$).

In the sense of [74, Section 10.2.4] we additionally assume that the base problem (6.23) and problem (6.12) are homotopically connected, i.e., there exists a family $(H_t, \langle\,\cdot\,,\,\cdot\,\rangle_t)_{t\in[0,1]}$ of separable (complex) Hilbert spaces and a family $(M_t)_{t\in[0,1]}$ of bounded, positive definite hermitian sesquilinear forms $M_t\colon H_t \times H_t \to \mathbb{C}$ such that $(H_1, \langle\,\cdot\,,\,\cdot\,\rangle_1) = (H, \langle\,\cdot\,,\,\cdot\,\rangle)$ and $M_1 = M$, as well as for all $0 \leq s \leq t \leq 1$ we assume

$$H_s \supseteq H_t \quad \text{and} \quad \frac{M_s(u,u)}{\langle u,u\rangle_s} \leq \frac{M_t(u,u)}{\langle u,u\rangle_t} \quad \text{for all } u \in H_t \setminus \{0\}\,. \tag{6.24}$$

Then, for $t \in [0,1]$ we consider the family of eigenvalue problems

$$u \in H_t, \quad M_t(u,\varphi) = \lambda^{(t)}\langle u,\varphi\rangle_t \quad \text{for all } \varphi \in H_t. \tag{6.25}$$

Now, (for fixed $t \in [0,1]$) let $\lambda_1^{(t)} \leq \lambda_2^{(t)} \cdots$ denote the eigenvalues of problem (6.25) below $\sigma_0^{(t)}$ (with $\sigma_0^{(t)}$ defined again as the infimum of the essential spectrum of the associated self-adjoint operator). Then, by assumptions (6.24) Poincaré's min-max-principle (see [109, Chapter 2] and [74, Theorem 10.33]) implies for $1 \leq s \leq t \leq 1$:

$$\lambda_j^{(s)} \leq \lambda_j^{(t)} \quad \text{for all } j \in \mathbb{N} \text{ such that } \lambda_j^{(t)} < \sigma_0^{(t)} \text{ exists.} \tag{6.26}$$

Note that in the affirmative case the existence of $\lambda_j^{(t)} < \sigma_0^{(t)}$ directly implies the existence of $\lambda_j^{(s)}$ below the essential spectrum.

To start the homotopy (if $n_0 \geq 1$ holds true) we suppose that the gap between $\lambda_{n_0}^{(0)}$ and ρ_0 is not "too small". For some $t_1 > 0$, by standard (i.e., non-verified) numerical means, we compute approximate eigenpairs $(\tilde{\lambda}_j^{(t_1)}, \tilde{u}_j^{(t_1)})$ for $j = 1, \ldots, n_0$ of problem (6.25) (with $t = t_1$) such that $\tilde{\lambda}_1^{(t_1)}, \ldots, \tilde{\lambda}_{n_0}^{(t_1)}$ are ordered by magnitude. If t_1 was chosen not "too large", we might expect to find all n_0 eigenvalues below ρ_0 and we might assume that the Rayleigh quotient computed with $\tilde{u}_{n_0}^{(t_1)}$ satisfies

$$\frac{M_{t_1}(\tilde{u}_{n_0}^{(t_1)}, \tilde{u}_{n_0}^{(t_1)})}{\langle \tilde{u}_{n_0}^{(t_1)}, \tilde{u}_{n_0}^{(t_1)} \rangle_{t_1}} < \rho_0. \tag{6.27}$$

Additionally, we are interested in choosing t_1 "almost" maximal satisfying property (6.27), i.e., such that the inequality in (6.27) is almost an equality, or $t_1 = 1$ (where the argumentation further below finishes the homotopy already). In the case $t_1 < 1$ we use the approximate eigenvalues $\tilde{\lambda}_1^{(t_1)}, \ldots, \tilde{\lambda}_{n_0}^{(t_1)}$ to guess if the exact eigenvalue $\lambda^{(t_1)}$ belongs to a cluster of eigenvalues (or is a multiple eigenvalue respectively) or is a well-separated (single) eigenvalue.

In the latter case, we apply Corollary 6.9 to problem (6.25) with $t = t_1$ and with $(H, \langle \cdot, \cdot \rangle) = (H_{t_1}, \langle \cdot, \cdot \rangle_{t_1})$, $v = \tilde{u}_{n_0}^{(t_1)}$ as well as $\rho = \rho_0$ (provided an appropriate Goerisch setting X, b, T and $w_{n_0}^{(t_1)}$ is at hand for this problem as required in Corollary 6.9) and obtain that there exists an eigenvalue $\lambda^{(t_1)}$ of the given problem (for $t = t_1$) with

$$\rho_1 := \frac{\rho_0 \langle \tilde{u}_{n_0}^{(t_1)}, \tilde{u}_{n_0}^{(t_1)} \rangle_{t_1} - M_{t_1}(\tilde{u}_{n_0}^{(t_1)}, \tilde{u}_{n_0}^{(t_1)})}{\rho_0 b(w_{n_0}^{(t_1)}, w_{n_0}^{(t_1)}) - \langle \tilde{u}_{n_0}^{(t_1)}, \tilde{u}_{n_0}^{(t_1)} \rangle_{t_1}} \leq \lambda^{(t_1)} < \rho_0 \tag{6.28}$$

(cf. (6.21)), where $w_{n_0}^{(t_1)}$ satisfies $b(T\varphi, w_{n_0}^{(t_1)}) = \langle \varphi, \tilde{u}_{n_0}^{(t_1)} \rangle_H$ for all $\varphi \in H$ (cf. assumptions of Corollary 6.9). Moreover, since we know by assumption that the base problem has precisely n_0 eigenvalues in the interval $(0, \rho_0)$, the monotonicity of the eigenvalues (cf. (6.26)) implies that problem (6.25) with $t = t_1$ has at most n_0 eigenvalues in $(0, \rho_0)$, which together with the enclosure in (6.28) yields:

$$\text{problem (6.25) with } t = t_1 \text{ has at most } n_0 - 1 \text{ eigenvalues in } (0, \rho_1).$$

If the approximate eigenfunction $\tilde{u}_{n_0}^{(t_1)}$ is computed with sufficient accuracy, the structure of ρ_1 implies that ρ_1 is not "far below" ρ_0. Hence, if the gap between $\lambda_{n_0}^{(t_1)}$ and $\lambda_{n_0-1}^{(t_1)}$ is not too small (which is assumed in the current case), we expect $\lambda_{n_0}^{(t_1)}$ to be the only eigenvalue in $[\rho_1, \rho_0)$, and thus, that problem (6.25) with $t = t_1$ has exactly $n_0 - 1$ eigenvalues in $(0, \rho_1)$.

In the case where $\lambda_{n_0}^{(t_1)}$ seems to be part of an eigenvalue cluster or appears to be an eigenvalue with higher multiplicity, instead of Corollary 6.9 we apply Theorem 6.8 directly with $n = n_c \geq 2$ being the expected size of the cluster. Then, we get lower bounds for the eigenvalues $\lambda_{n_0-n_c+1}^{(t_1)}, \ldots, \lambda_{n_0}^{(t_1)}$ and by ρ_1 we denote the (lower) bound for the smallest of these, i.e., $\rho_1 \leq \lambda_{n_0-n_c+1}^{(t_1)}$. Again, since the base problem has exactly n_0 eigenvalues in $(0, \rho_0)$, the monotonicity of the eigenvalues (see (6.26)) implies that problem (6.25) with $t = t_1$ has at most $n_0 - n_c$ eigenvalues in $(0, \rho_1)$. If furthermore $\lambda_{n_0-n_c}^{(t_1)}$ and $\lambda_{n_0-n_c+1}^{(t_1)}$ are well separated (we expect this to be the case since the cluster size was assumed to be n_c) and ρ_1 is not too far below $\lambda_{n_0-n_c+1}^{(t_1)}$, we might expect that the only eigenvalues in $[\rho_1, \rho_0)$ are $\lambda_{n_0-n_c+1}^{(t_1)}, \ldots, \lambda_{n_0}^{(t_1)}$ and thus, that problem (6.25) with $t = t_1$ has exactly $n_0 - n_c$ eigenvalues in $(0, \rho_1)$.

Hence, with

$$n_1 := \begin{cases} 1, & \text{if } \lambda_{n_0}^{(t_1)} \text{ and } \lambda_{n_0-1}^{(t_1)} \text{ are well separated,} \\ n_c, & \text{otherwise,} \end{cases}$$

we obtain that problem (6.25) with $t = t_1$ has at most $n_0 - n_1$ eigenvalues in $(0, \rho_1)$, and we expect that it has exactly $n_0 - n_1$ eigenvalues in $(0, \rho_1)$. Already at this stage we could check if this expectation is true using a Rayleigh-Ritz computation, but this is not necessary. We simply continue on the basis of this expectation, and perform a final Rayleigh-Ritz computation at the end of the homotopy method which proves the expectations a posteriori or shows that the homotopy was not successful (cf. [74, Section 10.2.4]).

Remark 6.10. *To the best of our knowledge there exists no analytical theory for the computation of an "optimal" new homotopy parameter t_1. Nevertheless, in Section 9.5.3 we present an algorithm to obtain the desired homotopy parameter using approximate eigenvalues.*

In the second homotopy step (taking place if $n_0 - n_1 \geq 1$ and $t_1 < 1$), we repeat the procedure above with 0 replaced by t_1 and ρ_0 replaced by ρ_1: For some $t_2 > t_1$ we again

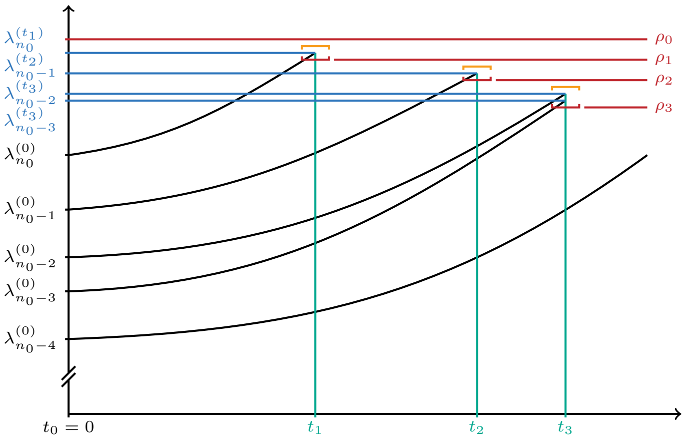


Figure 6.1: Possible course of the first homotopy steps

compute approximate eigenpairs $(\tilde{\lambda}_j^{(t_2)}, \tilde{u}_j^{(t_2)})$ for $j = 1, \ldots, n_0 - n_1$ of problem (6.25) (with $t = t_2$) such that

$$\frac{M_{t_2}(\tilde{u}_{n_0-n_1}^{(t_2)}, \tilde{u}_{n_0-n_1}^{(t_2)})}{\langle \tilde{u}_{n_0-n_1}^{(t_2)}, \tilde{u}_{n_0-n_1}^{(t_2)} \rangle_{t_2}} < \rho_1,$$

and such that t_2 is "almost" maximal in the sense described above. The same arguments as in the previous step imply the existence of a constant

$$n_2 := \begin{cases} n_1 + 1, & \text{if } \lambda_{n_0-n_1}^{(t_2)} \text{ and } \lambda_{n_0-n_1-1}^{(t_2)} \text{ are well separated,} \\ n_1 + n_c, & \text{otherwise,} \end{cases}$$

where n_c denotes the size of an eigenvalue cluster which $\lambda_{n_0-n_1}^{(t_2)}$ possibly belongs to. Then either Corollary 6.9 or Theorem 6.8 respectively, yields a lower bound ρ_2 (cf. (6.28)) such that there are at least $n_2 - n_1$ eigenvalues of problem (6.25) with $t = t_2$ in $[\rho_2, \rho_1)$, and hence, we conclude

$$\text{problem (6.25) with } t = t_2 \text{ has at most } n_0 - n_2 \text{ eigenvalues in } (0, \rho_2).$$

As in the first step, we expect that problem (6.25) with $t = t_2$ has exactly $n_0 - n_2$ eigenvalues in $(0, \rho_2)$.

We go on with this algorithm until for some $r \in \mathbb{N}_0$ either $t_r = 1$ and $n_r \leq n_0$ or $t_r < 1$ and $n_r = 0$ (in which case the homotopy cannot be continued). For the case $t_r = 1$ the same procedure as above yields that

$$\text{problem (6.25) with } t = t_r = 1 \text{ has at most } n := n_0 - n_r \text{ eigenvalues in } (0, \rho_r). \quad (6.29)$$

Hence $\rho := \rho_r$ is a lower bound for the $(n+1)$st eigenvalue of (6.12), but, if $n \geq 1$, it is possibly smaller than the n-th eigenvalue.

Thus, if $t_r = 1$ and $n \geq 1$, we start a (final) Rayleigh-Ritz computation for problem (6.12) and check, if $\hat{\lambda}_n < \rho$ (cf. (6.19)) holds true (it will be satisfied if all our expectations during the homotopy steps are correct). If this check is successful, we obtain that problem (6.12) has at least n eigenvalues in $(0, \rho)$ implying, together with (6.29), that problem (6.12) has precisely n eigenvalues in $(0, \rho)$. Finally, having ρ in hand, by Theorem 6.8 we can compute the desired lower bounds for the n smallest eigenvalues of (6.12).

Finally, if $t_r < 1$ and $n = 0$ the homotopy needs to be restarted with larger values of n_0 and ρ_0.

Remark 6.11. (i) *Introductory examples to the homotopy method presented above can be found for instance in [74, Section 10.2.5]. For more examples we refer the reader to several computer-assisted works (like [14], [82], [117], ...) where the homotopy method was applied successfully.*

(ii) *As mentioned earlier, at several stages the homotopy method heavily relies on the possibility of computing verified eigenvalue enclosures of matrix eigenvalue problems (cf. Theorem 6.7, Theorem 6.8 and Corollary 6.9). For more details about the computation of these verified enclosures we refer the reader to Section 9.5.1.*

(iii) *If the base problem* (6.23) *is not "too far away" from our problem* (6.12) *it can directly be used as a comparison problem, i.e., our homotopy then consists of one single step (cf. [74, beginning of Section 10.2.4]), i.e., if*

$$\hat{\lambda}_{n_0} < \lambda^{(0)}_{n_0+1} \leq \lambda_{n_0+1}$$

(with $\hat{\lambda}_{n_0}$ denoting the upper bound for the nth eigenvalue of problem (6.12) *provided by the Rayleigh-Ritz method) holds true we can directly use $\rho := \lambda^{(0)}_{n_0+1}$ as the desired "rough" lower bound for the* $(n+1)$*st eigenvalue (of problem* (6.12)*) required in Theorem 6.8.*

In the further course, we apply the methods presented above to our Navier-Stokes equations to obtain the desired eigenvalue enclosures which then, together with the information about the essential spectrum (cf. Section 6.2.2), results in the required norm bounds K and K^* respectively.

6.2.1.1 Eigenvalue Bounds

Now, we apply the eigenvalue methods presented in the beginning of this Section to treat the eigenvalue problems (6.8) and (6.10) respectively. To be able to apply the algorithms we first introduce shift parameters $\nu > 0$ and $\hat{\nu} > 0$ respectively which results in the following shifted eigenvalue problems

$$\begin{aligned} u \in H(\Omega),\ &\langle \Phi^{-1} \mathrm{L}_{U+\omega}\, u, \Phi^{-1} \mathrm{L}_{U+\omega}\, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\ &+ \nu \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \lambda_\nu \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega) \end{aligned} \tag{6.30}$$

and

$$\begin{aligned} u \in H(\Omega),\ &\langle \Phi^{-1} \hat{\mathrm{L}}_{U+\omega}\, u, \Phi^{-1} \hat{\mathrm{L}}_{U+\omega}\, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\ &+ \hat{\nu} \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \lambda_{\hat{\nu}} \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega) \end{aligned} \tag{6.31}$$

with $\lambda_\nu := \lambda + \nu$ and $\lambda_{\hat{\nu}} := \lambda + \hat{\nu}$ denoting the new (shifted) eigenvalue parameters.

Obviously, due to $\nu > 0$ and $\hat{\nu} > 0$ respectively the left-hand side of the eigenvalue problems (6.30) and (6.31) respectively defines a bounded, positive definite and symmetric bilinear form on $H(\Omega)$, i.e., an inner product on $H(\Omega)$, which is a crucial assumption for the eigenvalue methods introduced above.

In the sense of the abstract setting presented at the beginning of this Section we define the left-hand side of (6.30) by

$$M\colon H(\Omega) \times H(\Omega) \to \mathbb{R},\ M(u,\varphi) := \langle \Phi^{-1} \mathrm{L}_{U+\omega}\, u, \Phi^{-1} \mathrm{L}_{U+\omega}\, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \nu \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)},$$

i.e., our (shifted) eigenvalue problem (6.30) is now of the form (6.12) considered in the abstract setting. Analogously, with the definition

$$\hat{M}\colon H(\Omega) \times H(\Omega) \to \mathbb{R},\ \hat{M}(u,\varphi) := \langle \Phi^{-1} \hat{\mathrm{L}}_{U+\omega}\, u, \Phi^{-1} \hat{\mathrm{L}}_{U+\omega}\, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \hat{\nu} \langle u, \varphi \rangle_{H_0^1(\Omega,\mathbb{R}^2)}$$

the second eigenvalue problem is of the form (6.12) as well.

Since both problems (6.30) and (6.31) are of the same form, in the following we restrict our explanations mainly to the first eigenvalue problem (6.30), whereas the second problem can be treated mutatis mutandis. Only, at some stages we point out some fundamental differences between both cases.

Upper Eigenvalue Bounds

At first glance, the computation of upper eigenvalue bounds seems redundant since all eigenvalues of our problems (6.8) and (6.9) are positive and we are interested to bound the eigenvalues away from zero (cf. beginning of this Section). Nevertheless, to ensure that the Lehmann-Goerisch method indeed encloses the lowest (starting from the first) eigenvalues we need index information which is provided by the Rayleigh-Ritz Theorem 6.7.

Now, let $\tilde{u}_1, \dots, \tilde{u}_n \in H(\Omega)$ denote approximate eigenfunctions to the first n eigenvalues of the shifted problem (6.30) (which are also eigenfunctions to the first n eigenvalues of problem (6.8)). Thus, applying Theorem 6.7 we obtain upper bounds for the first n eigenvalues of problem (6.30) if they exist below the essential spectrum (or below σ_0 respectively). Having a closer look at Theorem 6.7 we see that the matrices

$$\begin{aligned} A_0 &= (M(\tilde{u}_i, \tilde{u}_j))_{i,j=1,\dots,n} \\ &= \big(\langle \Phi^{-1}\,\mathrm{L}_{U+\omega}\,\tilde{u}_i, \Phi^{-1}\,\mathrm{L}_{U+\omega}\,\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \nu\langle \tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\big)_{i,j=1,\dots,n} \end{aligned} \tag{6.32}$$

and $A_1 := \big(\langle \tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\big)_{i,j=1,\dots,n}$ have to be computed with the approximate eigenfunctions used as (linearly independent) trial functions. Obviously, since $\Phi^{-1}\,\mathrm{L}_{U+\omega}\,\tilde{u}_i$ is not explicitly known for all $i = 1, \dots, n$ the computation of the matrix A_0 (in contrast to the computation of A_1) is not possible directly. Thus, Rayleigh-Ritz' Theorem 6.7 cannot be applied in our situation without further effort.

To overcome these difficulties and to obtain upper eigenvalue bounds for our eigenvalue problem (6.30) anyway, we follow the lines by Plum in [74, Section 9.4.1.2]. As described in [74], replacing the matrix A_0 in Theorem 6.7 by any hermitian matrix $\tilde{A}_0$ such that $\tilde{A}_0 - A_0$ is positive semi-definite, the Rayleigh-Ritz Theorem 6.7 still provides upper bounds for the eigenvalues of the shifted problem (6.30). Taking a look at the proof of the Rayleigh-Ritz method we see that the new bounds are not much worse than the original ones if the difference $\tilde{A}_0 - A_0$ is “small” (cf. [74, Section 9.4.1.2]). Hence, in the further course we present one possibility to replace A_0 by a suitable hermitian matrix $\tilde{A}_0$ for which its entries are computable explicitly.

Following the lines in [74, Section 9.4.1.2] we use the definition of $\mathrm{L}_{U+\omega}$ (see (3.9)) and Φ (see (2.14)) to rewrite the entries of A_0 defined in (6.32) as follows:

$$\begin{aligned} (A_0)_{i,j} &= \langle \Phi^{-1}((-\Delta+\sigma)\tilde{u}_i + (\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i), \Phi^{-1}((-\Delta+\sigma)\tilde{u}_j + (\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j)\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\ &\quad + \nu\langle \tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\ &= (1+\nu)\langle \tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\ &\quad + \langle \tilde{u}_i, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \end{aligned}$$

for all $i, j = 1, \dots, n$.

Applying (2.16) and inserting the definition of $\mathrm{B}_{U+\omega}$ (see (3.2)) the terms containing Φ

only on one side of the inner product can equivalently be written as

$$\begin{aligned}
&\langle \tilde{u}_i, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad = ((\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j)[\tilde{u}_i] + ((\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i)[\tilde{u}_j] \\
&\quad = \int_\Omega (Re[(\tilde{u}_j \cdot \nabla)(U+\omega) + ((U+\omega)\cdot\nabla)\tilde{u}_j] - \sigma\tilde{u}_j) \cdot \tilde{u}_i \,\mathrm{d}(x,y) \\
&\qquad\quad + \int_\Omega (Re[(\tilde{u}_i \cdot \nabla)(U+\omega) + ((U+\omega)\cdot\nabla)\tilde{u}_i] - \sigma\tilde{u}_i) \cdot \tilde{u}_j \,\mathrm{d}(x,y).
\end{aligned}$$

Moreover, using Lemma A.10 (i) we calculate

$$\int_\Omega [((U+\omega)\cdot\nabla)\tilde{u}_j]\cdot\tilde{u}_i \,\mathrm{d}(x,y) = -\int_\Omega [((U+\omega)\cdot\nabla)\tilde{u}_i]\cdot\tilde{u}_j \,\mathrm{d}(x,y),$$

which together with the formulations above implies

$$\langle \tilde{u}_i, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \int_\Omega \tilde{u}_i^T \,\mathrm{G}_{U+\omega}\,\tilde{u}_j \,\mathrm{d}(x,y).$$

where we use the abbreviation

$$\mathrm{G}_{U+\omega} := Re[\nabla(U+\omega) + (\nabla(U+\omega))^T] - 2\sigma\,\mathrm{id}\,. \tag{6.33}$$

Altogether, we obtain the following representation for the entries of A_0:

$$\begin{aligned}
(A_0)_{i,j} = M(\tilde{u}_i,\tilde{u}_j) = (1+\nu)\langle \tilde{u}_i,\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \int_\Omega \tilde{u}_i^T \,\mathrm{G}_{U+\omega}\,\tilde{u}_j \,\mathrm{d}(x,y) \\
+ \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)}
\end{aligned} \tag{6.34}$$

for all $i,j = 1,\ldots,n$.

With the matrix

$$D := \left(\langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\right)_{i,j=1,\ldots,n}$$

introduced in [74, Section 9.4.1.2] as well, [74, Lemma 9.22] reads as follows:

Lemma 6.12. *For $i = 1,\ldots,n$ let $\rho_i^{(x)}, \rho_i^{(y)} \in L^2(\Omega,\mathbb{R}^2)$ with $\frac{\partial\rho_i^{(x)}}{\partial x} + \frac{\partial\rho_i^{(y)}}{\partial y} \in L^2(\Omega,\mathbb{R}^2)$ and define the matrix*

$$\begin{aligned}
\tilde{D} := \Bigg(&\langle \rho_i^{(x)}, \rho_j^{(x)}\rangle_{L^2(\Omega,\mathbb{R}^2)} + \langle \rho_i^{(y)}, \rho_j^{(y)}\rangle_{L^2(\Omega,\mathbb{R}^2)} \\
&+ \frac{1}{\sigma}\left\langle \frac{\partial\rho_i^{(x)}}{\partial x} + \frac{\partial\rho_i^{(y)}}{\partial y} + (\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \frac{\partial\rho_j^{(x)}}{\partial x} + \frac{\partial\rho_j^{(y)}}{\partial y} + (\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j \right\rangle_{L^2(\Omega,\mathbb{R}^2)} \Bigg)_{i,j=1,\ldots,n}.
\end{aligned}$$

Then the following assertions hold true:

(i) *$\tilde{D} - D$ is positive semi-definite.*

(ii) *$\tilde{D} = D$ if $\rho_i^{(z)} := \frac{\partial}{\partial z}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i \in L^2(\Omega,\mathbb{R}^2)$ for all $i = 1,\ldots,n$ and $z \in \{x,y\}$. Note that for this choice we have*

$$\frac{\partial\rho_i^{(x)}}{\partial x} + \frac{\partial\rho_i^{(y)}}{\partial y} = \sigma\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i - (\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i \in L^2(\Omega,\mathbb{R}^2) \text{ for all } i = 1,\ldots,n.$$

Proof. (i) The proof is similar to [74, Proof of Lemma 9.22].

(ii) Let $\rho_i^{(z)} := \frac{\partial}{\partial z}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i \in L^2(\Omega, \mathbb{R}^2)$ for all $i = 1, \dots, n$ and $z \in \{x, y\}$. Then, using the definition of Φ (see (2.14)) we calculate

$$\begin{aligned}\frac{\partial \rho_i^{(x)}}{\partial x} + \frac{\partial \rho_i^{(y)}}{\partial y} &= \frac{\partial^2}{\partial x^2}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i + \frac{\partial^2}{\partial y^2}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i \\ &= (\sigma - (-\Delta + \sigma))\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i \\ &= \sigma\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i - (\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i\end{aligned}$$

for all $i = 1, \dots, n$. Moreover, using this equality and the definition of the inner product on $H(\Omega)$ (see (2.5)) we obtain

$$\begin{aligned}(\tilde{D})_{i,j} &= \left\langle \frac{\partial}{\partial x}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \frac{\partial}{\partial x}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j \right\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &\quad + \left\langle \frac{\partial}{\partial y}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \frac{\partial}{\partial y}\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j \right\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &\quad + \frac{1}{\sigma}\langle \sigma\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \sigma\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j \rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &= \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_j \rangle_{H_0^1(\Omega,\mathbb{R}^2)} = D_{i,j}\end{aligned}$$

for all $i, j = 1, \dots, n$ which proves the assertion.

□

Thus, Lemma 6.12 together with definition (6.34) implies that for

$$(\tilde{A}_0)_{i,j} := (1 + \nu)\langle \tilde{u}_i, \tilde{u}_j \rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \int_\Omega \tilde{u}_i^T \, \mathrm{G}_{U+\omega} \, \tilde{u}_j \, \mathrm{d}(x, y) + \tilde{D}_{i,j}$$

for all $i, j = 1, \dots, n$ (with $\tilde{D}$ defined as in Lemma 6.12) the difference $\tilde{A}_0 - A_0$ is positive semi-definite for arbitrary $\rho_i^{(x)}, \rho_i^{(y)} \in L^2(\Omega, \mathbb{R}^2)$ such that $\frac{\partial \rho_i^{(x)}}{\partial x} + \frac{\partial \rho_i^{(y)}}{\partial y} \in L^2(\Omega, \mathbb{R}^2)$. Hence, as mentioned above performing a Rayleigh-Ritz computation with A_0 replaced by $\tilde{A}_0$ we obtain larger but computable upper eigenvalue bounds for our eigenvalue problem (6.30).

To get only “slightly” larger bounds we have to chose $\tilde{A}_0$, i.e., $\rho_i^{(x)}, \rho_i^{(y)} \in L^2(\Omega, \mathbb{R}^2)$, such that the difference $\tilde{A}_0 - A_0$ is “small”. Therefore, we use the strategy presented by Plum in [74, p. 338], i.e., we approximately minimize the diagonal entries $\tilde{D}_{i,i}$ for all $i = 1, \dots, n$ over all admissible $\rho_i^{(x)}, \rho_i^{(y)} \in L^2(\Omega, \mathbb{R}^2)$ such that $\frac{\partial \rho_i^{(x)}}{\partial x} + \frac{\partial \rho_i^{(y)}}{\partial y} \in L^2(\Omega, \mathbb{R}^2)$. However, in our minimization process the space of admissible $\rho_i^{(x)}, \rho_i^{(y)} \in L^2(\Omega, \mathbb{R}^2)$ is restricted to a suitable subspace, i.e., we minimize the functionals $J_i \colon H(\mathrm{div}, \Omega, \mathbb{R}^{2\times 2}) \to \mathbb{R}$ defined by

$$J_i\left[(\rho_i^{(x)}, \rho_i^{(y)})\right] := \left\|\rho_i^{(x)}\right\|^2_{L^2(\Omega,\mathbb{R}^2)} + \left\|\rho_i^{(y)}\right\|^2_{L^2(\Omega,\mathbb{R}^2)} + \frac{1}{\sigma}\left\|\frac{\partial \rho_i^{(x)}}{\partial x} + \frac{\partial \rho_i^{(y)}}{\partial y} + (\mathrm{B}_{U+\omega} - \sigma)\tilde{u}_i\right\|^2_{L^2(\Omega,\mathbb{R}^2)}$$

for all $i = 1, \dots, n$. We note that with the choice $H(\mathrm{div}, \Omega, \mathbb{R}^{2\times 2})$ as domain for the functionals J_i the assumption $\frac{\partial \rho_i^{(x)}}{\partial x} + \frac{\partial \rho_i^{(y)}}{\partial y} \in L^2(\Omega, \mathbb{R}^2)$ required for Lemma 6.12 is satisfied by construction. In practice, in our computations we actually choose an appropriate (finite-dimensional) finite element subspace of $H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2})$, for instance, we can use Raviart Thomas finite elements in our approximation procedure.

We close this paragraph with some remarks on our second (shifted) eigenvalue problem (6.31). Similar calculations as above but now using the definition of $\hat{\mathrm{B}}_{U+\omega}$ (see (6.2)) yield

$$\langle \tilde{u}_i, \Phi^{-1}(\hat{\mathrm{B}}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}(\hat{\mathrm{B}}_{U+\omega} - \sigma)\tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \int_\Omega \tilde{u}_i^T \, \mathrm{G}_{U+\omega} \, \tilde{u}_j \, \mathrm{d}(x,y).$$

Thus, in the adjoint case the matrix A_0 defined in (6.34) is replaced by the following definition:

$$\begin{aligned}(A_0)_{i,j} = (1+\nu)\langle \tilde{u}_i, \tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)} &+ \int_\Omega \tilde{u}_i^T \, \mathrm{G}_{U+\omega} \, \tilde{u}_j \, \mathrm{d}(x,y) \\ &+ \langle \Phi^{-1}(\hat{\mathrm{B}}_{U+\omega} - \sigma)\tilde{u}_i, \Phi^{-1}(\hat{\mathrm{B}}_{U+\omega} - \sigma)\tilde{u}_j\rangle_{H_0^1(\Omega,\mathbb{R}^2)}.\end{aligned} \tag{6.35}$$

Finally, we can adapt Lemma 6.12 suitably to treat our second problem (6.31) similarly to the first one (see (6.30)).

Remark 6.13. *Having computed upper bounds $(\overline{\lambda}_\nu)_1, \dots, (\overline{\lambda}_\nu)_n$ for the shifted problem* (6.30) *(with eigenvalues $(\lambda_\nu)_1, \dots, (\lambda_\nu)_n$) the retransformation $\lambda = \lambda_\nu - \nu$ yields associated upper bounds $\overline{\lambda}_1, \dots, \overline{\lambda}_n$ for the original problem* (6.8)*, i.e., $\overline{\lambda}_i = (\overline{\lambda}_\nu)_i - \nu$ for all $i = 1, \dots, n$. Similarly, we obtain upper bounds for the eigenvalues of the second eigenvalue problem* (6.9) *provided upper bounds for the shifted problem* (6.31) *are computed already.*

Lower Eigenvalue Bounds

To obtain the desired lower eigenvalue bounds for both problems (6.30) and (6.31) respectively using the methods presented at the beginning of this Subsection we have to find an appropriate Goerisch setting for Theorem 6.8 (cf. "XbT-setting" introduced in [74, Lemma 10.25]).

First of all, we use similar computations as above (cf. (6.34) with $\tilde{u}_i, \tilde{u}_j$ replaced by arbitrary $u, \varphi \in H(\Omega)$) to obtain an equivalent formulation for the left-hand side of our shifted eigenvalue problem (6.30):

$$\begin{aligned}M(u,\varphi) = (1+\nu)\langle u, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} &+ \int_\Omega u^T \, \mathrm{G}_{U+\omega} \, \varphi \, \mathrm{d}(x,y) \\ &+ \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)u, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u, \varphi \in H(\Omega),\end{aligned} \tag{6.36}$$

where we use the same abbreviation as in the previous Section (cf. (6.33)).

Thus, in the sense of the Goerisch extension of Temple-Lehmann's method (cf. Theorem 6.8) we choose $X := L^2(\Omega, \mathbb{R}^{2\times 2}) \times L^2(\Omega, \mathbb{R}^2) \times H(\Omega) \times L^2(\Omega, \mathbb{R}^2)$. Moreover, the equality (6.36) suggests to define the bilinear form $b \colon X \times X \to \mathbb{R}$ needed in Theorem 6.8 as follows

$$\begin{aligned}b(w, \hat{w}) := (1+\nu)&\langle w_1, \hat{w}_1\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \nu\sigma\langle w_2, \hat{w}_2\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &+ \int_\Omega w_2^T \, \mathrm{G}_{U+\omega} \, \hat{w}_2 \, \mathrm{d}(x,y) + \langle w_3, \hat{w}_3\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \sigma\langle w_4, \hat{w}_4\rangle_{L^2(\Omega,\mathbb{R}^2)}.\end{aligned} \tag{6.37}$$

Since $\mathrm{G}_{U+\omega}$ is symmetric we directly obtain that b defines a symmetric bilinear form on X. Having a closer look at the definition of b we see that b is positive semi-definite if we have

$$\nu\sigma\langle v, v\rangle_{L^2(\Omega,\mathbb{R}^2)} + \int_\Omega v^T \mathrm{G}_{U+\omega} v \,\mathrm{d}(x,y) \geq 0 \quad \text{for all } v \in L^2(\Omega, \mathbb{R}^2). \tag{6.38}$$

At this stage, we note that we have to choose $\sigma > 0$ because otherwise (if the first term vanishes) the inequality above might be wrong. Therefore, in the further course we assume $\sigma > 0$. Then, using the definition of $\mathrm{G}_{U+\omega}$ (see (6.33)) and Lemma A.9 (ii) we calculate

$$\begin{aligned}\int_\Omega v^T \mathrm{G}_{U+\omega} v \,\mathrm{d}(x,y) &= 2\left(Re\int_\Omega v^T(\nabla(U+\omega))v \,\mathrm{d}(x,y) - \sigma\langle v, v\rangle_{L^2(\Omega,\mathbb{R}^2)}\right) \\ &\geq -2\left(Re\|\nabla(U+\omega)\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})} + \sigma\right)\langle v, v\rangle_{L^2(\Omega,\mathbb{R}^2)}\end{aligned} \tag{6.39}$$

for all $v \in L^2(\Omega, \mathbb{R}^2)$. Hence, (6.38) is satisfied and thus b is positive semi-definite on X if we fix our shift parameter ν such that

$$\nu \geq 2\left(\frac{Re}{\sigma}\|\nabla(U+\omega)\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})} + 1\right) > 0.$$

Finally, the operator

$$T\colon H(\Omega) \to X,\ Tu := (\nabla u, u, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)u, u)^T \tag{6.40}$$

completes our Goerisch setting for the computation of lower eigenvalue bounds for problem (6.30). Moreover, using (6.36) we calculate

$$\begin{aligned}b(Tu, T\varphi) &= (1+\nu)\langle\nabla u, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \nu\sigma\langle u, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} + \int_\Omega u^T \mathrm{G}_{U+\omega}\varphi \,\mathrm{d}(x,y) \\ &\quad + \langle\Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)u, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \sigma\langle u, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &= M(u, \varphi) \quad \text{for all } u, \varphi \in H(\Omega),\end{aligned}$$

hence, condition (6.15) in Theorem 6.8 is satisfied.

The remaining task required to apply Theorem 6.8 to our eigenvalue problem is the computation of functions $w_i = ((w_i)_1, (w_i)_2, (w_i)_3, (w_i)_4)^T \in X$ for each $i = 1, \ldots, n$ such that $b(w_i, T\varphi) = \langle\tilde{u}_i, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)}$ for all $\varphi \in H(\Omega)$ (cf. (6.16)). Thus, in the following we fix $i \in \{1, \ldots, n\}$ and omit the index i in the further course, i.e., condition (6.16) now reads as

$$\begin{aligned}&(1+\nu)\langle w_1, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \nu\sigma\langle w_2, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} + \int_\Omega w_2^T \mathrm{G}_{U+\omega}\varphi \,\mathrm{d}(x,y) \\ &\quad + \langle w_3, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \sigma\langle w_4, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} = \langle\tilde{u}, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega).\end{aligned} \tag{6.41}$$

Before we can solve the equation above for one of the components, using (2.16), the definition of $\mathrm{B}_{U+\omega}$ (see (3.2)), Lemma A.10 (ii) and the definition of $\hat{\mathrm{B}}_{U+\omega}$ (see (6.2)) we calculate (cf. (6.1))

$$\begin{aligned}&\langle w_3, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\ &\quad = ((\mathrm{B}_{U+\omega} - \sigma)\varphi)[w_3] = \langle(\mathrm{B}_{U+\omega} - \sigma)\varphi, w_3\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &\quad = \int_\Omega (Re[(\varphi\cdot\nabla)(U+\omega) + ((U+\omega)\cdot\nabla)\varphi] - \sigma\varphi)\cdot w_3 \,\mathrm{d}(x,y) \\ &\quad = \int_\Omega \varphi^T\left(Re[(\nabla(U+\omega))^T w_3 - (\nabla w_3)(U+\omega)] - \sigma w_3\right)\mathrm{d}(x,y) \\ &\quad = \langle(\hat{\mathrm{B}}_{U+\omega} - \sigma)w_3, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega).\end{aligned} \tag{6.42}$$

Inserting this result into (6.41) and using the symmetry of $\mathrm{G}_{U+\omega}$ we obtain the following condition for w:

$$\begin{aligned}0 = &\langle (1+\nu)w_1 - \nabla\tilde{u}, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} \\ &+ \langle \nu\sigma w_2 + \mathrm{G}_{U+\omega}\, w_2 + (\hat{\mathrm{B}}_{U+\omega} - \sigma)w_3 + \sigma w_4 - \sigma\tilde{u}, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega)\end{aligned} \tag{6.43}$$

In the further course we consider the new variable $w5 \in H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2})$, i.e., $w_5 \in L^2(\Omega, \mathbb{R}^{2\times 2})$ such that the "row-wise" divergence is an element of $L^2(\Omega, \mathbb{R}^2)$ (c.f. (2.8)) . In turn, we fix

$$w_1 := \frac{1}{1+\nu}(w_5 + \nabla\tilde{u}) \tag{6.44}$$

and thus, using integration by parts we obtain

$$\begin{aligned}\langle (1+\nu)w_1 - \nabla\tilde{u}, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} &= \langle w_5, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} \\ &= \int_\Omega (w_5)_1 \cdot \nabla\varphi_1 \,\mathrm{d}(x,y) + \int_\Omega (w_5)_2 \cdot \nabla\varphi_2 \,\mathrm{d}(x,y) \\ &= -\int_\Omega \operatorname{div}(w_5)_1\varphi_1 \,\mathrm{d}(x,y) - \int_\Omega \operatorname{div}(w_5)_2\varphi_2 \,\mathrm{d}(x,y) \\ &= -\langle \operatorname{div} w_5, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)}\end{aligned} \tag{6.45}$$

for all $\varphi \in H(\Omega)$ where $(w_5)_j$ denotes the j-th row of w_5 (cf. Section 2.1). This equality together with condition (6.43) implies

$$0 = \langle -\operatorname{div} w_5 + \nu\sigma w_2 + \mathrm{G}_{U+\omega}\, w_2 + (\hat{\mathrm{B}}_{U+\omega} - \sigma)w_3 + \sigma w_4 - \sigma\tilde{u}, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)}$$

for all $\varphi \in H(\Omega)$ which can be "solved" for the fourth component, i.e., we choose

$$w_4 := \frac{1}{\sigma}\Big(\operatorname{div} w_5 - \nu\sigma w_2 - \mathrm{G}_{U+\omega}\, w_2 - (\hat{\mathrm{B}}_{U+\omega} - \sigma)w_3 + \sigma\tilde{u}\Big). \tag{6.46}$$

Finally, for arbitrary components $w_2 \in L^2(\Omega, \mathbb{R}^2)$, $w_3 \in H(\Omega)$ and $w_5 \in L^2(\Omega, \mathbb{R}^2)$ such that $\frac{\partial (w_5)_1}{\partial x} + \frac{\partial (w_5)_2}{\partial y} \in L^2(\Omega, \mathbb{R}^2)$ Theorem 6.8 provides lower bounds for our eigenvalue problem (6.30). However, to obtain "good" bounds we follow the lines in [74, Remark 10.26 (c)], i.e., we approximately minimize $b(w, w)$ over w_2, w_3 and w_5 in a suitable finite element subspace. Here, b is defined by (6.37) where the choices for w_1 (see (6.44)) and w_4 (see (6.46)) are used to obtain a functional only in the variables w_2, w_3 and w_5. This minimization strategy will be discussed in the further course.

Therefore, inserting the definitions of w_1 and w_4 into the bilinear form b (cf. (6.37)) we obtain

$$\begin{aligned}b(w,w) = &\frac{1}{1+\nu}\|w_5 + \nabla\tilde{u}\|^2_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \nu\sigma\|w_2\|^2_{L^2(\Omega,\mathbb{R}^2)} \\ &+ \int_\Omega w_2^T\, \mathrm{G}_{U+\omega}\, w_2 \,\mathrm{d}(x,y) + \|w_3\|^2_{H_0^1(\Omega,\mathbb{R}^2)} \\ &+ \frac{1}{\sigma}\|\operatorname{div} w_5 - \nu\sigma w_2 - \mathrm{G}_{U+\omega}\, w_2 - (\hat{\mathrm{B}}_{U+\omega} - \sigma)w_3 + \sigma\tilde{u}\|^2_{L^2(\Omega,\mathbb{R}^2)},\end{aligned}$$

i.e., we have to minimize the functional $J\colon H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2}) \times L^2(\Omega, \mathbb{R}^2) \times H(\Omega) \to \mathbb{R}$ defined by the right-hand side of the previous identity (with the variables ordered as

$(w_5, w_2, w_3))$. Again, the additional assumption on w_5 is satisfied by construction. Moreover, similar as in the minimization procedure for the Rayleigh-Ritz method we choose a suitable finite element subspace of $H(\mathrm{div}, \Omega, \mathbb{R}^{2\times 2}) \times L^2(\Omega, \mathbb{R}^2) \times H(\Omega)$ to run our numerical algorithm. In our applications we actually choose quadratic Lagrangian finite elements for each component which yields a solution in the desired spaces by construction.

Remark 6.14. *We note that minimizing b over all variables (w_5, w_2, w_3) results in a relatively high computational effort for the computation of w (corresponding to $\tilde{u}$). To slightly reduce the computational effort we might use the strategy presented for example in [14, middle of p. 77]. Therefore, let $\tilde{\lambda}$ denote the approximate eigenvalue corresponding to the approximate eigenfunction $\tilde{u}$ of the shifted eigenvalue problem* (6.30)*. Then, we can choose $w_2 = \frac{1}{\tilde{\lambda}}\tilde{u} \in L^2(\Omega, \mathbb{R}^2)$ and minimize b only over the remaining variables (w_5, w_3) which in our applications results in "slightly worse" lower bounds in the Goerisch method.*

To the end of this paragraph we shortly present the strategy for the second eigenvalue problem (6.31). Analogously to (6.36) (cf. (6.35)) we obtain

$$\begin{aligned} \hat{M}(u,\varphi) = (1+\nu)\langle u,\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \int_\Omega u^T\,\mathrm{G}_{U+\omega}\,\varphi\,\mathrm{d}(x,y) \\ + \langle \Phi^{-1}(\hat{\mathrm{B}}_{U+\omega}-\sigma)u, \Phi^{-1}(\hat{\mathrm{B}}_{U+\omega}-\sigma)\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u,\varphi \in H(\Omega), \end{aligned} \tag{6.47}$$

where we again use the abbreviation $\mathrm{G}_{U+\omega}$ defined in (6.33).

Therefore, we can use the Goerisch setting presented above (cf. (6.37) and (6.40)) with the third component of Tu replaced by $\Phi^{-1}(\hat{\mathrm{B}}_{U+\omega}-\sigma)u$. The same arguments as in (6.42) yield $\langle \Phi^{-1}(\hat{\mathrm{B}}_{U+\omega}-\sigma)w_3, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \langle(\mathrm{B}_{U+\omega}-\sigma)\varphi, w_3\rangle_{L^2(\Omega,\mathbb{R}^2)}$ for all $\varphi \in H(\Omega)$ which directly implies that condition (6.16) of Theorem 6.8 now reads as

$$\begin{aligned} 0 = \langle(1+\nu)w_1 - \nabla\tilde{u}, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} \\ + \langle \nu\sigma w_2 + \mathrm{G}_{U+\omega}\,w_2 + (\mathrm{B}_{U+\omega}-\sigma)w_3 + \sigma w_4 - \sigma\tilde{u}, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega), \end{aligned}$$

i.e., for our second problem we obtain the required condition by replacing $(\hat{\mathrm{B}}_{U+\omega}-\sigma)w_3$ by $(\mathrm{B}_{U+\omega}-\sigma)w_3$ in (6.43). The same replacement in (6.46) yields the formula for w_4 whereas the formula for w_1 remains the same as in the previous case. Finally, we are left with a minimization procedure for $b(w,w)$ over w_5, w_2 and w_3 on a suitable finite dimensional subspace of $H^1(\Omega, \mathbb{R}^{2\times 2}) \times L^2(\Omega, \mathbb{R}^2) \times H(\Omega)$ again to obtain "good" lower bounds for our eigenvalue problem (6.31).

Having a closer look at the remarks presented after Theorem 6.8 we see that we finally have to compute a "rough" lower bound ρ for the $(n+1)$st eigenvalue if such an eigenvalue exists below the essential spectrum. Thus, in the following paragraph we describe the appliance of the homotopy method presented at the beginning of this Subsection to obtain the desired bound ρ for both eigenvalue problems (6.30) and (6.31) respectively.

6.2.1.2 A Collection of Homotopy Methods

In the further course we present a procedure to compute the desired "rough" lower bound ρ for the $(n+1)$st eigenvalue (needed to apply Theorem 6.8) for the eigenvalue problems (6.30) and (6.31) respectively by a homotopy method. Actually, we perform several different homotopy methods successively starting with a coefficient homotopy followed by a domain deformation homotopy and finally, we perform a homotopy to fade out the solenoidal condition in the space (Note that this is the order of constructing the different homotopy methods and not the order of application).

Although the abstract description of the homotopy method presented above requires the base problem in advance, in this Section we start our explanations with the homotopy of the bilinear forms (cf. (6.24)) since the appropriate base problem is somehow a direct consequence of choosing appropriate bilinear forms for the homotopy methods.

At this stage, we briefly recall the structure of the approximate solution

$$\tilde{\omega} = \begin{cases} \tilde{\omega}_0, & \text{in } \Omega_0, \\ 0, & \text{in } \Omega \setminus \Omega_0, \end{cases}$$

introduced in Section 3.2 (cf. (3.7)). Hence, there exists some radius $R > 0$ such that $\Omega_0 \subseteq S_R \cap \Omega =: \Omega_R$ with $S_R := (-R, R) \times (0, 1)$ (see also Remark 4.2). Moreover, from the construction of the auxiliary function V (cf. Section 4.1) and the definition of ω (see (3.8)) we conclude that $\operatorname{supp}(\omega)$ is contained in the bounded part $\overline{\Omega_R}$. Since during our homotopy we are going to enlarge our domain to the entire strip S we also extend our approximation ω by zero on $S \setminus \Omega$. We note that in the further course also the extended approximation will be denoted by ω where it becomes clear from the context whether the approximation on Ω or on the entire strip S is meant.

In the following, we present two possibilities to perform a homotopy for our eigenvalue problems (6.8) and (6.9). Both versions of our homotopy result in a base problem of the same structure which will be discussed in the next Section.

A Straightforward Coefficient Homotopy (First Approach)

For our first approach we use the same strategy as in the computation of the upper and lower eigenvalue bounds (cf. Section 6.2.1.1), i.e., we use the representation

$$\begin{aligned} M(u, \varphi) = (1+\nu)\langle u, \varphi \rangle_{H_0^1(\Omega, \mathbb{R}^2)} &+ \int_\Omega u^T \, \mathrm{G}_{U+\omega} \, \varphi \, \mathrm{d}(x, y) \\ &+ \langle \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)u, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)\varphi \rangle_{H_0^1(\Omega, \mathbb{R}^2)} \quad \text{for all } u, \varphi \in H(\Omega), \end{aligned}$$

of the bilinear form M again (cf. (6.36)).

In the further course we are going to estimate the term $\int_\Omega u^T \, \mathrm{G}_{U+\omega} \, u \, \mathrm{d}(x, y)$ from below. Therefore, introducing the abbreviation $\tilde{\mathrm{G}}_{U+\omega} := \mathrm{G}_{U+\omega} + 2\sigma \, \mathrm{id}$ and using the fact that ω vanishes outside of Ω_R (cf. (6.33)) we obtain

$$\tilde{\mathrm{G}}_{U+\omega} = Re[\nabla U + (\nabla U)^T] = \begin{pmatrix} 0 & Re(1-2y) \\ Re(1-2y) & 0 \end{pmatrix} \quad \text{in } \Omega \setminus \Omega_R.$$

Hence, for fixed $(x, y) \in \Omega \setminus \Omega_R$ we get the representation $\lambda_{\min}(\tilde{\mathrm{G}}_{U+\omega}(x, y)) = -Re\,|1 - 2y|$ for the minimal eigenvalue of $\tilde{\mathrm{G}}_{U+\omega}(x, y)$ which directly implies

$$\inf_{(x,y) \in \Omega \setminus \Omega_R} \lambda_{\min}(\tilde{\mathrm{G}}_{U+\omega}(x, y)) = -Re.$$

Thus, together with the definition of $\tilde{\mathrm{G}}_{U+\omega}$ and $\mathrm{G}_{U+\omega}$ respectively (cf. (6.33)) we conclude

$$\int_{\Omega\setminus\Omega_R} u^T \, \mathrm{G}_{U+\omega} \, u \, \mathrm{d}(x,y) \geq -(2\sigma + Re) \int_{\Omega\setminus\Omega_R} |u|^2 \, \mathrm{d}(x,y). \tag{6.48}$$

For the remaining region Ω_R we calculate

$$\begin{aligned}\tilde{\mathrm{G}}_{U+\omega} &= Re[\nabla(U+\omega) + (\nabla(U+\omega))^T] \\ &= \begin{pmatrix} 2Re\frac{\partial\omega_1}{\partial x} & Re\left(1 - 2y + \frac{\partial\omega_1}{\partial y} + \frac{\partial\omega_2}{\partial x}\right) \\ Re\left(1 - 2y + \frac{\partial\omega_1}{\partial y} + \frac{\partial\omega_2}{\partial x}\right) & -2Re\frac{\partial\omega_1}{\partial x} \end{pmatrix} \quad \text{in } \Omega_R,\end{aligned}$$

where we used the fact that ω is a solenoidal function, i.e., $0 = \mathrm{div}(\omega) = \frac{\partial\omega_1}{\partial x} + \frac{\partial\omega_2}{\partial y}$ holds true.

Similar as above we compute the minimal eigenvalue

$$\lambda_{\min}(\tilde{\mathrm{G}}_{U+\omega}(x,y)) = -Re\sqrt{4\left(\frac{\partial\omega_1}{\partial x}\right)^2 + \left(1 - 2y + \frac{\partial\omega_1}{\partial y} + \frac{\partial\omega_2}{\partial x}\right)^2} \quad \text{for all } (x,y) \in \Omega_R$$

and define

$$-\tau := \frac{1}{Re} \inf_{(x,y)\in\Omega_R} \lambda_{\min}(\tilde{\mathrm{G}}_{U+\omega}(x,y)) \leq 0.$$

The same arguments as above imply

$$\int_{\Omega_R} u^T \, \mathrm{G}_{U+\omega} \, u \, \mathrm{d}(x,y) \geq -(2\sigma + Re\,\tau) \int_{\Omega_R} |u|^2 \, \mathrm{d}(x,y). \tag{6.49}$$

Remark 6.15. *To obtain the desired lower bound $-\tau$ we compute enclosures of the ranges of $\frac{\partial\omega_1}{\partial x}$, $\frac{\partial\omega_1}{\partial y}$ and $\frac{\partial\omega_2}{\partial x}$ respectively by the techniques presented in Section 5.1. Therefore, we again use the Bernstein expansion for the (polynomial) terms and apply the enclosure result in* (5.7)*.*

Combining both estimates (6.48) and (6.49) we obtain the following estimate on the entire domain Ω:

$$\int_{\Omega} u^T \, \mathrm{G}_{U+\omega} \, u \, \mathrm{d}(x,y) \geq -(2\sigma + Re) \int_{\Omega} u^2 \, \mathrm{d}(x,y) - Re(\tau - 1) \int_{\Omega_R} |u|^2 \, \mathrm{d}(x,y).$$

Hence, using Sobolev's embedding constant C_2 (see (2.12)) we are in a position to estimate our bilinear form M from below by

$$M(u,u) \geq (1 + \nu - {C_2}^2(2\sigma + Re))\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - Re(\tau - 1) \int_{\Omega_R} |u|^2 \, \mathrm{d}(x,y) \tag{6.50}$$

for all $u \in H(\Omega)$. Introducing the constants $\gamma_1 := 1 - {C_2}^2(2\sigma + Re)$ and $\gamma_2 := Re(\tau - 1)$ inspired by the right-hand side we define the bilinear form $M_0\colon H(\Omega) \times H(\Omega) \to \mathbb{R}$ for the base problem of our first (coefficient) homotopy by

$$M_0(u,\varphi) := (\gamma_1 + \nu)\langle u,\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - \gamma_2 \int_{\Omega_R} u \cdot \varphi \, \mathrm{d}(x,y) \quad \text{for all } u,\varphi \in H(\Omega) \tag{6.51}$$

which now (by (6.50)) satisfies the crucial inequality (6.22) needed for the homotopy method. Note that in this case the spaces $H_0 = H(\Omega)$ and $H = H(\Omega)$ (and thus the inner products) coincide.

In the context of our base problem (cf. Section 6.2.1.3) we require the constant γ_1 to be positive. Therefore, in the further course we assume that $\gamma_1 > 0$. Having a closer look at the definition of γ_1 again, we realize that $\gamma_1 > 0$ obviously cannot hold true for arbitrarily large Reynolds numbers, i.e., our homotopy approach only applies to the case of "small" Reynolds numbers. To overcome this difficulty we introduce a more complex version in the subsequent Section.

Moreover, if $\gamma_2 \leq 0$ we can estimate the second term from below by 0 as well which will result in the bilinear form $M_0(u,\varphi) := (\gamma_1+\nu)\langle u,\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)}$. In the affirmative case, $\gamma_1+\nu$ is a lower bound for the spectral points of the shifted eigenvalue problem (6.30). Since we assumed $\gamma_1 > 0$ it is the desired lower bound for the spectral points of the original eigenvalue problem (6.8). However, since in all our applications the case $\gamma_2 \leq 0$ is not appearing we do not go into further details and assume the "difficult" case $\gamma_2 > 0$ in the following.

In the case $\gamma_2 > 0$ we actually have to perform a (coefficient) homotopy to obtain the desired lower bound for the eigenvalues of our original problem (6.8). Therefore, in view of the abstract setting for the homotopy (see (6.24)) we define the family of spaces $(H_t, \langle\,\cdot\,,\,\cdot\,\rangle_t) := (H(\Omega), \langle\,\cdot\,,\,\cdot\,\rangle_{H_0^1(\Omega,\mathbb{R}^2)})$ for all $t \in [0,1]$ and the family $(M_t)_{t\in[0,1]}$ of bilinear forms by

$$M_t(u,\varphi) := t\,M(u,\varphi) + (1-t)\,M_0(u,\varphi) \quad \text{for all } u,\varphi \in H_t = H(\Omega). \tag{6.52}$$

Hence, applying the inequality above again for fixed $u \in H(\Omega)$ we calculate

$$M_t(u,u) - M_s(u,u) = (t-s)(M(u,u) - M_0(u,u)) \geq 0 \quad \text{for all } 0 \leq s \leq t \leq 1$$

which implies the desired inequalities for the quotients $\frac{M_t(u,u)}{\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}}$ (cf. (6.24)).

Thus, to apply the homotopy method with the setting presented above we need information about the eigenvalues of the base problem

$$\begin{aligned} &u \in H(\Omega),\\ &(\gamma_1+\nu)\langle u,\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - \gamma_2\int_{\Omega_R} u\cdot\varphi\,\mathrm{d}(x,y) = \lambda_\nu\langle u,\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega). \end{aligned} \tag{6.53}$$

However, due to the "complicated" domain Ω we cannot compute (or enclose) the eigenvalues of this (first) base problem directly. Thus, we perform a (second) domain deformation homotopy (see corresponding Section afterwards) to obtain the desired information about the eigenvalues of problem (6.53).

Finally, we shortly present the Goerisch setting (cf. Theorem 6.8 and Corollary 6.9 respectively) which is needed to obtain the lower bounds in each homotopy step (cf. description of the homotopy method presented in the beginning of Section 6.2.1). Therefore, we first fix the vector space $X := L^2(\Omega,\mathbb{R}^{2\times2}) \times L^2(\Omega,\mathbb{R}^2) \times H(\Omega) \times L^2(\Omega,\mathbb{R}^2)$ and

$$T\colon H(\Omega) \to X,\ Tu := (\nabla u, u, \Phi^{-1}(\mathrm{B}_{U+\omega} - \sigma)u, u)^T$$

which coincide with the definitions for the Goerisch setting for the shifted eigenvalue problem (6.30) presented in Section 6.2.1.1. Moreover, for any $t \in [0,1]$ we consider the bilinear form $b_t \colon X \times X \to \mathbb{R}$ given by

$$\begin{aligned} b_t(w,\hat{w}) &:= (t+(1-t)\gamma_1+\nu)\langle w_1,\hat{w}_1\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \sigma(t+(1-t)\gamma_1+\nu-1)\langle w_2,\hat{w}_2\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &\quad + t\int_\Omega w_2^T \, \mathrm{G}_{U+\omega}\, \hat{w}_2 \,\mathrm{d}(x,y) - (1-t)\gamma_2 \int_{\Omega_R} w_2\cdot \hat{w}_2 \,\mathrm{d}(x,y) \\ &\quad + t\langle w_3,\hat{w}_3\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \sigma\langle w_4,\hat{w}_4\rangle_{L^2(\Omega,\mathbb{R}^2)}. \end{aligned}$$

Then, for fixed $t \in [0,1]$ from the definition of b_t we conclude that b_t is positive semi-definite if

$$\begin{aligned} 0 \le \sigma(t+(1-t)\gamma_1+\nu-1)\langle v,v\rangle_{L^2(\Omega,\mathbb{R}^2)} + t\int_\Omega v^T \, \mathrm{G}_{U+\omega}\, v \,\mathrm{d}(x,y) \\ - (1-t)\gamma_2 \int_{\Omega_R} v\cdot v \,\mathrm{d}(x,y) \end{aligned}$$

holds true for all $v \in L^2(\Omega,\mathbb{R}^2)$ (recall that $\gamma_1, \gamma_2 > 0$ in our considerations). Using inequality (6.39) from the previous Section again we obtain

$$\begin{aligned} &\sigma(t+(1-t)\gamma_1+\nu-1)\langle v,v\rangle_{L^2(\Omega,\mathbb{R}^2)} + t\int_\Omega v^T \, \mathrm{G}_{U+\omega}\, v \,\mathrm{d}(x,y) \\ &\quad - (1-t)\gamma_2 \int_{\Omega_R} v\cdot v \,\mathrm{d}(x,y) \\ &\quad \ge \left(\sigma(\nu-1) - 2(Re\|\nabla(U+\omega)\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})}+\sigma) - \gamma_2\right)\langle v,v\rangle_{L^2(\Omega,\mathbb{R}^2)} \end{aligned}$$

for all $v \in L^2(\Omega,\mathbb{R}^2)$. We note that this estimate could be sightly improved, but in all our examples it turned out that our “rough” estimate is sufficient. Hence, fixing the shift parameter ν such that

$$\nu \ge \frac{2Re\|\nabla(U+\omega)\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})} + \gamma_2}{\sigma} + 3 > 0$$

directly implies that b_t is positive semi-definite.

Furthermore, using the definition of M_t (see (6.52)) as well as of its components M (see (6.36)) and M_0 (see (6.51)) we calculate

$$\begin{aligned} b_t(Tu,T\varphi) &= (t+(1-t)\gamma_1+\nu)\langle \nabla u,\nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} + \sigma(t+(1-t)\gamma_1+\nu-1)\langle u,\varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &\quad + t\int_\Omega u^T \, \mathrm{G}_{U+\omega}\, \varphi \,\mathrm{d}(x,y) - (1-t)\gamma_2 \int_{\Omega_R} u\cdot \varphi \,\mathrm{d}(x,y) \\ &\quad + t\langle \Phi^{-1}(\mathrm{B}_{U+\omega}-\sigma)u, \Phi^{-1}(\mathrm{B}_{U+\omega}-\sigma)\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \sigma\langle u,\varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} \\ &= t\,M(u,\varphi) + (1-t)\,M_0(u,\varphi) = M_t(u,\varphi) \quad \text{for all } u,\varphi \in H(\Omega) \end{aligned}$$

proving that condition (6.15) from Theorem 6.8 is satisfied for our bilinear form b_t, i.e., we have defined a suitable “XbT-setting” for our homotopy method. We note that b_1 indeed coincides with the bilinear form b (cf. (6.37)) introduced in the Goerisch setting for computing lower eigenvalue bounds of the shifted eigenvalue problem (6.30).

Next, for some $v \in H(\Omega)$ we shortly present a possible choice for a function $w \in X$ (corresponding to v) satisfying (6.16) in Theorem 6.8. Again, we omit the index j in the further course and thus, we have to consider assumption (6.16) with v_j replaced by v and w_j replaced by w, i.e., our condition for the desired function $w = (w_1, w_2, w_3, w_4)^T \in X$ now reads as

$$\begin{aligned}(t + (1-t)\gamma_1 + \nu)\langle w_1, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})} &+ \sigma(t + (1-t)\gamma_1 + \nu - 1)\langle w_2, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)}\\ &+ t\int_\Omega w_2^T \operatorname{G}_{U+\omega} \varphi \,\mathrm{d}(x,y) - (1-t)\gamma_2 \int_{\Omega_R} w_2 \cdot \varphi \,\mathrm{d}(x,y)\\ &+ t\langle w_3, \Phi^{-1}(\operatorname{B}_{U+\omega} - \sigma)\varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \sigma\langle w_4, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} = \langle v, \varphi\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\end{aligned}$$

for all $\varphi \in H(\Omega)$.

To solve for one of the components of w we follow the lines already described in the previous Section about the computation of lower eigenvalue bounds for the shifted eigenvalue problem (6.30). Using the identity (6.42) together with the symmetry of $\operatorname{G}_{U+\omega}$ again we get the following condition for w:

$$\begin{aligned}0 = \langle (t + (1-t)\gamma_1 + \nu)w_1 - \nabla v, \nabla\varphi\rangle_{L^2(\Omega,\mathbb{R}^{2\times 2})}&\\ + \langle (\sigma(t + (1-t)\gamma_1 + \nu - 1) - (1-t)\gamma_2\chi_{\Omega_R})w_2 + t\operatorname{G}_{U+\omega} w_2, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)}&\\ + \langle t(\hat{\operatorname{B}}_{U+\omega} - \sigma)w_3 + \sigma w_4 - \sigma v, \varphi\rangle_{L^2(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega).&\end{aligned}$$

Similar as in the previous considerations we introduce the new variable $w_5 \in H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2})$ and set

$$w_1 := \frac{1}{t + (1-t)\gamma_1 + \nu}\left(w_5 + \nabla v\right).$$

Hence, inserting this identity into the condition for w from above integration by parts (cf. (6.45)) yields

$$\begin{aligned}0 = \Big\langle -\operatorname{div} w_5 + (\sigma(t + (1-t)\gamma_1 + \nu - 1) - (1-t)\gamma_2\chi_{\Omega_R})w_2&\\ + t\operatorname{G}_{U+\omega} w_2 + t(\hat{\operatorname{B}}_{U+\omega} - \sigma)w_3 + \sigma w_4 - \sigma v, \varphi\Big\rangle_{L^2(\Omega,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(\Omega)&\end{aligned}$$

which can be solved for the fourth component again, i.e., we fix

$$\begin{aligned}w_4 := \frac{1}{\sigma}\big(\operatorname{div} w_5 - (\sigma(t + (1-t)\gamma_1 + \nu - 1) - (1-t)\gamma_2\chi_{\Omega_R})w_2&\\ - t\operatorname{G}_{U+\omega} w_2 - t(\hat{\operatorname{B}}_{U+\omega} - \sigma)w_3 + \sigma v\big).&\end{aligned}$$

Finally, the components $w_2 \in L^2(\Omega, \mathbb{R}^2)$, $w_3 \in H(\Omega)$ and $w_5 \in H(\operatorname{div}, \Omega, \mathbb{R}^{2\times 2})$ can be chosen arbitrary in each homotopy step. However, to obtain “tight” lower bounds during our homotopy we follow the lines in [74, Remark 10.26 (c)] again, i.e., we approximately minimize

$$\begin{aligned}b_t(w,w) = \frac{1}{t + (1-t)\gamma_1 + \nu}\|w_5 + \nabla v\|^2_{L^2(\Omega,\mathbb{R}^{2\times 2})} &+ \sigma(t + (1-t)\gamma_1 + \nu - 1)\|w_2\|^2_{L^2(\Omega,\mathbb{R}^2)}\\ &+ t\int_\Omega w_2^T \operatorname{G}_{U+\omega} w_2 \,\mathrm{d}(x,y) - (1-t)\gamma_2\int_{\Omega_R} |w_2|^2 \,\mathrm{d}(x,y) + t\|w_3\|^2_{H_0^1(\Omega,\mathbb{R}^2)}\\ &+ \frac{1}{\sigma}\big\|\operatorname{div} w_5 - (\sigma(t + (1-t)\gamma_1 + \nu - 1) - (1-t)\gamma_2\chi_{\Omega_R})w_2\\ &\qquad - t\operatorname{G}_{U+\omega} w_2 - t(\hat{\operatorname{B}}_{U+\omega} - \sigma)w_3 + \sigma v\big\|^2_{L^2(\Omega,\mathbb{R}^2)}\end{aligned}$$

over the free variables w_2, w_3 and w_5 in a suitable finite element subspace (cf. previous Section).

To the end of this Subsection, we will have a closer look at the "adjoint" eigenvalue problem (6.31) with its bilinear form $\hat{M}$ given in (6.47). Comparing the definitions of M and $\hat{M}$ carefully we see that the same arguments as above imply

$$\hat{M}(u,u) \geq (1+\nu - {C_2}^2(2\sigma + Re))\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - Re(\tau - 1)\int_{\Omega_R} |u|^2 \,\mathrm{d}(x,y)$$

for all $u \in H(\Omega)$, i.e., after estimating the bilinear form of the "adjoint" problem we are in the same situation already considered for problem (6.30). Hence, the methods described previously (starting subsequent to (6.50)) are applicable in the setting of the "adjoint" problem which results in the same base problem (6.53), i.e., the same constants γ_1 and γ_2. Therefore, the required information about the eigenvalues of the base problem in the "adjoint" case coincides with the information needed in the "original" case. The Goerisch setting needed for the homotopy method in the "adjoint" case can be formulated mutatis mutandis to the setting for the shifted problem (6.30) presented above.

A More Complex Coefficient Homotopy (Second Approach)

Before coming to the domain deformation homotopy mentioned above we first present a second approach for the coefficient homotopy which will result in a similar base problem compared to the first approach (cf. (6.53)). For a comparison of both approaches we refer the reader to Section 8.1. As a first step for this extended coefficient homotopy we prove some technical preliminaries needed to formulate the homotopy setting.

Lemma 6.16. *The following equalities hold true for all $u \in H(\Omega)$:*

(i) $$\begin{aligned}&\langle \Phi^{-1}\,\mathrm{L}_{U+\omega}\,u, \Phi^{-1}\,\mathrm{L}_{U+\omega}\,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad= \langle \Phi^{-1}\,\mathrm{L}_U\,u, \Phi^{-1}\,\mathrm{L}_U\,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}\,\mathrm{B}_\omega\,u, \Phi^{-1}(\mathrm{B}_\omega + 2\,\mathrm{B}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\qquad+ Re\int_{\Omega_R} u^T((\nabla\omega) + (\nabla\omega)^T)u\,\mathrm{d}(x,y)\end{aligned}$$

(ii) $$\begin{aligned}&\langle \Phi^{-1}\,\hat{\mathrm{L}}_{U+\omega}\,u, \Phi^{-1}\,\hat{\mathrm{L}}_{U+\omega}\,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad= \langle \Phi^{-1}\,\hat{\mathrm{L}}_U\,u, \Phi^{-1}\,\hat{\mathrm{L}}_U\,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}\,\hat{\mathrm{B}}_\omega\,u, \Phi^{-1}(\hat{\mathrm{B}}_\omega + 2\,\hat{\mathrm{B}}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\qquad+ Re\int_{\Omega_R} u^T((\nabla\omega) + (\nabla\omega)^T)u\,\mathrm{d}(x,y)\end{aligned}$$

Proof. First, we can treat both cases simultaneously. Therefore, for some $w \in W(\Omega)$ let $(\mathcal{L}_w, \mathcal{B}_w)$ either be $(\mathrm{L}_w, \mathrm{B}_w)$ or $(\hat{\mathrm{L}}_w, \hat{\mathrm{B}}_w)$. Then, for all $u \in H(\Omega)$ we get:

$$\begin{aligned}&\langle \Phi^{-1}\mathcal{L}_{U+\omega}u, \Phi^{-1}\mathcal{L}_{U+\omega}u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - \langle \Phi^{-1}\mathcal{L}_U u, \Phi^{-1}\mathcal{L}_U u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad= \langle \Phi^{-1}((-\Delta+\sigma)u + (\mathcal{B}_{U+\omega}-\sigma)u), \Phi^{-1}((-\Delta+\sigma)u + (\mathcal{B}_{U+\omega}-\sigma)u)\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\qquad- \langle \Phi^{-1}((-\Delta+\sigma)u + (\mathcal{B}_U-\sigma)u), \Phi^{-1}((-\Delta+\sigma)u + (\mathcal{B}_U-\sigma)u)\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad= \langle u + \Phi^{-1}(\mathcal{B}_{U+\omega}-\sigma)u, u + \Phi^{-1}(\mathcal{B}_{U+\omega}-\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\qquad- \langle u + \Phi^{-1}(\mathcal{B}_U-\sigma)u, u + \Phi^{-1}(\mathcal{B}_U-\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\end{aligned}$$

$$\begin{aligned}
&= 2\langle \Phi^{-1}(\mathcal{B}_{U+\omega} - \sigma)u, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \langle \Phi^{-1}(\mathcal{B}_{U+\omega} - \sigma)u, \Phi^{-1}(\mathcal{B}_{U+\omega} - \sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad - 2\langle \Phi^{-1}(\mathcal{B}_{U} - \sigma)u, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - \langle \Phi^{-1}(\mathcal{B}_{U} - \sigma)u, \Phi^{-1}(\mathcal{B}_{U} - \sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&= 2((\mathcal{B}_{U+\omega} - \mathcal{B}_U)u)[u] \\
&\quad + \langle \Phi^{-1}((\mathcal{B}_{U+\omega} - \sigma) - (\mathcal{B}_U - \sigma))u, \Phi^{-1}((\mathcal{B}_{U+\omega} - \sigma) + (\mathcal{B}_U - \sigma))u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&= 2(\mathcal{B}_\omega u)[u] + \langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}(\mathcal{B}_\omega + 2\mathcal{B}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}.
\end{aligned}$$

Next, we treat both cases separately.

(i) In the first case, i.e., $(\mathcal{L}_w, \mathcal{B}_w) = (\mathrm{L}_w, \mathrm{B}_w)$ holds true, we obtain

$$(\mathcal{B}_\omega u)[u] = (\mathrm{B}_\omega\, u)[u] = Re \int_\Omega [(u \cdot \nabla)\omega] \cdot u \,\mathrm{d}(x,y) + Re \int_\Omega [(\omega \cdot \nabla)u] \cdot u \,\mathrm{d}(x,y).$$

Using the compact support of ω, we can replace the integration domain Ω by the computational domain $\Omega_R \supseteq \mathrm{supp}(\omega)$. Moreover, Lemma A.10 (iii) directly yields $\int_\Omega [(\omega \cdot \nabla)u] \cdot u \,\mathrm{d}(x,y) = 0$. Thus, we obtain

$$\begin{aligned}
(\mathrm{B}_\omega\, u)[u] &= Re \int_{\Omega_R} [(u \cdot \nabla)\omega] \cdot u \,\mathrm{d}(x,y) = Re \int_{\Omega_R} u^T (\nabla\omega) u \,\mathrm{d}(x,y) \\
&= \frac{1}{2} Re \int_{\Omega_R} u^T ((\nabla\omega) + (\nabla\omega)^T) u \,\mathrm{d}(x,y).
\end{aligned}$$

(ii) For $(\mathcal{L}_w, \mathcal{B}_w) = (\hat{\mathrm{L}}_w, \hat{\mathrm{B}}_w)$ we get

$$(\mathcal{B}_\omega)[u] = (\hat{\mathrm{B}}_\omega\, u)[u] = Re \int_\Omega u^T (\nabla\omega) u \,\mathrm{d}(x,y) - Re \int_\Omega [(\omega \cdot \nabla)u] \cdot u \,\mathrm{d}(x,y).$$

The same arguments as in part (i) yield

$$(\hat{\mathrm{B}}_\omega\, u)[u] = Re \int_{\Omega_R} u^T (\nabla\omega) u \,\mathrm{d}(x,y) = \frac{1}{2} Re \int_{\Omega_R} u^T ((\nabla\omega) + (\nabla\omega)^T) u \,\mathrm{d}(x,y).$$

Finally, in both cases the assertion follows. □

Additionally, we need the following Lemma for the estimation of the bilinear form M.

Lemma 6.17. *Let $\tau_1, \tau_2 \in (0,\infty)$ such that $\frac{1}{\tau_1} + \frac{\sigma}{\tau_2} \geq 1$. Then, the following inequalities hold true for all $u \in H(\Omega)$:*

(i) $\langle \Phi^{-1}\, \mathrm{B}_\omega\, u, \Phi^{-1}(\mathrm{B}_\omega + 2\,\mathrm{B}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}$

$$\geq \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right) 4Re^2 \|\omega\|^2_{L^\infty(\Omega_R,\mathbb{R}^2)} \|u\|^2_{L^2(\Omega_R,\mathbb{R}^2)} - \left(\frac{\tau_1}{4} Re^2 + \tau_2 \sigma {C_2}^2\right) \|u\|^2_{L^2(\Omega,\mathbb{R}^2)}$$

(ii) $\langle \Phi^{-1}\, \hat{\mathrm{B}}_\omega\, u, \Phi^{-1}(\hat{\mathrm{B}}_\omega + 2\,\hat{\mathrm{B}}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}$

$$\begin{aligned}
&\geq \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right) Re^2 \left(\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)} + C_2 \|\nabla\omega\|_{L^\infty(\Omega_R,\mathbb{R}^{2\times 2})}\right)^2 \|u\|^2_{L^2(\Omega_R,\mathbb{R}^2)} \\
&\quad - \left(\tau_1 Re^2 \left(C_2 + \frac{1}{4}\right)^2 + \tau_2 \sigma {C_2}^2\right) \|u\|^2_{L^2(\Omega,\mathbb{R}^2)}
\end{aligned}$$

Proof. Again, in the beginning we can treat both cases simultaneously. Therefore, for some $w \in W(\Omega)$ let $\mathcal{B}_w$ either be B_w or $\hat{\mathrm{B}}_w$. Applying Cauchy-Schwarz' inequality we obtain

$$2\left|\langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}\mathcal{B}_U u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\right| \le 2\|\Phi^{-1}\mathcal{B}_\omega u\|_{H_0^1(\Omega,\mathbb{R}^2)}\|\Phi^{-1}\mathcal{B}_U u\|_{H_0^1(\Omega,\mathbb{R}^2)}.$$

Hence, Young's inequality implies

$$2\left|\langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}\mathcal{B}_U u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\right| \le \frac{1}{\tau_1}\|\Phi^{-1}\mathcal{B}_\omega u\|^2_{H_0^1(\Omega,\mathbb{R}^2)} + \tau_1\|\Phi^{-1}\mathcal{B}_U u\|^2_{H_0^1(\Omega,\mathbb{R}^2)}$$

for arbitrary $\tau_1 > 0$. In almost the same manner we estimate

$$2\left|\langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\right| \le \frac{1}{\tau_2}\|\Phi^{-1}\mathcal{B}_\omega u\|^2_{H_0^1(\Omega,\mathbb{R}^2)} + \tau_2\|\Phi^{-1}u\|^2_{H_0^1(\Omega,\mathbb{R}^2)}$$

for arbitrary $\tau_2 > 0$. Using the isometric property of Φ and the estimate introduced in Section 2.2 (directly following equation (2.13)) we obtain

$$\|\Phi^{-1}u\|_{H_0^1(\Omega,\mathbb{R}^2)} = \|u\|_{H(\Omega)'} \le C_2\|u\|_{L^2(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in L^2(\Omega,\mathbb{R}^2)$$

which then implies

$$2\left|\langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\right| \le \frac{1}{\tau_2}\|\Phi^{-1}\mathcal{B}_\omega u\|^2_{H_0^1(\Omega,\mathbb{R}^2)} + \tau_2{C_2}^2\|u\|^2_{L^2(\Omega,\mathbb{R}^2)}.$$

Putting everything together, we obtain

$$\begin{aligned}&\langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}(\mathcal{B}_\omega + 2\mathcal{B}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad = \|\Phi^{-1}\mathcal{B}_\omega u\|^2_{H_0^1(\Omega,\mathbb{R}^2)} + 2\langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}\mathcal{B}_U u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - 2\sigma\langle \Phi^{-1}\mathcal{B}_\omega u, \Phi^{-1}u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad \ge \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right)\|\Phi^{-1}\mathcal{B}_\omega u\|^2_{H_0^1(\Omega,\mathbb{R}^2)} - \tau_1\|\Phi^{-1}\mathcal{B}_U u\|^2_{H_0^1(\Omega,\mathbb{R}^2)} - \sigma\tau_2{C_2}^2\|u\|^2_{L^2(\Omega,\mathbb{R}^2)}.\end{aligned}$$

In the following we distinguish the two cases again and treat them separately.

(i) In the first case we have $\mathcal{B}_w = \mathrm{B}_w$. Hence, Lemma A.11 (i) yields

$$\begin{aligned}&\langle \Phi^{-1}\,\mathrm{B}_\omega\, u, \Phi^{-1}(\mathrm{B}_\omega + 2\,\mathrm{B}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad \ge \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right)4Re^2\|\omega\|^2_{L^\infty(\Omega_R,\mathbb{R}^2)}\|u\|^2_{L^2(\Omega_R,\mathbb{R}^2)}\\ &\qquad - \tau_1 4Re^2\|U\|^2_{L^\infty(\Omega,\mathbb{R}^2)}\|u\|^2_{L^2(\Omega,\mathbb{R}^2)} - \sigma\tau_2{C_2}^2\|u\|^2_{L^2(\Omega,\mathbb{R}^2)}\end{aligned}$$

Directly from the definition of U (cf. (1.10)) we obtain

$$\|U\|_{L^\infty(\Omega,\mathbb{R}^2)} \le \|U\|_{L^\infty(S,\mathbb{R}^2)} = \frac{1}{4}$$

and thus, the first assertion follows.

(ii) In the second case let $\mathcal{B}_w = \hat{\mathrm{B}}_w$. Thus, Lemma A.11 (ii) implies

$$\begin{aligned}&\langle \Phi^{-1}\,\hat{\mathrm{B}}_\omega\, u, \Phi^{-1}(\hat{\mathrm{B}}_\omega + 2\,\hat{\mathrm{B}}_U - 2\sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\\ &\quad \ge \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right)Re^2\left(\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)} + C_2\|\nabla\omega\|_{L^\infty(\Omega_R,\mathbb{R}^{2\times 2})}\right)^2\|u\|^2_{L^2(\Omega_R,\mathbb{R}^2)}\\ &\qquad - \tau_1\left(\|U\|_{L^\infty(\Omega,\mathbb{R}^2)} + C_2\|\nabla U\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})}\right)^2\|u\|^2_{L^2(\Omega_R,\mathbb{R}^2)} - \sigma\tau_2{C_2}^2\|u\|^2_{L^2(\Omega,\mathbb{R}^2)}\end{aligned}$$

Again, by (1.10), we get $\|U\|_{L^\infty(\Omega,\mathbb{R}^2)} \leq \frac{1}{4}$ and

$$\|\nabla U\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})} = \left\| \frac{\partial U_1}{\partial y} \right\|_{L^\infty(\Omega)} = \|1-2y\|_{L^\infty(\Omega)} \leq \|1-2y\|_{L^\infty(S)} = 1,$$

yielding the second assertion.

□

Now, we consider the shifted eigenvalue problem (6.30) and present an appropriate homotopy setting. Similar as in the previous approach (cf. computation of τ and Remark 6.15) we compute the lower bound

$$\gamma_0 := \inf_{(x,y)\in\Omega_R} \lambda_{\min}((\nabla\omega)(x,y) + ((\nabla\omega)(x,y))^T).$$

Having computed such a lower bound, Lemma 6.16 (i) and Lemma 6.17 (i) imply

$$\begin{aligned}
&\langle \Phi^{-1} \mathrm{L}_{U+\omega}\, u, \Phi^{-1} \mathrm{L}_{U+\omega}\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\geq \langle \Phi^{-1} \mathrm{L}_U\, u, \Phi^{-1} \mathrm{L}_U\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - \left(\frac{\tau_1}{4} Re^2 + \tau_2\sigma {C_2}^2\right) \langle u,u\rangle_{L^2(\Omega,\mathbb{R}^2)} \\
&\quad + \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right) 4Re^2 \|\omega\|^2_{L^\infty(\Omega_R,\mathbb{R}^2)} \langle u,u\rangle_{L^2(\Omega_R,\mathbb{R}^2)} \\
&\quad + Re \int_{\Omega_R} u^T((\nabla\omega) + (\nabla\omega)^T) u \,\mathrm{d}(x,y) \\
&\geq \langle \Phi^{-1} \mathrm{L}_U\, u, \Phi^{-1} \mathrm{L}_U\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} - {C_2}^2 \left(\frac{\tau_1}{4} Re^2 + \tau_2\sigma {C_2}^2\right) \langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad + \left[Re\gamma_0 + \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right) 4Re^2 \|\omega\|^2_{L^\infty(\Omega_R,\mathbb{R}^2)} \right] \langle u,u\rangle_{L^2(\Omega_R,\mathbb{R}^2)}
\end{aligned}$$

for all $\tau_1, \tau_2 \in (0,\infty)$ such that $\frac{1}{\tau_1} + \frac{\sigma}{\tau_2} \geq 1$.

In the further course we suppose that a constant $\kappa > 0$ with

$$\langle \Phi^{-1} \mathrm{L}_U\, u, \Phi^{-1} \mathrm{L}_U\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq \kappa \langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega)$$

is in hand explicitly. For more details about the computation of the desired lower bound κ we refer the reader to Section 6.2.1.4. Then, we calculate

$$\begin{aligned}
&\langle \Phi^{-1} \mathrm{L}_{U+\omega}\, u, \Phi^{-1} \mathrm{L}_{U+\omega}\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \nu \langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\geq \left[\nu + \kappa - {C_2}^2 \left(\frac{\tau_1}{4} Re^2 + \tau_2\sigma {C_2}^2\right)\right] \langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad + \left[Re\gamma_0 + \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right) 4Re^2 \|\omega\|^2_{L^\infty(\Omega_R,\mathbb{R}^2)} \right] \langle u,u\rangle_{L^2(\Omega_R,\mathbb{R}^2)}.
\end{aligned}$$

Hence, we can fix the constants $\tau_1, \tau_2 \in (0,\infty)$ with $\frac{1}{\tau_1} + \frac{\sigma}{\tau_2} \geq 1$ such that

$$\gamma_1 := \kappa - {C_2}^2 \left(\frac{\tau_1}{4} Re^2 + \tau_2\sigma {C_2}^2\right) > 0 \tag{6.54}$$

is not much smaller than κ, i.e., $\gamma_1 \approx \kappa$. Again, if the constant

$$-\gamma_2 := Re\gamma_0 + \left(1 - \frac{1}{\tau_1} - \frac{\sigma}{\tau_2}\right) 4Re^2 \|\omega\|^2_{L^\infty(\Omega_R,\mathbb{R}^2)} \tag{6.55}$$

is non-negative, $\gamma_1 + \nu$ is the desired bound for the spectral points of the shifted eigenvalue problem (6.30), i.e., for the original eigenvalue problem (6.8) we obtain the lower bound γ_1. Otherwise, (for $\gamma_2 > 0$) the corresponding base problem is again of the form (6.53).

Remark 6.18. *Since we have to subtract the shift parameter at the end of our calculations and γ_1 characterizes the essential spectrum of the base problem (cf. Section 6.2.1.3) it makes sense to define γ_1 (and thus γ_2) independent of the shift parameter ν.*

In the following, we present a strategy how to choose the parameters τ_1 and τ_2 respectively needed for the definition of the constants γ_1 and γ_2. Therefore, we first fix some constant $0 < \gamma_1 < \kappa$ "sufficiently close" to κ. Since $\sigma > 0$ we can rearrange the terms in equation (6.54) and obtain $\tau_2 = \frac{\kappa - \gamma_1 - \frac{1}{4}{C_2}^2 Re^2 \tau_1}{\sigma {C_2}^4}$. Assumption $\tau_2 > 0$ from Lemma 6.17 now implies a condition on τ_1 which reads as

$$\tau_1 < \frac{\kappa - \gamma_1}{\frac{1}{4}{C_2}^2 Re^2}. \tag{6.56}$$

Finally, using the defining equation for γ_2 (see (6.55)) we are left with minimizing

$$\gamma_2 = \gamma_2(\tau_1) = -Re\gamma_0 - \left(1 - \frac{1}{\tau_1} - \frac{\sigma^2 {C_2}^4}{\kappa - \gamma_1 - \frac{1}{4}{C_2}^2 Re^2 \tau_1}\right) 4Re^2 \|\omega\|^2_{L^\infty(\Omega_R, \mathbb{R}^2)}$$

together with the constraints $\tau_1 > 0$, (6.56) as well as

$$\frac{1}{\tau_1} + \frac{\sigma^2 {C_2}^4}{\kappa - \gamma_1 - \frac{1}{4}{C_2}^2 Re^2 \tau_1} \geq 1 \tag{6.57}$$

(cf. Lemma 6.17). Performing the minimization procedure described in Appendix A.6 in Lemma A.12 (with $a = \sigma^2 {C_2}^4$, $b = \kappa - \gamma_1$ and $c = \frac{1}{4}{C_2}^2 Re^2$; cf. (6.56)) we obtain the desired constant $\tau_1 = \frac{\kappa - \gamma_1}{\frac{1}{2}{C_2}^2 Re(\frac{1}{2}Re + \sigma C_2)}$ and the side condition (6.57) requires the lower bound

$$\gamma_1 \geq \kappa - {C_2}^2 \left(\frac{1}{2}Re + \sigma C_2\right)^2 \tag{6.58}$$

which finally yields an "a posteriori" constraint on γ_1, i.e., we have to fix $\gamma_1 < \kappa$ "sufficiently close" to κ such that this inequality is satisfied. Inserting the results in the definition of γ_2 we get

$$\gamma_2 = -Re\gamma_0 - \left(1 - \frac{{C_2}^2(\frac{1}{2}Re + \sigma C_2)^2}{\kappa - \gamma_1}\right) 4Re^2 \|\omega\|^2_{L^\infty(\Omega_R, \mathbb{R}^2)}.$$

Remark 6.19. *In some of our applications it turned out to be more efficient to choose a slightly smaller constant γ_1 (but still satisfying condition* (6.58)*) which then results in a larger constant γ_2 (cf. computation of γ_1 and γ_2 from above). Hence, by the structure of the base problem* (6.53) *we see that this strategy reduces the number of eigenvalues which have to be considered in the homotopy method. Thus, in our applications this effect massively reduces the computational effort for the homotopy.*

We close this Subsection with some final remarks on the "adjoint" problem (6.31). Therefore, we first assume that we have computed a constant $\hat{\kappa} > 0$ explicitly such that

$$\langle \Phi^{-1}\,\hat{\mathrm{L}}_U\, u, \Phi^{-1}\,\hat{\mathrm{L}}_U\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq \hat{\kappa}\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega).$$

Again, for more details about the computation procedure we refer the reader to Section 6.2.1.4.

Next, we can apply Lemma 6.16 (ii) and Lemma 6.17 (ii) to obtain

$$\begin{aligned}
&\langle \Phi^{-1}\,\hat{\mathrm{L}}_{U+\omega}\, u, \Phi^{-1}\,\hat{\mathrm{L}}_{U+\omega}\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \hat{\nu}\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad\geq \left[\hat{\nu} + \hat{\kappa} - {C_2}^2\left(\tau_1 Re^2\left(C_2+\frac{1}{4}\right)^2 + \tau_2\sigma {C_2}^2\right)\right]\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\qquad + \left[Re\gamma_0 + \left(1-\frac{1}{\hat{\tau}_1}-\frac{\sigma}{\hat{\tau}_2}\right)Re^2\left(\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)} + C_2\|\nabla\omega\|_{L^\infty(\Omega_R,\mathbb{R}^{2\times 2})}\right)^2\right]\langle u,u\rangle_{L^2(\Omega_R,\mathbb{R}^2)}
\end{aligned}$$

for all $\hat{\tau}_1, \hat{\tau}_2 \in (0,\infty)$ such that $\frac{1}{\hat{\tau}_1} + \frac{\sigma}{\hat{\tau}_2} \geq 1$.

Similar as above, we choose $\hat{\tau}_1, \hat{\tau}_2 \in (0,\infty)$ (with $\frac{1}{\hat{\tau}_1} + \frac{\sigma}{\hat{\tau}_2} \geq 1$) such that

$$\hat{\gamma}_1 := \hat{\kappa} - {C_2}^2\left(\tau_1 Re^2\left(C_2+\frac{1}{4}\right)^2 + \tau_2\sigma {C_2}^2\right) > 0 \tag{6.59}$$

and $\hat{\gamma}_1 \approx \hat{\kappa}$. Additionally, we define

$$-\hat{\gamma}_2 := Re\gamma_0 + \left(1-\frac{1}{\hat{\tau}_1}-\frac{\sigma}{\hat{\tau}_2}\right)Re^2\left(\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)} + C_2\|\nabla\omega\|_{L^\infty(\Omega_R,\mathbb{R}^{2\times 2})}\right)^2. \tag{6.60}$$

Again, if $\hat{\gamma}_2 \leq 0$ we obtain the desired lower bound $\hat{\gamma}_1 + \hat{\nu}$ for the spectral points of the shifted eigenvalue problem (6.31) which results in the lower bound $\hat{\gamma}_1$ for the eigenvalue problem (6.9). Otherwise, we consider the base problem (6.53) again but now with γ_1 replaced by $\hat{\gamma}_1$ and γ_2 replaced by $\hat{\gamma}_2$.

To compute the constants $\hat{\tau}_1, \hat{\tau}_2$ needed in the definition of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ respectively we proceed in the same way as mentioned above for the original problem (6.30), i.e., we fix some $0 < \hat{\gamma}_1 < \hat{\kappa}$ "sufficiently close" to $\hat{\kappa}$ as mentioned previously. Finally, a similar minimization procedure (cf. Lemma A.12 with $a = \sigma^2 {C_2}^4$, $b = \hat{\kappa} - \hat{\gamma}_1$, $c = {C_2}^2(C_2+\frac{1}{4})^2 Re^2$) yields

$$\hat{\gamma}_2 = -Re\gamma_0 - \left(1 - \frac{{C_2}^2((C_2+\frac{1}{4})Re + \sigma C_2)^2}{\hat{\kappa}-\hat{\gamma}_1}\right)Re^2\left(\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)} + C_2\|\nabla\omega\|_{L^\infty(\Omega_R,\mathbb{R}^{2\times 2})}\right)^2$$

with the "a posteriori" constraint on $\hat{\gamma}_1$ given by

$$\hat{\gamma}_1 \geq \hat{\kappa} - {C_2}^2\left(\left(C_2+\frac{1}{4}\right)Re + \sigma C_2\right)^2.$$

Domain Deformation Homotopy

To compute the indispensable information about the eigenvalues of the (first) base problem (6.53) (formulated on our domain Ω) we perform a domain deformation homotopy (cf. [74, Section 10.2.5.2] which will result in a (second) base problem on the entire strip S.

Therefore, we choose a family of domains $(\Omega^{(t)})_{t\in[0,1]}$ such that $\Omega^{(0)} = S$ and $\Omega^{(1)} = \Omega$ as well as $\Omega^{(s)} \supseteq \Omega^{(t)}$ for all $0 \le s \le t \le 1$. As mentioned in [74] for the computation of the homotopy steps we actually do not need the entire family of domains but only finitely many of them (starting with S and ending with Ω) are sufficient.

Now, we choose the required families $(H_t, \langle\,\cdot\,,\,\cdot\,\rangle_t)_{t\in[0,1]}$ and $(M_t)_{t\in[0,1]}$ as follows:

$$\begin{aligned}
&H_t := \left\{u \in H(S)\colon u = 0 \text{ on } S \setminus \Omega^{(t)}\right\},\\
&\langle u, \varphi\rangle_t := \langle u, \varphi\rangle_{H_0^1(\Omega^{(t)},\mathbb{R}^2)} \quad \text{for all } u, \varphi \in H_t,\\
&M_t(u, \varphi) := (\gamma_1 + \nu)\langle u, \varphi\rangle_{H_0^1(\Omega^{(t)},\mathbb{R}^2)} - \gamma_2 \int_{S_R\cap\Omega^{(t)}} u \cdot \varphi \,\mathrm{d}(x, y) \quad \text{for all } u, \varphi \in H_t
\end{aligned}$$

for all $0 \le t \le 1$.

Then, due to $\Omega^{(s)} \supseteq \Omega^{(t)}$ we directly see that $H_s \supseteq H_t$ for all $0 \le s \le t \le 1$. Moreover, for all $0 \le s \le t \le 1$ we have $\langle u, \varphi\rangle_s = \langle u, \varphi\rangle_t$ and $M_s(u, u) = M_t(u, u)$ for all $u \in H_t$.

Hence, condition (6.24) holds for the homotopy setting introduced above and (before performing this domain deformation homotopy) we are left with finding appropriate information about the eigenvalues of the following (second) base problem on the strip S:

$$\begin{aligned}
&u \in H(S),\\
&(\gamma_1 + \nu)\langle u, \varphi\rangle_{H_0^1(S,\mathbb{R}^2)} - \gamma_2 \int_{S_R} u \cdot \varphi \,\mathrm{d}(x, y) = \lambda_\nu \langle u, \varphi\rangle_{H_0^1(S,\mathbb{R}^2)} \quad \text{for all } \varphi \in H(S).
\end{aligned} \tag{6.61}$$

Again, we are not in a position to obtain the desired information about the eigenvalues of this base problem. Hence, we perform a third homotopy which fades in the solenoidal property of the space $H(S)$.

Prior to that we shortly introduce the Goerisch setting needed in the domain deformation homotopy. Therefore, for any homotopy parameter $t \in [0, 1]$ we consider the space $X_t := L^2(\Omega^{(t)}, \mathbb{R}^{2\times 2}) \times L^2(\Omega^{(t)}, \mathbb{R}^2) \times L^2(\Omega^{(t)}, \mathbb{R}^2)$. Moreover, we define the bilinear form $b_t\colon X_t \times X_t \to \mathbb{R}$ by

$$\begin{aligned}
b_t(w, \hat{w}) := (\gamma_1 + \nu)\langle w_1, \hat{w}_1\rangle_{L^2(\Omega^{(t)},\mathbb{R}^{2\times 2})} &+ \sigma(\gamma_1 + \nu - 1)\langle w_2, \hat{w}_2\rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)}\\
- \gamma_2 \int_{S_R\cap\Omega^{(t)}} w_2 \cdot \hat{w}_2 \,\mathrm{d}(x, y) &+ \sigma\langle w_3, \hat{w}_3\rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)}
\end{aligned} \tag{6.62}$$

and the Goerisch operator $T_t\colon H_t \to X_t$, $T_t u := (\nabla u, u, u)^T$. Then, obviously b_t defines a symmetric bilinear form on X_t. Additionally, we need to fix the shift parameter $\nu > 0$ such that b_t is positive semi-definite, i.e., such that

$$0 \le \sigma(\gamma_1 + \nu - 1)\langle v, v\rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)} - \gamma_2 \int_{S_R\cap\Omega^{(t)}} |v|^2 \,\mathrm{d}(x, y) \quad \text{for all } v \in L^2(\Omega^{(t)}, \mathbb{R}^2),$$

which can be achieved by choosing

$$\nu \geq \frac{\gamma_2}{\sigma} - \gamma_1 + 1.$$

In addition to that, we calculate

$$\begin{aligned} b_t(Tu, T\varphi) &= (\gamma_1 + \nu)\langle \nabla u, \nabla \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^{2\times 2})} + \sigma(\gamma_1 + \nu - 1)\langle u, \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)} \\ &\quad - \gamma_2 \int_{S_R \cap \Omega^{(t)}} u \cdot \varphi \, \mathrm{d}(x,y) + \sigma \langle u, \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)} \\ &= M_t(u, \varphi) \quad \text{for all } u, \varphi \in H_t, \end{aligned}$$

i.e., condition (6.15) in Theorem 6.8 is satisfied as well.

In view of Corollary 6.9 (or Theorem 6.8 respectively in the case of clustering eigenvalues) in the following we fix $v \in H_t$ and present a strategy to compute a corresponding function $w = (w_1, w_2, w_3)^T \in X_t$ such that $b_t(w, T\varphi) = \langle v, \varphi \rangle_{H_0^1(\Omega^{(t)},\mathbb{R}^2)}$ for all $\varphi \in H_t$ (cf. (6.16)). Using the definition of b_t (see (6.62)) and the Goerisch operator T_t, condition (6.16) reads as

$$\begin{aligned} &(\gamma_1 + \nu)\langle w_1, \nabla \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^{2\times 2})} + \sigma(\gamma_1 + \nu - 1)\langle w_2, \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)} \\ &\quad - \gamma_2 \int_{S_R \cap \Omega^{(t)}} w_2 \cdot \varphi \, \mathrm{d}(x,y) + \sigma \langle w_3, \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)} = \langle v, \varphi \rangle_{H_0^1(\Omega^{(t)},\mathbb{R}^2)} \quad \text{for all } \varphi \in H_t. \end{aligned}$$

which implies the equivalent condition for w:

$$\begin{aligned} 0 &= \langle (\gamma_1 + \nu) w_1 - \nabla v, \nabla \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^{2\times 2})} \\ &\quad + \big\langle (\sigma(\gamma_1 + \nu - 1) - \gamma_2 \chi_{S_R \cap \Omega^{(t)}}) w_2 + \sigma w_3 - \sigma v, \varphi \big\rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)} \quad \text{for all } \varphi \in H_t. \end{aligned} \tag{6.63}$$

Analogously as before, to be able to solve for one of the components of w we introduce a new variable $w_4 \in H(\mathrm{div}, \Omega^{(t)}, \mathbb{R}^{2\times 2})$ and define

$$w_1 := \frac{1}{\gamma_1 + \nu} (w_4 + \nabla v). \tag{6.64}$$

Then, integration by parts yields

$$\langle (\gamma_1 + \nu) w_1 - \nabla \tilde{u}, \nabla \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^{2\times 2})} = \langle w_4, \nabla \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^{2\times 2})} = -\langle \mathrm{div}\, w_4, \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)}$$

for all $\varphi \in H_t$. Using this identity together with (6.63) we obtain the following constraint to the components w_2, w_3 and w_4

$$0 = \langle -\mathrm{div}\, w_4 + (\sigma(\gamma_1 + \nu - 1) - \gamma_2 \chi_{S_R \cap \Omega^{(t)}}) w_2 + \sigma w_3 - \sigma v, \varphi \rangle_{L^2(\Omega^{(t)},\mathbb{R}^2)}$$

for all $\varphi \in H_0^1(\Omega^{(t)}, \mathbb{R}^2)$ which can be solved for the third component, i.e., we fix

$$w_3 := \frac{1}{\sigma} \left(\mathrm{div}\, w_4 - (\sigma(\gamma_1 + \nu - 1) - \gamma_2 \chi_{S_R \cap \Omega^{(t)}}) w_2 + \sigma v \right). \tag{6.65}$$

Now, we are left with the computation of the remaining components $w_2 \in L^2(\Omega^{(t)}, \mathbb{R}^2)$ and $w_4 \in H(\mathrm{div}, \Omega^{(t)}, \mathbb{R}^{2\times 2})$ such that Theorem 6.8 provides “good” lower bounds. Therefore, we follow the lines in [74, Remark 10.26 (c)] again and approximately minimize $b_t(w, w)$

over w_2 and w_4 in a suitable finite element subspace (using the constraint (6.65)). Again, we use the definition of w_1 (cf. (6.64)) to obtain a functional only depending on the variables w_2 and w_4.

Finally, we present the functional used for the minimization process. Therefore, we insert the definitions of w_1 (see (6.64)) and w_3 (see (6.65)) into the bilinear form b_t (see (6.62)) which results in

$$\begin{aligned} b_t(w,w) = {} & \frac{1}{1+\nu}\|w_4 + \nabla v\|^2_{L^2(\Omega^{(t)},\mathbb{R}^{2\times 2})} + \sigma(\gamma_1 + \nu - 1)\|w_2\|^2_{L^2(\Omega^{(t)},\mathbb{R}^2)} \\ & - \gamma_2 \int_{S_R \cap \Omega^{(t)}} |w_2|^2 \,\mathrm{d}(x,y) \\ & + \frac{1}{\sigma}\|\operatorname{div} w_4 - (\sigma(\gamma_1 + \nu - 1) - \gamma_2 \chi_{S_R \cap \Omega^{(t)}})w_2 + \sigma v\|^2_{L^2(\Omega^{(t)},\mathbb{R}^2)}. \end{aligned}$$

Again, we can use the right-hand side of this identity to define a functional on the space $H(\operatorname{div}, \Omega^{(t)}, \mathbb{R}^{2\times 2}) \times L^2(\Omega^{(t)}, \mathbb{R}^2)$ which can be minimized on a suitable finite dimensional subspace. We note that by the choice of the space the additional smoothness assumptions on w_4 are satisfied by construction. In our applications we choose quadratic Lagrangian finite elements for each component again which yields an approximation in the desired space.

Remark 6.20. *Similar as already mentioned in Remark 6.14 we can reduce the computational effort by fixing $w_2 = \frac{1}{\lambda}v$ (with $\frac{1}{\lambda}$ denoting the approximate eigenvalue corresponding to v). In our applications in the finite element computation this strategy reduces the degrees of freedom on each nodal point from* 6 *to* 4 *which results in a faster homotopy computation. Nevertheless, in our applications it turned out that the computational time for the complete minimizing procedure is "reasonable" (compared to the "small" minimization procedure). Thus, we do not make use of this abbreviation since the lower bounds provided by the complete minimization process are "slightly" better.*

Constraint Homotopy

To get rid of the divergence condition in the space we perform a third homotopy. Therefore, we first define the following homotopy setting:

$$\begin{aligned} & H_t := H^1_0(S,\mathbb{R}^2) \text{ for all } t \in [0,1), \qquad H_1 := H(S), \\ & \langle u,\varphi\rangle_t := \langle u,\varphi\rangle_{H^1_0(S,\mathbb{R}^2)} \quad \text{for all } u,\varphi \in H_t,\ t \in [0,1], \\ & M_t(u,\varphi) := (\gamma_1+\nu)\langle u,\varphi\rangle_{H^1_0(S,\mathbb{R}^2)} - \gamma_2 \int_{S_R} u\cdot\varphi \,\mathrm{d}(x,y) \\ & \qquad\qquad + \frac{t}{1-t}\int_S \operatorname{div} u \operatorname{div}\varphi \,\mathrm{d}(x,y) \quad \text{for all } u,\varphi \in H_t,\ t\in[0,1), \\ & M_1(u,\varphi) := (\gamma_1+\nu)\langle u,\varphi\rangle_{H^1_0(S,\mathbb{R}^2)} - \gamma_2 \int_{S_R} u\cdot\varphi \,\mathrm{d}(x,y) \quad \text{for all } u,\varphi \in H_1 = H(S). \end{aligned} \tag{6.66}$$

To see that condition (6.24) is satisfied for this homotopy setting as well, we first observe that the mapping $t \mapsto \frac{t}{1-t}$ is monotonically increasing on $[0,1)$. Thus, for all $0 \le s \le t < 1$ we conclude

$$M_t(u,u) - M_s(u,u) = \left(\frac{t}{1-t} - \frac{s}{1-s}\right)\int_S (\operatorname{div} u)^2 \,\mathrm{d}(x,y) \ge 0$$

for all $u \in H_t = H_0^1(S, \mathbb{R}^2)$. Moreover, by the definition of $H(S)$ we have $\operatorname{div} u = 0$ for all $u \in H(S)$ which implies

$$M_1(u,u) - M_s(u,u) = 0 \quad \text{for all } u \in H_1 = H(S),\ 0 \le s \le 1.$$

Finally, the new (third) base problem reads as

$$\begin{aligned}
&u \in H_0^1(S, \mathbb{R}^2),\\
&(\gamma_1 + \nu)\langle u, \varphi\rangle_{H_0^1(S,\mathbb{R}^2)} - \gamma_2 \int_{S_R} u \cdot \varphi \,\mathrm{d}(x,y) = \lambda_\nu \langle u, \varphi\rangle_{H_0^1(S,\mathbb{R}^2)} \quad \text{for all } \varphi \in H_0^1(S, \mathbb{R}^2).
\end{aligned}$$

Splitting the strip into the three subdomains $(-\infty, -R) \times (0,1)$, $(R, \infty) \times (0,1)$ and S_R, using separation of variables and computing fundamental systems for the resulting ordinary differential equations on each of the subintervals $(-\infty, -R)$, (R, ∞) and $(-R, R)$ separately together with suitable matching conditions at the interfaces yields the desired eigenvalue information of the "final" base problem. For more details about the investigation of the base problem we refer the reader to Section 6.2.1.3. However, before considering the "final" base problem in detail we will have a closer look at the Goerisch setting needed for the constraint homotopy.

Remark 6.21. *Usually, in applications a Goerisch setting for the bilinear form M_1 is not needed since the final step is actually the original problem or the first problem of a new homotopy method (which is the case in our considerations), i.e., we hope that the required rough lower bound needed for the original eigenvalue problem (or the next homotopy as in our case) can be obtained from a problem with a sufficiently large homotopy parameter $t \in [0,1)$.*

In the further course, according to the previous Remark let $t \in [0,1)$ be a fixed homotopy parameter. In view of Goerisch's extension of Temple-Lehmann's Theorem (cf. Theorem 6.8) and its application in the homotopy method presented in Section 6.2.1 we choose $X := L^2(S, \mathbb{R}^{2\times 2}) \times L^2(S, \mathbb{R}^2) \times L^2(S) \times L^2(S, \mathbb{R}^2)$. Moreover, using the definition of the bilinear form M_t (cf. (6.66)) we define the bilinear form $b_t \colon X \times X \to \mathbb{R}$ needed in Corollary 6.9 by

$$\begin{aligned}
b_t(w, \hat{w}) := {}& (\gamma_1 + \nu)\langle w_1, \hat{w}_1\rangle_{L^2(S,\mathbb{R}^{2\times 2})} + \sigma(\gamma_1 + \nu - 1)\langle w_2, \hat{w}_2\rangle_{L^2(S,\mathbb{R}^2)}\\
& - \gamma_2 \int_{S_R} w_2 \cdot \hat{w}_2 \,\mathrm{d}(x,y) + \frac{t}{1-t}\langle w_3, \hat{w}_3\rangle_{L^2(S)} + \sigma\langle w_4, \hat{w}_4\rangle_{L^2(S,\mathbb{R}^2)}.
\end{aligned} \tag{6.67}$$

Then, obviously b_t is a symmetric bilinear form on X. Similar to the domain deformation homotopy case, we choose the shift parameter $\nu > 0$ such that

$$\nu \ge \frac{\gamma_2}{\sigma} - \gamma_1 + 1$$

which directly implies

$$\begin{aligned}
0 &\le (\sigma(\gamma_1 + \nu - 1) - \gamma_2)\langle v, v\rangle_{L^2(S,\mathbb{R}^2)}\\
&\le \sigma(\gamma_1 + \nu - 1)\langle v, v\rangle_{L^2(S,\mathbb{R}^2)} - \gamma_2 \int_{S_R} |v|^2 \,\mathrm{d}(x,y) \quad \text{for all } v \in L^2(S, \mathbb{R}^2)
\end{aligned}$$

proving that b_t is a positive semi-definite bilinear form (recall $t \in [0,1)$ and $\gamma_2 > 0$). Finally, we complete our Goerisch setting for the constraint homotopy with defining the operator T required for Theorem 6.8 and Corollary 6.9 respectively as follows

$$T\colon H_0^1(S,\mathbb{R}^2) \to X,\ Tu := (\nabla u, u, \operatorname{div} u, u)^T.$$

Furthermore, we compute

$$\begin{aligned} b_t(Tu,T\varphi) &= (\gamma_1+\nu)\langle \nabla u, \nabla\varphi\rangle_{L^2(S,\mathbb{R}^{2\times 2})} + \sigma(\gamma_1+\nu-1)\langle u,\varphi\rangle_{L^2(S,\mathbb{R}^2)} \\ &\qquad - \gamma_2 \int_{S_R} u\cdot\varphi \,\mathrm{d}(x,y) + \frac{t}{1-t}\langle \operatorname{div} u, \operatorname{div}\varphi\rangle_{L^2(S)} + \sigma\langle u,\varphi\rangle_{L^2(S,\mathbb{R}^2)} \\ &= M_t(u,\varphi) \quad \text{for all } u,\varphi \in H_0^1(S,\mathbb{R}^2), \end{aligned}$$

i.e., condition (6.15) in Theorem 6.8 is satisfied as well.

To apply Corollary 6.9 (or Theorem 6.8 respectively in the case of clustering eigenvalues), for some fixed $v \in H_0^1(S,\mathbb{R}^2)$ we need to compute $w = (w_1,w_2,w_3,w_4)^T \in X$ such that $b_t(w,T\varphi) = \langle v,\varphi\rangle_{H_0^1(S,\mathbb{R}^2)}$ for all $\varphi \in H_0^1(S,\mathbb{R}^2)$ (cf. (6.16)). Using the definition of b_t (see (6.67)) and the Goerisch operator T, condition (6.16) reads as

$$\begin{aligned} &(\gamma_1+\nu)\langle w_1,\nabla\varphi\rangle_{L^2(S,\mathbb{R}^{2\times 2})} + \sigma(\gamma_1+\nu-1)\langle w_2,\varphi\rangle_{L^2(S,\mathbb{R}^2)} - \gamma_2\int_{S_R} w_2\cdot\varphi\,\mathrm{d}(x,y) \\ &\quad + \frac{t}{1-t}\langle w_3,\operatorname{div}\varphi\rangle_{L^2(S,\mathbb{R}^2)} + \sigma\langle w_4,\varphi\rangle_{L^2(S,\mathbb{R}^2)} = \langle v,\varphi\rangle_{H_0^1(S,\mathbb{R}^2)} \quad \text{for all } \varphi \in H_0^1(S,\mathbb{R}^2). \end{aligned}$$

Reordering the terms above implies the following equivalent condition for the computation of w:

$$\begin{aligned} 0 &= \langle (\gamma_1+\nu)w_1 - \nabla v, \nabla\varphi\rangle_{L^2(S,\mathbb{R}^{2\times 2})} + \frac{t}{1-t}\langle w_3,\operatorname{div}\varphi\rangle_{L^2(S,\mathbb{R}^2)} \\ &\quad + \langle (\sigma(\gamma_1+\nu-1) - \gamma_2\chi_{S_R})w_2 + \sigma w_4 - \sigma v, \varphi\rangle_{L^2(S,\mathbb{R}^2)} \quad \text{for all } \varphi \in H_0^1(S,\mathbb{R}^2). \end{aligned} \tag{6.68}$$

Next, to be able to solve for one of the components of w we additionally assume that $w_3 \in H^1(S)$ and use integration by parts to obtain $\langle w_3, \operatorname{div}\varphi\rangle_{L^2(S,\mathbb{R}^2)} = -\langle \nabla w_3,\varphi\rangle_{L^2(S,\mathbb{R}^2)}$ for all $\varphi \in H_0^1(S,\mathbb{R}^2)$.

Moreover, similar to the previous Sections we introduce a new variable $w_5 \in H(\operatorname{div}, S, \mathbb{R}^{2\times 2})$. Then, we fix

$$w_1 := \frac{1}{\gamma_1+\nu}\,(w_5 + \nabla v) \tag{6.69}$$

and thus, using integration by parts we obtain

$$\langle (\gamma_1+\nu)w_1 - \nabla v, \nabla\varphi\rangle_{L^2(S,\mathbb{R}^{2\times 2})} = \langle w_5,\nabla\varphi\rangle_{L^2(S,\mathbb{R}^{2\times 2})} = -\langle \operatorname{div} w_5,\varphi\rangle_{L^2(S,\mathbb{R}^2)}$$

for all $\varphi \in H_0^1(S,\mathbb{R}^2)$. The previous identities together with condition (6.68) above imply

$$0 = \left\langle -\operatorname{div} w_5 + (\sigma(\gamma_1+\nu-1) - \gamma_2\chi_{S_R})w_2 - \frac{t}{1-t}\nabla w_3 + \sigma w_4 - \sigma v, \varphi \right\rangle_{L^2(S,\mathbb{R}^2)}$$

for all $\varphi \in H_0^1(S,\mathbb{R}^2)$ which can be solved for the fourth component again, i.e., we are in a position to define

$$w_4 := \frac{1}{\sigma}\left(\operatorname{div} w_5 - (\sigma(\gamma_1+\nu-1) - \gamma_2\chi_{S_R})w_2 + \frac{t}{1-t}\nabla w_3 + \sigma v\right). \tag{6.70}$$

Finally, we are left with the computation of the remaining components $w_2 \in L^2(S, \mathbb{R}^2)$, $w_3 \in H^1(S)$ and $w_5 \in H(\operatorname{div}, S, \mathbb{R}^{2\times 2})$ such that Theorem 6.8 provides "good" lower bounds. Therefore, we follow the lines in [74, Remark 10.26 (c)] again and approximately minimize $b_t(w, w)$ over w_2, w_3 and w_5 in a suitable finite element subspace with the constraint (6.70). We note that in this case as well we use the definition of w_1 (cf. (6.69)) to obtain a functional only depending on the variables w_2, w_3 and w_5.

To the end of this Section we shortly state the functional used in the minimization procedure. Therefore, we insert the definitions of w_1 (see (6.69)) and w_4 (see (6.70)) into the bilinear form b_t (see (6.67)) and obtain

$$\begin{aligned} &b_t(w, w) \\ &\quad = \frac{1}{\gamma_1 + \nu} \|w_5 + \nabla v\|^2_{L^2(S,\mathbb{R}^{2\times 2})} + \sigma(\gamma_1 + \nu - 1)\|w_2\|^2_{L^2(S,\mathbb{R}^2)} \\ &\quad\quad - \gamma_2 \int_{S_R} |w_2|^2 \,\mathrm{d}(x, y) + \frac{t}{1-t}\|w_3\|^2_{L^2(S)} \\ &\quad\quad + \frac{1}{\sigma} \left\| \operatorname{div} w_5 - (\sigma(\gamma_1 + \nu - 1) - \gamma_2 \chi_{S_R}) w_2 + \frac{t}{1-t} \nabla w_3 + \sigma v \right\|^2_{L^2(S,\mathbb{R}^2)} . \end{aligned}$$

Again, we can use the right-hand side of this identity to define a functional on the space $H(\operatorname{div}, S, \mathbb{R}^{2\times 2}) \times L^2(S, \mathbb{R}^2) \times H^1(S)$ which can be minimized on a suitable finite dimensional subspace. We note that the additional assumptions for w_3 and w_5 are modeled in the space itself and thus are satisfied by construction. Again, we use quadratic Lagrangian finite element in our applications to obtain the desired approximations.

6.2.1.3 Base Problem

We are left with the investigation of the base problem

$$\begin{aligned} &u \in H^1_0(S, \mathbb{R}^2), \\ &\gamma_1 \langle u, \varphi\rangle_{H^1_0(S,\mathbb{R}^2)} - \gamma_2 \int_{S_R} u \cdot \varphi \,\mathrm{d}(x, y) = \mu \langle u, \varphi\rangle_{H^1_0(S,\mathbb{R}^2)} \quad \text{for all } \varphi \in H^1_0(S, \mathbb{R}^2) \end{aligned} \tag{6.71}$$

on the strip S with constants $\gamma_1, \gamma_2 > 0$.

Remark 6.22. *Since the shift parameters $\nu > 0$ and $\hat{\nu} > 0$ respectively introduced in the previous Sections only shift the eigenvalues of the base problem we can omit the shifts in the context of the base problem.*

First of all, we have a closer look at the essential spectrum. Therefore, by Riesz' Representation Lemma we see that for any $u \in H^1_0(S, \mathbb{R}^2)$ there exists a unique $w_u \in H^1_0(S, \mathbb{R}^2)$ such that

$$\langle w_u, \varphi\rangle_{H^1_0(S,\mathbb{R}^2)} = \int_{S_R} u \cdot \varphi \,\mathrm{d}(x, y) \quad \text{for all } \varphi \in H^1_0(S, \mathbb{R}^2)$$

(note that the right-hand side defines a bounded linear functional on $H^1_0(\Omega, \mathbb{R}^2)$). Moreover, by Sobolev's Embedding Theorem (cf. [2, Theorem 5.4]) there exists a constant $C > 0$ such that $\|\varphi\|_{L^2(S,\mathbb{R}^2)} \leq C\|\varphi\|_{H^1_0(S,\mathbb{R}^2)}$ for all $\varphi \in H^1_0(S, \mathbb{R}^2)$. Hence, for the operator

$\mathcal{K}\colon H_0^1(S,\mathbb{R}^2) \to H_0^1(S,\mathbb{R}^2)$, $\mathcal{K}u := w_u$ (with w_u given by Riesz' Representation Lemma as above) we calculate

$$\|\mathcal{K}u\|^2_{H_0^1(S,\mathbb{R}^2)} = \int_{S_R} u \cdot \mathcal{K}u \,\mathrm{d}(x,y) \leq C\|u\|_{L^2(S_R,\mathbb{R}^2)}\|\mathcal{K}u\|_{H_0^1(S,\mathbb{R}^2)} \quad \text{for all } u \in H_0^1(S,\mathbb{R}^2),$$

i.e., we get $\|\mathcal{K}u\|_{H_0^1(S,\mathbb{R}^2)} \leq C\|u\|_{L^2(S_R,\mathbb{R}^2)}$ for all $u \in H_0^1(S,\mathbb{R}^2)$.

Now, let $(u_n)_{n\in\mathbb{N}}$ denote a bounded sequence in $H_0^1(S,\mathbb{R}^2)$. Since we have $S_R \subseteq S$ we see that $(u_n)_{n\in\mathbb{N}}$ is also bounded in $H^1(S_R,\mathbb{R}^2)$ and thus, Sobolev-Kondrachev-Rellich's Embedding Theorem (cf. [26, Theorem 1; p. 272]) yields the existence of a subsequence $(u_{n_k})_{k\in\mathbb{N}}$ converging in $L^2(S_R,\mathbb{R}^2)$, i.e., $(u_{n_k})_{k\in\mathbb{N}}$ is a Cauchy sequence in $L^2(S_R,\mathbb{R}^2)$. Hence, by the inequality above we obtain

$$\|\mathcal{K}u_{n_k} - \mathcal{K}u_{n_l}\|_{H_0^1(S,\mathbb{R}^2)} \leq C\|u_{n_k} - u_{n_l}\|_{L^2(S_R,\mathbb{R}^2)} \to 0 \quad \text{as } k,l \to \infty.$$

Finally, we obtain that $(\mathcal{K}u_{n_k})_{k\in\mathbb{N}}$ is a Cauchy sequence and thus convergent in $H_0^1(S,\mathbb{R}^2)$ which implies that $\mathcal{K}$ is a compact operator.

Using the operator $\mathcal{K}$ our base problem (6.71) reads as $\gamma_1 u - \gamma_2\mathcal{K}u = \mu u$ which, due to the compactness of $\mathcal{K}$, implies that the essential spectrum of our base problem (6.71) consists of the single value γ_1, i.e., in the sense of the homotopy method presented in the beginning of this Section we have $\sigma_0^{(0)} = \gamma_1$.

Since we are interested in the eigenvalues of the base problem below the essential spectrum, we can introduce the new eigenvalue parameter $\tilde{\mu} := \frac{\gamma_2}{\gamma_1-\mu}$ which, together with the fact that $\gamma_2 > 0$, implies the equivalent formulation of the base problem:

$$u \in H_0^1(S,\mathbb{R}^2),\ \langle u,\varphi\rangle_{H_0^1(S,\mathbb{R}^2)} = \tilde{\mu}\int_{S_R} u\cdot\varphi\,\mathrm{d}(x,y) \quad \text{for all } \varphi \in H_0^1(S,\mathbb{R}^2)$$

which has the strong formulation

$$u \in H^2(S,\mathbb{R}^2) \cap H_0^1(S,\mathbb{R}^2),\ -\Delta u + \sigma u = \tilde{\mu}\chi_{S_R}u.$$

Moreover, if $(u_1,u_2) \in H^2(S,\mathbb{R}^2) \cap H_0^1(S,\mathbb{R}^2)$ denotes an eigenfunction of this eigenvalue problem we see that (u_2,u_1), $(u_1,-u_2)$ and $(u_2,-u_1)$ are eigenfunctions corresponding to the same eigenvalue. Moreover, we calculate

$$\operatorname{span}\left\{\begin{pmatrix}u_1\\u_2\end{pmatrix},\begin{pmatrix}u_2\\u_1\end{pmatrix},\begin{pmatrix}u_1\\-u_2\end{pmatrix},\begin{pmatrix}u_2\\-u_1\end{pmatrix}\right\} = \operatorname{span}\left\{\begin{pmatrix}u_1\\0\end{pmatrix},\begin{pmatrix}u_2\\0\end{pmatrix},\begin{pmatrix}0\\u_1\end{pmatrix},\begin{pmatrix}0\\u_2\end{pmatrix}\right\}$$

which implies that it suffices to consider the scalar valued eigenvalue problem

$$u \in H^2(S) \cap H_0^1(S),\ -\Delta u + \sigma u = \tilde{\mu}\chi_{S_R}u \tag{6.72}$$

and finally count each eigenvalue with doubled multiplicity.

In the following we use separation of variables to determine the (smallest) eigenvalues of the scalar valued (base) problem (6.72). Hence, inserting the ansatz $u(x,y) = v(x)w(y)$ for all $(x,y) \in S$ into our eigenvalue problem we obtain

$$-v''(x)w(y) - v(x)w''(y) + (\sigma - \tilde{\mu}\chi_{S_R})\,v(x)w(y) = 0 \quad \text{for all } (x,y) \in S$$

which leads to the equations

$$-v''(x) = (\tilde{\mu}\chi_{[-R,R]} - \sigma + \tau)\, v(x) \quad \text{for all } x \in \mathbb{R} \quad \text{and} \quad w''(y) = \tau w(y) \quad \text{for all } y \in (0,1)$$

for v and w respectively where τ denotes a real constant. Thus, applying the boundary conditions in y-direction $u(x,0) = u(x,1) = 0$ for all $x \in \mathbb{R}$, i.e., $w(0) = w(1) = 0$, for any $n \in \mathbb{N}$ we obtain a solution

$$w_n(y) = \sin(n\pi y) \quad \text{for all } y \in [0,1],$$

where $\tau_n := -n^2\pi^2$ and these are all solutions of the equation for w. Now, inserting this result into the differential equation for v yields

$$-v''(x) = (\tilde{\mu}\chi_{[-R,R]} - (\sigma + n^2\pi^2))\, v(x) \quad \text{for all } x \in \mathbb{R}. \tag{6.73}$$

First, for each fixed $n \in \mathbb{N}$, we see that the eigenvalues of this eigenvalue problem satisfy $\tilde{\mu} > \sigma + n^2\pi^2$. To show this fact, we assume that there exists an eigenvalue $\bar{\mu} \leq \sigma + n^2\pi^2$ corresponding to an eigenfunction $\bar{v} \neq 0$. Then, multiplying the differential equation with $\bar{v}$, integrating over $\mathbb{R}$ and applying integration by parts (note the “boundary conditions” on $\bar{v}$ originating from those one the strip) yield

$$\begin{aligned}\int_{\mathbb{R}} (\bar{v}')^2 \,\mathrm{d}x &= \bar{\mu} \int_{-R}^{R} \bar{v}^2 \,\mathrm{d}x - (\sigma + n^2\pi^2) \int_{\mathbb{R}} \bar{v}^2 \,\mathrm{d}x \\ &\leq (\sigma + n^2\pi^2) \left(\int_{-R}^{R} \bar{v}^2 \,\mathrm{d}x - \int_{\mathbb{R}} \bar{v}^2 \,\mathrm{d}x \right) \leq 0.\end{aligned}$$

Hence, we conclude $\bar{v} = 0$ which is a contradiction, i.e., $\tilde{\mu} > \sigma + n^2\pi^2$ holds true.

To obtain the desired eigenvalues $\tilde{\mu}$, for fixed $n \in \mathbb{N}$, we first compute fundamental systems of the differential equation (6.73) on the subintervals $(-\infty, -R)$, $(-R, R)$ and (R, ∞) separately which together with suitable matching conditions yields all solutions to (6.73). Finally, this (together with the set of solutions for $w'' = \tau w$) yields all solutions to our scalar valued eigenvalue problem (6.72) (Note that the coefficients of (6.72) are constant with respect to y).

Thus, we exploit the facts $\sigma + n^2\pi^2 > 0$ and $\hat{\mu} := \tilde{\mu} - (\sigma + n^2\pi^2) > 0$ as well as the boundary conditions on v (originating from those one the strip) to calculate the desired fundamental systems on the subintervals which gives the general solution of (6.73):

$$v(x) = \begin{cases} A_1 \mathrm{e}^{\sqrt{\sigma + n^2\pi^2}(x+R)}, & x \in (-\infty, -R], \\ B_1 \sin(\sqrt{\hat{\mu}}x) + B_2 \cos(\sqrt{\hat{\mu}}x), & x \in (-R, R), \\ A_2 \mathrm{e}^{-\sqrt{\sigma + n^2\pi^2}(x-R)}, & x \in [R, \infty) \end{cases}$$

for constants $A_1, A_2, B_1, B_2 \in \mathbb{R}$.

Since we are interested in (non-trivial) solutions $u \in H^2(S) \cap H_0^1(S)$ we have to find appropriate matching conditions at $x = \pm R$. First, we require continuity of v at $x = \pm R$ which leads to the equations

$$\begin{aligned} A_1 &= -B_1 \sin(\sqrt{\hat{\mu}}R) + B_2 \cos(\sqrt{\hat{\mu}}R), \\ A_2 &= B_1 \sin(\sqrt{\hat{\mu}}R) + B_2 \cos(\sqrt{\hat{\mu}}R). \end{aligned}$$

Moreover, to guarantee that u becomes an element in $H^2(S)\cap H_0^1(S)$ the first derivative

$$v'(x) = \begin{cases} A_1\sqrt{\sigma+n^2\pi^2}\mathrm{e}^{\sqrt{\sigma+n^2\pi^2}(x+R)}, & x\in(-\infty,-R],\\ B_1\sqrt{\hat{\mu}}\cos(\sqrt{\hat{\mu}}x) - B_2\sqrt{\hat{\mu}}\sin(\sqrt{\hat{\mu}}x)), & x\in(-R,R),\\ -A_2\sqrt{\sigma+n^2\pi^2}\mathrm{e}^{-\sqrt{\sigma+n^2\pi^2}(x-R)}, & x\in[R,\infty). \end{cases}$$

has to be continuous (at $x=\pm R$) as well. Hence, we obtain the following matching conditions:

$$\begin{aligned} A_1\sqrt{\sigma+n^2\pi^2} &= B_1\sqrt{\hat{\mu}}\cos(\sqrt{\hat{\mu}}R) + B_2\sqrt{\hat{\mu}}\sin(\sqrt{\hat{\mu}}R),\\ -A_2\sqrt{\sigma+n^2\pi^2} &= B_1\sqrt{\hat{\mu}}\cos(\sqrt{\hat{\mu}}R) - B_2\sqrt{\hat{\mu}}\sin(\sqrt{\hat{\mu}}R)). \end{aligned}$$

Next, we rewrite the matching conditions as matrix vector product

$$\begin{pmatrix} 1 & 0 & \sin(\sqrt{\hat{\mu}}R) & -\cos(\sqrt{\hat{\mu}}R)\\ 0 & 1 & -\sin(\sqrt{\hat{\mu}}R) & -\cos(\sqrt{\hat{\mu}}R)\\ 1 & 0 & -\frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\cos(\sqrt{\hat{\mu}}R) & -\frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\sin(\sqrt{\hat{\mu}}R)\\ 0 & -1 & -\frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\cos(\sqrt{\hat{\mu}}R) & \frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\sin(\sqrt{\hat{\mu}}R) \end{pmatrix} \cdot \begin{pmatrix} A_1\\ A_2\\ B_1\\ B_2 \end{pmatrix} = 0. \tag{6.74}$$

Since there are non-trivial solutions of (6.73) if and only if the determinant of the matrix in (6.74) vanishes, we obtain the condition:

$$0 = \left(\frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\sin(\sqrt{\hat{\mu}}R) - \cos(\sqrt{\hat{\mu}}R)\right)\cdot\left(\frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\cos(\sqrt{\hat{\mu}}R) + \sin(\sqrt{\hat{\mu}}R)\right),$$

i.e.,

$$0 = \frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\sin(\sqrt{\hat{\mu}}R) - \cos(\sqrt{\hat{\mu}}R) \quad \text{or} \quad 0 = \frac{\sqrt{\hat{\mu}}}{\sqrt{\sigma+n^2\pi^2}}\cos(\sqrt{\hat{\mu}}R) + \sin(\sqrt{\hat{\mu}}R).$$

Since the roots of sin and cos do not coincide we conclude $\cos(\sqrt{\hat{\mu}}R)\neq 0$ and thus, we obtain the following transcendental equations for $\tilde{\mu}$ (recall $\hat{\mu}=\tilde{\mu}-(\sigma+n^2\pi^2)$):

$$0 = \tan(\sqrt{\tilde{\mu}-(\sigma+n^2\pi^2)}R) - \frac{\sqrt{\sigma+n^2\pi^2}}{\sqrt{\tilde{\mu}-(\sigma+n^2\pi^2)}} \tag{6.75}$$

or

$$0 = \tan(\sqrt{\tilde{\mu}-(\sigma+n^2\pi^2)}R) + \frac{\sqrt{\tilde{\mu}-(\sigma+n^2\pi^2)}}{\sqrt{\sigma+n^2\pi^2}}. \tag{6.76}$$

Remark 6.23. *We note that solutions $\tilde{\mu}$ of* (6.75) *correspond to symmetric solutions v of* (6.73)*, whereas solutions $\tilde{\mu}$ of* (6.76) *correspond to antisymmetric solutions v of* (6.73)*. Since this fact is not required in the further course of this thesis we omit the proof. However, it justifies the notation used in the further course.*

At this stage we want to emphasize that in all calculations above the parameter $n\in\mathbb{N}$ appears. Hence, the solutions $\tilde{\mu}$ of (6.75) and (6.76) actually depend on n.

To obtain the desired eigenvalues $\tilde{\mu}$ we treat both equations (6.75) and (6.76) separately. Starting with the "symmetric" case, we consider the transcendental equation (6.75) for

$\tilde{\mu}$. To solve this equation we first introduce the abbreviation $\zeta := \sqrt{\tilde{\mu} - (\sigma + n^2\pi^2)R} > 0$ which yields the equivalent equation

$$0 = \tan(\zeta) - \frac{\sqrt{\sigma + n^2\pi^2 R}}{\zeta} =: g_n^s(\zeta). \tag{6.77}$$

Hence, we are left with the computation of all roots of the functions $g_1^s, \dots, g_N^s \colon (0, \infty) \to \mathbb{R}$ within some suitable compact interval and an appropriate $N \in \mathbb{N}$. We want to emphasize that the computation (or at least enclosure) of all these (finitely many) roots can be done with the computer using interval arithmetic algorithms (cf. Section Interval Newton Method on p. 107).

To localize the roots of a function $g_n^s \colon (0, \infty) \to \mathbb{R}$ (recall that $\zeta > 0$ due to the fact $\tilde{\mu} > \sigma + n^2\pi^2$) for fixed $n \in \mathbb{N}$ we compute its derivative

$$(g_n^s)'(\zeta) = \frac{1}{\cos^2(\zeta)} + \frac{\sqrt{\sigma + n^2\pi^2 R}}{\zeta^2} > 0 \quad \text{for all } \zeta \in (0, \infty) \setminus \left\{ \frac{\pi}{2} + k\pi \colon k \in \mathbb{N}_0 \right\}.$$

which implies that g_n^s is strictly increasing on the intervals $(0, \frac{\pi}{2})$ and $(\frac{\pi}{2} + k\pi,\ \frac{\pi}{2} + (k+1)\pi)$ for all $k \in \mathbb{N}_0$. Moreover, for all $k \in \mathbb{N}_0$ we calculate

$$\lim_{\zeta \to 0+} g_n^s(\zeta) = -\infty, \qquad \lim_{\zeta \to (\frac{\pi}{2} + k\pi)-} g_n^s(\zeta) = +\infty, \qquad \lim_{\zeta \to (\frac{\pi}{2} + k\pi)+} g_n^s(\zeta) = -\infty,$$

i.e., by the intermediate value theorem we conclude that g_n^s has exactly one root in each of the intervals $(0, \frac{\pi}{2})$ and $(\frac{\pi}{2} + k\pi,\ \frac{\pi}{2} + (k+1)\pi)$ for all $k \in \mathbb{N}_0$. Hence, there are countably many roots of g_n^s which in the following will be denoted by $\zeta_{n,k}^s$ for all $k \in \mathbb{N}$. Using the definition of ζ we obtain the eigenvalues

$$\tilde{\mu}_{n,k}^s := \frac{(\zeta_{n,k}^s)^2}{R^2} + \sigma + n^2\pi^2 \quad \text{for all } n, k \in \mathbb{N}. \tag{6.78}$$

In the further course we treat the "antisymmetric" case analogously. Therefore, we investigate (6.76) which together with the abbreviation $\zeta := \sqrt{\tilde{\mu} - (\sigma + n^2\pi^2)R} > 0$ results in

$$0 = \tan(\zeta) + \frac{\zeta}{\sqrt{\sigma + n^2\pi^2 R}} =: g_n^a(\zeta). \tag{6.79}$$

Hence, we are left with the computation of all roots of the functions $g_1^a, \dots, g_N^a \colon (0, \infty) \to \mathbb{R}$ within some suitable compact interval and some appropriate $N \in \mathbb{N}$ again. As before, for fixed $n \in \mathbb{N}$ we compute the derivative of $g_n^a \colon (0, \infty) \to \mathbb{R}$:

$$(g_n^a)'(\zeta) = \frac{1}{\cos^2(\zeta)} + \frac{1}{\sqrt{\sigma + n^2\pi^2 R}} > 0 \quad \text{for all } \zeta \in (0, \infty) \setminus \left\{ \frac{\pi}{2} + k\pi \colon k \in \mathbb{N}_0 \right\}$$

which implies that g_n^a is strictly increasing on the intervals $(0, \frac{\pi}{2})$ and $(\frac{\pi}{2} + k\pi,\ \frac{\pi}{2} + (k+1)\pi)$ for all $k \in \mathbb{N}_0$. In almost the same manner as above, we obtain

$$\lim_{\zeta \to 0+} g_n^a(\zeta) = 0, \qquad \lim_{\zeta \to (\frac{\pi}{2} + k\pi)-} g_n^a(\zeta) = +\infty, \qquad \lim_{\zeta \to (\frac{\pi}{2} + k\pi)+} g_n^a(\zeta) = -\infty,$$

i.e., g_n^a has no root in $(0, \frac{\pi}{2})$ and again by the intermediate value theorem exactly one root in each of the intervals $(\frac{\pi}{2} + k\pi,\ \frac{\pi}{2} + (k+1)\pi)$ for all $k \in \mathbb{N}_0$. Thus, by $\zeta_{n,k}^a$ for all $k \in \mathbb{N}$

we denote the countably many roots of g_n^a which (after a retransformation) results in the eigenvalues

$$\tilde{\mu}_{n,k}^a := \frac{(\zeta_{n,k}^a)^2}{R^2} + \sigma + n^2\pi^2 \quad \text{for all } n, k \in \mathbb{N} \tag{6.80}$$

similar as before.

As already mentioned, we are interested in the (smallest) eigenvalues of our base problem (6.71) located below some constant $\rho_0 < \sigma_0^{(0)} = \gamma_1$. Using the definition of the new eigenvalue parameter $\tilde{\mu}$ we obtain the corresponding bound $\frac{\gamma_2}{\gamma_1 - \rho_0}$ for the eigenvalues of our scalar value eigenvalue problem (6.72). Thus, it suffices to consider those eigenvalues $\tilde{\mu}_{n,k}^s$ and $\tilde{\mu}_{n,k}^a$ such that

$$\tilde{\mu}_{n,k}^s \leq \frac{\gamma_2}{\gamma_1 - \rho_0} \quad \text{and} \quad \tilde{\mu}_{n,k}^a \leq \frac{\gamma_2}{\gamma_1 - \rho_0} \tag{6.81}$$

which yields that only finitely many values of n have to be considered for the computation of our desired eigenvalues, i.e., for all $n \in \mathbb{N}$ with $\sigma + n^2\pi^2 > \frac{\gamma_2}{\gamma_1 - \rho_0}$ condition (6.81) is not satisfied anyway (cf. (6.78) and (6.80)). This leads to the upper bound

$$n_{\max} := \left\lfloor \sqrt{\frac{1}{\pi^2}\left(\frac{\gamma_2}{\gamma_1 - \rho_0} - \sigma\right)} \right\rfloor$$

for the values of n, i.e., for all $1 \leq n \leq n_{\max}$ we have to compute the first roots $\tilde{\mu}_{n,k}^s$ and $\tilde{\mu}_{n,k}^a$ respectively (i.e., for finitely many $1 \leq k \leq K_s(n)$ and $1 \leq k \leq K_a(n)$ respectively) such that (6.81) holds true. This strategy results in finitely many eigenvalues $\tilde{\mu}_1, \ldots, \tilde{\mu}_N$ (for some $N \in \mathbb{N}$) of the scalar valued problem (6.72). Using the transformation $\mu = \gamma_1 - \frac{\gamma_2}{\tilde{\mu}}$ we obtain the desired first eigenvalues $\mu_1, \ldots, \mu_N$ of our base problem (6.71) located below the given bound ρ_0.

Finally, we are left with the computation of the roots $\tilde{\mu}_{n,k}^s$ for all $1 \leq n \leq n_{\max}$, $1 \leq k \leq K_s(n)$ and $\tilde{\mu}_{n,k}^a$ for all $1 \leq n \leq n_{\max}$, $1 \leq k \leq K_a(n)$ respectively (which a posteriori characterizes the suitable compact interval and the number $N \in \mathbb{N}$ mentioned after (6.77) and (6.79) respectively). However, since equations (6.77) and (6.79) are transcendental finding the exact roots is challenging or even impossible. Nevertheless, since we only need to determine finitely many roots we can use an Interval Newton Method to enclose these (first) roots which is sufficient to provide the desired information (especially an index information is guaranteed) about the eigenvalues of the base problem (6.71).

Interval Newton Method

To the end of this Section we briefly recall some details of the interval Newton method needed to enclose all zeros of a function in a given compact interval. Since in our applications above all functions have simple roots we only present a version of the interval Newton method dealing with this case. However, in the literature there are more general versions treating the case of multiple roots as well (see e.g. [3]). The version we are going to present can be found for instance in [38]. Moreover, in the following by $[x] \subseteq \mathbb{R}$ we denote a compact real interval and by $\text{mid}([x])$ its midpoint.

Now, let $f\colon \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function and $[x]_0 \subseteq \mathbb{R}$ denote a compact real interval such that $0 \notin f'([x]_0)$. From the latter condition we conclude that f has at most one zero $x^* \in [x]_0$. Then, with

$$N([x]) := \operatorname{mid}([x]) - \frac{f(\operatorname{mid}([x])}{f'([x])}$$

the $(k+1)$st iterate of the interval Newton method is defined by

$$[x]_{k+1} := [x]_k \cap N([x]_k) \quad \text{for all } k \in \mathbb{N}_0.$$

Due to the intersection of $N([x]_k)$ with $[x]_k$ the interval Newton method cannot diverge, i.e., each iterate of the method remains bounded. Furthermore, by [38, Theorem 6.1] we obtain:

(a) Every zero $x^* \in [x]$ of f satisfies $x^* \in N([x])$.

(b) If $N([x]) \cap [x] = \emptyset$, then there exists no zero of f in $[x]$.

(c) If $N([x]) \overset{\circ}{\subset} [x]$ (i.e., the interval $N([x])$ is contained in the interior of $[x]$; cf. [38, Section 3.1]), then there exists a unique zero of f in $[x]$ and hence in $N([x])$.

In particular, conditions (a) and (b) yield that if $[x]_{k_0} = \emptyset$ for some $k_0 \in \mathbb{N}$ then $[x]_0$ does not contain a zero of f.

As mentioned above, we are interested in enclosing the "first" roots of the functions g_s and g_a respectively. Therefore, for each open interval (on which the functions g_s and g_a are strictly increasing; cf. previous Section) we choose some compact subinterval $[x]_0$ and check

$$0 \in f([x]_0), \quad 0 \notin f'([x]_0) \quad \text{and} \quad f(\inf([x]_0))f(\sup([x]_0)) < 0$$

a priori. If these conditions are satisfied, we perform the interval Newton algorithm presented above until either $[x]_{k_1} = [x]_{k_1+1}$ for some $k_1 \in \mathbb{N}$, or the diameter of $[x]_{k_1+1}$ is smaller than a prescribed tolerance. Then, $[x]_{k_1+1}$ encloses the (unique) root in the given interval.

Moreover, from the previous Sections it is known that in each interval (on which g_s and g_a are strictly increasing) there exists a single root. Hence, we might expect that for a suitably large initial subinterval $[x]_0$ the conditions above are actually satisfied, i.e., if our a priori check yields $0 \notin f([x]_0)$ we have to enlarge the "initial interval" $[x]_0$ a bit and check the a priori conditions again with this larger initial interval $[x]_0$.

6.2.1.4 Computation of the Lower Bounds κ and $\hat{\kappa}$

To guarantee the success of the extended coefficient homotopy (second approach in Section 6.2.1.2) we are left with the computation of constants $\kappa, \hat{\kappa} > 0$ such that

$$\langle \Phi^{-1} \mathrm{L}_U u, \Phi^{-1} \mathrm{L}_U u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq \kappa \langle u, u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega) \tag{6.82}$$

and

$$\langle \Phi^{-1} \hat{\mathrm{L}}_U u, \Phi^{-1} \hat{\mathrm{L}}_U u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq \hat{\kappa} \langle u, u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega) \tag{6.83}$$

respectively.

Exemplary, we start with the computation of κ and make some remarks on the procedure for the "adjoint" constant $\hat{\kappa}$ afterwards. The same calculations as in Section 6.2.1.1 (cf. definition (6.33) and the calculations before) show

$$\begin{aligned}
&\langle \Phi^{-1} \operatorname{L}_U u, \Phi^{-1} \operatorname{L}_U u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\qquad = \langle u, u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \int_\Omega u^T G_U u \,\mathrm{d}(x,y) + \langle \Phi^{-1}(\operatorname{B}_U - \sigma) u, \Phi^{-1}(\operatorname{B}_U - \sigma) u \rangle_{H_0^1(\Omega,\mathbb{R}^2)}
\end{aligned}$$

with $G_U := Re[\nabla U + (\nabla U)^T] - 2\sigma \,\mathrm{id}$.

For the moment, let $u \in H(\Omega) \subseteq H_0^1(\Omega, \mathbb{R}^2)$ be an arbitrary fixed function and let u_S denote its extension by zero which can be read as a function in $H_0^1(S, \mathbb{R}^2)$. Thus, we directly obtain the equality $\langle u, u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \langle u_S, u_S \rangle_{H_0^1(S,\mathbb{R}^2)}$.

Using the Fourier transform $\mathcal{F}_x$ in x-direction introduced in Section 2.4 and its isometric property (cf. Remark 2.6 (i)) together with Lemma 2.7 (i) we obtain

$$\begin{aligned}
\langle u, u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} &= \langle u_S, u_S \rangle_{H_0^1(S,\mathbb{R}^2)} \\
&= \int_{-\infty}^{\infty} \int_0^1 \left(\left| \frac{\partial u_S}{\partial x}(x,y) \right|^2 + \left| \frac{\partial u_S}{\partial y}(x,y) \right|^2 + \sigma \left| u_S(x,y) \right|^2 \right) \mathrm{d}y \,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \int_0^1 \left(\left| \mathcal{F}_x\left[\frac{\partial u_S}{\partial x} \right](\xi,y) \right|^2 + \left| \mathcal{F}_x\left[\frac{\partial u_S}{\partial y} \right](\xi,y) \right|^2 + \sigma \left| \mathcal{F}_x[u_S](\xi,y) \right|^2 \right) \mathrm{d}y \,\mathrm{d}\xi \\
&= \int_{-\infty}^{\infty} \int_0^1 \left(\left| \mathrm{i}\xi \mathcal{F}_x[u_S](\xi,y) \right|^2 + \left| \frac{\partial \mathcal{F}_x[u_S]}{\partial y}(\xi,y) \right|^2 + \sigma \left| \mathcal{F}_x[u_S](\xi,y) \right|^2 \right) \mathrm{d}y \,\mathrm{d}\xi \\
&= \int_{-\infty}^{\infty} \int_0^1 \left((\xi^2 + \sigma) \left| \mathcal{F}_x[u_S](\xi,y) \right|^2 + \left| \frac{\partial \mathcal{F}_x[u_S]}{\partial y}(\xi,y) \right|^2 \right) \mathrm{d}y \,\mathrm{d}\xi.
\end{aligned}$$

Next, we can expand $\mathcal{F}_x[u_S]$ in y-direction via Fourier series. Since $\mathcal{F}_x[u_S]$ satisfies Dirichlet boundary conditions in y-direction, i.e., $\mathcal{F}_x[u_S](\xi,0) = \mathcal{F}_x[u_S](\xi,1) = 0$ for all $\xi \in \mathbb{R}$, we can use the basis functions φ_n defined as follows

$$\varphi_n \colon [0,1] \to \mathbb{R}^2,\ \varphi_n(y) := \begin{cases} \begin{pmatrix} \sqrt{2}\sin(k\pi y) \\ 0 \end{pmatrix}, & n = 2k-1,\ k \in \mathbb{N}, \\ \begin{pmatrix} 0 \\ \sqrt{2}\sin(k\pi y) \end{pmatrix}, & n = 2k,\ k \in \mathbb{N} \end{cases} \tag{6.84}$$

for all $n \in \mathbb{N}$ to obtain the representation formula

$$\mathcal{F}_x[u_S](\xi,y) = \sum_{n=1}^{\infty} u_n(\xi) \varphi_n(y)$$

(converging in $L^2(S, \mathbb{C}^2)$) with "coefficients" $u_n \in L^2(\mathbb{R}, \mathbb{C})$ satisfying

$$\int_{-\infty}^{\infty} (1 + \xi^2) \left| u_n(\xi) \right|^2 \,\mathrm{d}\xi < \infty.$$

Furthermore, direct calculations show

$$\int_0^1 \varphi_n(y)\varphi_m(y)\,\mathrm{d}y = \delta_{n,m} \quad \text{and} \quad \int_0^1 \varphi_n'(y)\varphi_m'(y)\,\mathrm{d}y = \left\lceil \frac{n}{2} \right\rceil^2 \pi^2 \delta_{n,m}$$

for all $n, m \in \mathbb{N}$ which together with the previous identities implies

$$\begin{aligned}\langle u, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} &= \sum_{n,m=1}^{\infty} \int_{-\infty}^{\infty} \left(u_n(\xi)\overline{u_m(\xi)} \int_0^1 \left[(\xi^2+\sigma)\varphi_n(y)\varphi_m(y) + \varphi_n'(y)\varphi_m'(y)\right]\,\mathrm{d}y \right) \mathrm{d}\xi \\ &= \sum_{n=1}^{\infty} \int_{-\infty}^{\infty} |u_n(\xi)|^2 \left(\xi^2 + \sigma + \left\lceil \frac{n}{2} \right\rceil^2 \pi^2 \right) \mathrm{d}\xi.\end{aligned}$$

With new “coefficients”

$$v_n(\xi) \coloneqq u_n(\xi)\sqrt{\xi^2 + \sigma + \left\lceil \frac{n}{2} \right\rceil^2 \pi^2} \quad \text{for all } \xi \in \mathbb{R},\ n \in \mathbb{N}$$

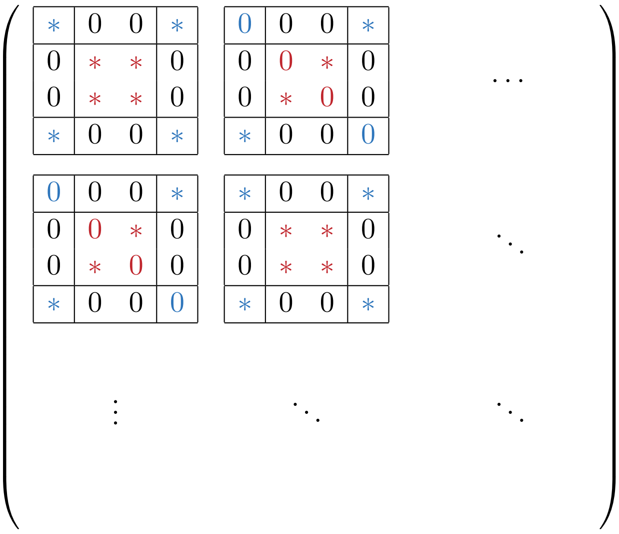
we directly obtain the identity

$$\langle u, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} = \sum_{n=1}^{\infty} \int_{-\infty}^{\infty} |v_n(\xi)|^2 \,\mathrm{d}\xi. \tag{6.85}$$

Since u_S is the extension of u by zero we have $\int_\Omega u^T G_U u\,\mathrm{d}(x,y) = \int_S u_S^T G_U u_S\,\mathrm{d}(x,y)$. Hence, using the definition of G_U together the same techniques as above we calculate

$$\begin{aligned}&\int_\Omega u^T G_U u\,\mathrm{d}(x,y) \\ &\quad= \int_S u_S^T G_U u_S\,\mathrm{d}(x,y) \\ &\quad= Re \sum_{n,m=1}^{\infty} \int_{-\infty}^{\infty} u_n(\xi)\overline{u_m(\xi)} \int_0^1 \varphi_n(y)^T \begin{pmatrix} 0 & 1-2y \\ 1-2y & 0 \end{pmatrix} \varphi_m(y)\,\mathrm{d}y\,\mathrm{d}\xi \\ &\qquad - 2\sigma \sum_{n,m=1}^{\infty} \int_{-\infty}^{\infty} u_n(\xi)\overline{u_m(\xi)} \int_0^1 \varphi_n(y)\varphi_m(y)\,\mathrm{d}y\,\mathrm{d}\xi \\ &\quad= Re \sum_{n,m=1}^{\infty} \int_{-\infty}^{\infty} \frac{v_n(\xi)}{\sqrt{\xi^2+\sigma+\lceil\frac{n}{2}\rceil^2\pi^2}} \frac{\overline{v_m(\xi)}}{\sqrt{\xi^2+\sigma+\lceil\frac{m}{2}\rceil^2\pi^2}} \int_0^1 \varphi_n(y)^T \begin{pmatrix} 0 & 1-2y \\ 1-2y & 0 \end{pmatrix} \varphi_m(y)\,\mathrm{d}y\,\mathrm{d}\xi \\ &\qquad - 2\sigma \sum_{n,m=1}^{\infty} \int_{-\infty}^{\infty} \frac{v_n(\xi)}{\sqrt{\xi^2+\sigma+\lceil\frac{n}{2}\rceil^2\pi^2}} \frac{\overline{v_m(\xi)}}{\sqrt{\xi^2+\sigma+\lceil\frac{m}{2}\rceil^2\pi^2}} \delta_{n,m}\,\mathrm{d}\xi.\end{aligned}$$

Again using the “relatively rough” estimate $\langle \Phi^{-1}(\mathrm{B}_U - \sigma)u, \Phi^{-1}(\mathrm{B}_U - \sigma)u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq 0$ and



$$\begin{pmatrix} \begin{array}{|c|cc|c|} \hline * & 0 & 0 & * \\ \hline 0 & * & * & 0 \\ 0 & * & * & 0 \\ \hline * & 0 & 0 & * \\ \hline \end{array} & \begin{array}{|c|cc|c|} \hline 0 & 0 & 0 & * \\ \hline 0 & 0 & * & 0 \\ 0 & * & 0 & 0 \\ \hline * & 0 & 0 & 0 \\ \hline \end{array} & \cdots \\ \begin{array}{|c|cc|c|} \hline 0 & 0 & 0 & * \\ \hline 0 & 0 & * & 0 \\ 0 & * & 0 & 0 \\ \hline * & 0 & 0 & 0 \\ \hline \end{array} & \begin{array}{|c|cc|c|} \hline * & 0 & 0 & * \\ \hline 0 & * & * & 0 \\ 0 & * & * & 0 \\ \hline * & 0 & 0 & * \\ \hline \end{array} & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}$$

Figure 6.2: Structure of the multiplication operator A

combining the results from above we end up with

$$\begin{aligned} &\langle \Phi^{-1}\,\mathrm{L}_U\,u, \Phi^{-1}\,\mathrm{L}_U\,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \\ &\quad \geq \sum_{n,m=1}^{\infty} \int_{-\infty}^{\infty} v_n(\xi)\overline{v_m(\xi)} \Bigg[\left(1 - \frac{2\sigma}{\sqrt{\xi^2+\sigma+\lceil\frac{n}{2}\rceil^2\pi^2}\sqrt{\xi^2+\sigma+\lceil\frac{m}{2}\rceil^2\pi^2}}\right)\delta_{n,m} \\ &\qquad + \frac{Re}{\sqrt{\xi^2+\sigma+\lceil\frac{n}{2}\rceil^2\pi^2}\sqrt{\xi^2+\sigma+\lceil\frac{m}{2}\rceil^2\pi^2}} \int_0^1 \varphi_n(y)^T \begin{pmatrix} 0 & 1-2y \\ 1-2y & 0 \end{pmatrix} \varphi_m(y)\,\mathrm{d}y \Bigg]\,\mathrm{d}\xi \\ &\quad =: \sum_{n,m=1}^{\infty} \int_{-\infty}^{\infty} v_n(\xi)\overline{v_m(\xi)}(A(\xi))_{n,m}\,\mathrm{d}\xi \end{aligned} \tag{6.86}$$

with an infinite dimensional multiplication operator A. Using Lemma A.13 we obtain

$$(A(\xi))_{n,m} = \begin{cases} \frac{\xi^2-\sigma+\lceil\frac{n}{2}\rceil^2\pi^2}{\xi^2+\sigma+\lceil\frac{n}{2}\rceil^2\pi^2}, & n=m, \\ \frac{16\lceil\frac{n}{2}\rceil\lceil\frac{m}{2}\rceil Re}{\pi^2\left(\lceil\frac{n}{2}\rceil^2-\lceil\frac{m}{2}\rceil^2\right)^2\sqrt{\xi^2+\sigma+\lceil\frac{n}{2}\rceil^2\pi^2}\sqrt{\xi^2+\sigma+\lceil\frac{m}{2}\rceil^2\pi^2}}, & (n+m) \bmod 4 = 1, \\ 0, & \text{otherwise} \end{cases} \tag{6.87}$$

for all $n,m \in \mathbb{N}$ and $\xi \in \mathbb{R}$, i.e., $A(\xi)$ is of the form presented in Figure 6.2.

Now, we are aiming at a real constant κ such that

$$x^T A(\xi)\overline{x} \geq \kappa\|x\|^2_{\ell^2(\mathbb{C})} \quad \text{for all } x \in \ell^2(\mathbb{C}),\ \xi \in \mathbb{R}. \tag{6.88}$$

If κ is explicitly at hand (and positive), together with the representation of the inner product (cf. (6.85)) and estimate (6.86) from above we calculate

$$\begin{aligned}\langle \Phi^{-1}\,\mathrm{L}_U\,u, \Phi^{-1}\,\mathrm{L}_U\,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} &\geq \sum_{n,m=1}^{\infty}\int_{-\infty}^{\infty} v_n(\xi)\overline{v_m(\xi)}(A(\xi))_{n,m}\,\mathrm{d}\xi \\ &\geq \kappa\sum_{n=1}^{\infty}\int_{-\infty}^{\infty} |v_n(\xi)|^2\,\mathrm{d}\xi \\ &= \kappa\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)}\end{aligned}$$

which (since $u \in H(\Omega)$ was arbitrary) directly implies the desired estimate (6.82).

In the further course we give some details about the computation of the constant κ satisfying (6.88). Therefore, we first exploit the structure of the multiplication operator A and define the "sub operator"

$$(\tilde{A}(\xi))_{k,l} := \begin{cases} \frac{\xi^2-\sigma+k^2\pi^2}{\xi^2+\sigma+k^2\pi^2}, & k=l, \\ \frac{16klRe}{\pi^2(k^2-l^2)^2\sqrt{\xi^2+\sigma+k^2\pi^2}\sqrt{\xi^2+\sigma+l^2\pi^2}}, & (k+l) \bmod 2 = 1, \\ 0, & \text{otherwise}\end{cases} \tag{6.89}$$

for all $k,l \in \mathbb{N}$ and $\xi \in \mathbb{R}$. In view of Figure 6.2 the operator $\tilde{A}(\xi)$ consists of the red or blue entries respectively, i.e., $\tilde{A}(\xi)$ is now of the form illustrated in Figure 6.3. Thus, splitting also an element $x \in \ell^2(\mathbb{C})$ into two parts $y, z \in \ell^2(\mathbb{C})$, i.e., we set

$$y_k := \begin{cases} x_{2k}, & k \bmod 2 = 0, \\ x_{2k-1}, & k \bmod 2 = 1\end{cases} \quad\text{and}\quad z_k := \begin{cases} x_{2k-1}, & k \bmod 2 = 0, \\ x_{2k}, & k \bmod 2 = 1,\end{cases}$$

leads to the identity

$$x^T A(\xi)\overline{x} = y^T\tilde{A}(\xi)\overline{y} + z^T\tilde{A}(\xi)\overline{z} \quad \text{for all } \xi \in \mathbb{R}. \tag{6.90}$$

We note that reordering the terms is possible since all series appearing in this situation are absolutely convergent which can be seen by the arguments presented in the further course.

Thus, to obtain the desired constant κ we split the operator $\tilde{A}(\xi)$ into finite dimensional and an infinite dimensional parts. Therefore, we fix some "size" $N \in \mathbb{N}$ even and for all $\xi \in \mathbb{R}$ we define $\tilde{A}_0(\xi) \in \mathbb{C}^{N\times N}$ with $(\tilde{A}_0(\xi))_{k,l} := (\tilde{A}(\xi))_{k,l}$ for all $k,l = 1,\dots,N$, $\tilde{A}_1(\xi)$ such that $(\tilde{A}_1(\xi))_{k,l} := (\tilde{A}(\xi))_{k,l+N}$ for all $k = 1,\dots,N$, $l \in \mathbb{N}$ and the infinite dimensional operator $\tilde{A}_2(\xi)$ with $(\tilde{A}_2(\xi))_{k,l} := (\tilde{A}(\xi))_{k+N,l+N}$ for all $k,l \in \mathbb{N}$. Thus, $\tilde{A}(\xi)$ can be represented in terms of $\tilde{A}_0(\xi)$, $\tilde{A}_1(\xi)$ and $\tilde{A}_2(\xi)$ (cf. Figure 6.3):

$$\tilde{A}(\xi) = \left(\begin{array}{c|c} \tilde{A}_0(\xi) & \tilde{A}_1(\xi) \\ \hline \overline{\tilde{A}_1(\xi)}^T & \tilde{A}_2(\xi)\end{array}\right) \quad \text{for all } \xi \in \mathbb{R}. \tag{6.91}$$

Moreover, by $\tilde{D}(\xi)$ we denote the diagonal part of $\tilde{A}_2(\xi)$, i.e., we have the identity $\tilde{D}_{k,k}(\xi) = \tilde{A}_{k+N,k+N}(\xi)$ for all $\xi \in \mathbb{R}$, $k \in \mathbb{N}$ and zero entries otherwise.

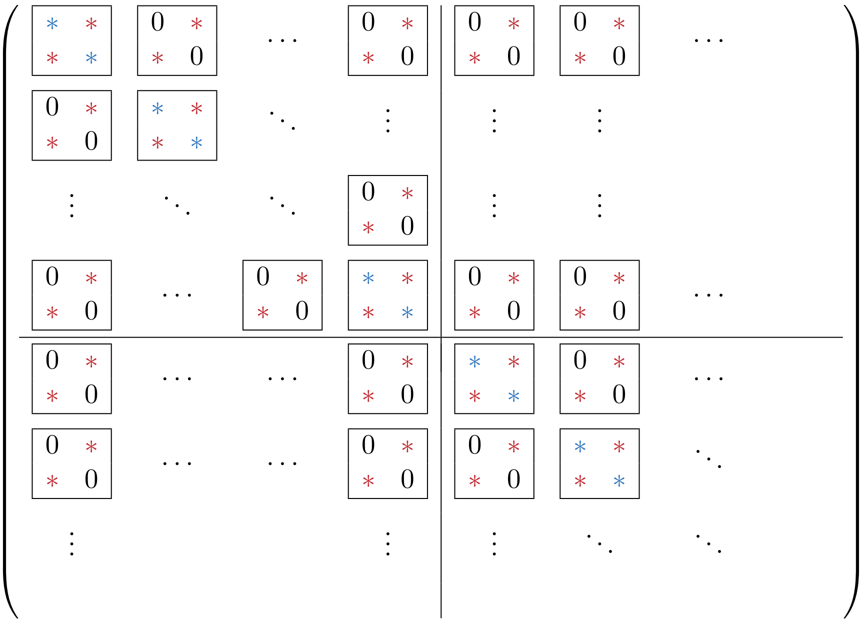


Figure 6.3: Structure of the multiplication operator $\tilde{A}$

In the following, we suppose that functions $\theta_0, \theta_1 \colon \mathbb{R} \to \mathbb{R}$ and $\theta_2, \theta_3 \colon \mathbb{R} \to [0,\infty)$ are known explicitly such that

1. $\lambda_{\min}(\tilde{A}_0(\xi)) \geq \theta_0(\xi),$ (6.92)

2. $\lambda_{\min}(\tilde{D}(\xi)) \geq \theta_1(\xi),$ (6.93)

3. $$\sum_{k=1}^{N}\sum_{l=1}^{\infty} |(\tilde{A}_1(\xi))_{k,l}|^2 = \sum_{k=1}^{N}\sum_{l=N+1}^{\infty} |(\tilde{A}(\xi))_{k,l}|^2 \leq \theta_2(\xi), \tag{6.94}$$

4. $$\sum_{k,l=1}^{\infty} |(\tilde{A}_2(\xi) - \tilde{D}(\xi))_{k,l}|^2 = \sum_{\substack{k,l=N+1\\ k\neq l}}^{\infty} |(\tilde{A}(\xi))_{k,l}|^2 \leq \theta_3(\xi) \tag{6.95}$$

for all $\xi \in \mathbb{R}$. We postpone the computation of the desired functions $\theta_1, \ldots, \theta_3$ to the end of this Subsection and present the abstract computation of κ first.

Remark 6.24. *We note that the absolute values $|(\tilde{A}(\xi))_{k,l}|$ of the off-diagonal entries, i.e., for $k \neq l$, become "smaller" as the values of $k, l \in \mathbb{N}$ increase. Thus, if the size N is "large" the error terms represented by our functions θ_2 and θ_3 also get "small".*

Now, let $y \in \ell^2(\mathbb{C})$ and define $v := (y_1, \ldots, y_N) \in \mathbb{C}^N$ and $w := (y_{N+1}, \ldots) \in \ell^2(\mathbb{C})$. Thus, we obtain the identity $\|y\|^2_{\ell^2(\mathbb{C})} = \|v\|_2^2 + \|w\|^2_{\ell^2(\mathbb{C})}$ and for all $\xi \in \mathbb{R}$ we calculate

$$\begin{aligned} y^T\tilde{A}(\xi)\overline{y} &= v^T\tilde{A}_0(\xi)\overline{v} + v^T\tilde{A}_1(\xi)\overline{w} + w^T\overline{\tilde{A}_1(\xi)}^T\overline{v} + w^T\tilde{A}_2(\xi)\overline{w} \qquad (6.96)\\ &= v^T\tilde{A}_0(\xi)\overline{v} + v^T\tilde{A}_1(\xi)\overline{w} + w^T\overline{\tilde{A}_1(\xi)}^T\overline{v} + w^T(\tilde{A}_2(\xi) - \tilde{D}(\xi))\overline{w} + w^T\tilde{D}(\xi)\overline{w}. \end{aligned}$$

Using Cauchy-Schwarz' inequality and assumption (6.94) we obtain

$$\begin{aligned} v^T\tilde{A}_1(\xi)\overline{w} &\le \|v\|_2\|\tilde{A}_1(\xi)\overline{w}\|_2 = \|v\|_2\left(\sum_{k=1}^{N}\left|\sum_{l=1}^{\infty}(\tilde{A}_1(\xi))_{k,l}\,\overline{w_l}\right|^2\right)^{\frac{1}{2}} \\ &\le \|v\|_2\left(\sum_{k=1}^{N}\left(\sum_{l=1}^{\infty}\left|(\tilde{A}_1(\xi))_{k,l}\right|^2\right)\left(\sum_{l=1}^{\infty}|w_l|^2\right)\right)^{\frac{1}{2}} \le \sqrt{\theta_2(\xi)}\|v\|_2\|w\|_{\ell^2(\mathbb{C})} \end{aligned}$$

for all $\xi\in\mathbb{R}$, directly implying $w^T\overline{\tilde{A}_1(\xi)}^T\overline{v} \le \sqrt{\theta_2(\xi)}\|v\|_2\|w\|_{\ell^2(\mathbb{C})}$. Similar arguments but now using (6.95) yield

$$w^T(\tilde{A}_2(\xi)-\tilde{D}(\xi))\overline{w} \le \sqrt{\theta_3(\xi)}\|w\|^2_{\ell^2(\mathbb{C})} \quad \text{for all } \xi\in\mathbb{R}.$$

Applying these estimates together with assumptions (6.92) and (6.93) from (6.96) we conclude

$$y^T\tilde{A}(\xi)\overline{y} \ge \theta_0(\xi)\|v\|_2^2 - 2\sqrt{\theta_2(\xi)}\|v\|_2\|w\|_{\ell^2(\mathbb{C})} - \sqrt{\theta_3}\|w\|^2_{\ell^2(\mathbb{C})} + \theta_1\|w\|^2_{\ell^2(\mathbb{C})}$$

for all $\xi\in\mathbb{R}$. Hence, using Young's inequality with an arbitrary constant $\eta(\xi)>0$ yields $2\|v\|_2\|w\|_{\ell^2(\mathbb{C})} \le \eta(\xi)\|v\|_2^2 + \frac{\|w\|^2_{\ell^2(\mathbb{C})}}{\eta(\xi)}$ and thus we obtain

$$y^T\tilde{A}(\xi)\overline{y} \ge (\theta_0(\xi)-\sqrt{\theta_2(\xi)}\eta(\xi))\|v\|_2^2 + \left(\theta_1(\xi)-\sqrt{\theta_3(\xi)}-\frac{\sqrt{\theta_2(\xi)}}{\eta(\xi)}\right)\|w\|^2_{\ell^2(\mathbb{C})}.$$

Now, we fix

$$\eta(\xi) := \frac{\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}}{2\sqrt{\theta_2(\xi)}} + \sqrt{\frac{\left(\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}\right)^2}{4\theta_2(\xi)}+1} > 0$$

implying $\theta_0(\xi)-\sqrt{\theta_2(\xi)}\eta(\xi) = \theta_1(\xi)-\sqrt{\theta_3(\xi)}-\frac{\sqrt{\theta_2(\xi)}}{\eta(\xi)}$. Hence, for all $\xi\in\mathbb{R}$ we obtain

$$\begin{aligned} y^T\tilde{A}(\xi)\overline{y} &\ge \left(\frac{\theta_0(\xi)+\theta_1(\xi)+\sqrt{\theta_3(\xi)}}{2} - \sqrt{\frac{\left(\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}\right)^2}{4}+\theta_2(\xi)}\right) \\ &\qquad \cdot (\|v\|_2^2+\|w\|^2_{\ell^2(\mathbb{C})}) \\ &= \frac{\theta_0(\xi)(\theta_1(\xi)-\sqrt{\theta_3(\xi)})-\theta_2(\xi)}{\frac{\theta_0(\xi)+\theta_1(\xi)-\sqrt{\theta_3(\xi)}}{2}+\sqrt{\frac{\left(\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}\right)^2}{4}+\theta_2(\xi)}}\|y\|^2_{\ell^2(\mathbb{C})}. \end{aligned}$$

Thus, combining this result with (6.90) we conclude

$$\begin{aligned} x^TA(\xi)\overline{x} &\ge \frac{\theta_0(\xi)(\theta_1(\xi)-\sqrt{\theta_3(\xi)})-\theta_2(\xi)}{\frac{\theta_0(\xi)+\theta_1(\xi)-\sqrt{\theta_3(\xi)}}{2}+\sqrt{\frac{\left(\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}\right)^2}{4}+\theta_2(\xi)}}(\|y\|^2_{\ell^2(\mathbb{C})}+\|z\|^2_{\ell^2(\mathbb{C})}) \\ &= \frac{\theta_0(\xi)(\theta_1(\xi)-\sqrt{\theta_3(\xi)})-\theta_2(\xi)}{\frac{\theta_0(\xi)+\theta_1(\xi)-\sqrt{\theta_3(\xi)}}{2}+\sqrt{\frac{\left(\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}\right)^2}{4}+\theta_2(\xi)}}\|x\|^2_{\ell^2(\mathbb{C})} \end{aligned}$$

for all $\xi \in \mathbb{R}$. Finally, the desired estimate (6.88) is satisfied with

$$\kappa := \inf_{\xi \in \mathbb{R}} \frac{\theta_0(\xi)(\theta_1(\xi) - \sqrt{\theta_3(\xi)}) - \theta_2(\xi)}{\frac{\theta_0(\xi)+\theta_1(\xi)-\sqrt{\theta_3(\xi)}}{2} + \sqrt{\frac{\left(\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}\right)^2}{4} + \theta_2(\xi)}} \tag{6.97}$$

(if the infimum is indeed finite).

Remark 6.25. *Recalling the structure of $\tilde{A}(\xi)$ (cf. (6.91)), having a closer look at the right-hand side of (6.97) and considering Remark 6.24 we see that for N sufficiently large the lower bound $\theta_0(\xi)$ for the smallest eigenvalue of $\tilde{A}_0(\xi)$ actually becomes the dominant value in the computation of the desired lower bound κ which one might have expected in advance.*

In view of the definition of $\tilde{A}(\xi)$ (cf. (6.89)) which shows that $\tilde{A}(\xi)$ is symmetric in ξ for the computation of κ it suffices to consider only the non-negative real axis, i.e., we only have to evaluate (or at least to find a lower bound for) the infimum $\inf_{\xi\in[0,\infty)} \theta(\xi)$ where θ is defined by the term on the right-hand side of (6.97). Additionally, this fact shows that it is sufficient to define the functions $\theta_0, \dots, \theta_3$ only on the non-negative interval $[0, \infty)$ as well.

To compute the desired lower bound κ for the infimum $\inf_{\xi\in[0,\infty)} \theta(\xi)$ we first of all fix some finite radius $\xi_0 > 0$.

On the "compact part" $[0, \xi_0]$ we divide the interval into several "small" closed subintervals $\mathcal{I}_1, \dots, \mathcal{I}_M$ where $M \in \mathbb{N}$ denotes the number of intervals and such that $\bigcup_{k=1}^{M} \mathcal{I}_k = [0, \xi_0]$ as well as $\mathcal{I}_k \cap \mathcal{I}_{k+1}$ contains a single "intersection" point for all $k \in \{1, \dots, M-1\}$. Then, on each subinterval $\mathcal{I}_k$ (for some $k \in \{1, \dots, M\}$) we use interval arithmetic computations (cf. Section 3.3) to evaluate $\theta(\mathcal{I}_k)$ to obtain a lower bound m_k for the range of θ on this subinterval $\mathcal{I}_k$, i.e., we define

$$m_k := \min \theta(\mathcal{I}_k) = \min \frac{\theta_0(\mathcal{I}_k) \odot (\theta_1(\mathcal{I}_k) \ominus \sqrt{\theta_3(\mathcal{I}_k)}) \ominus \theta_2(\mathcal{I}_k)}{\frac{\theta_0(\mathcal{I}_k)\oplus\theta_1(\mathcal{I}_k)\ominus\sqrt{\theta_3(\mathcal{I}_k)}}{2} \oplus \sqrt{\frac{\left(\theta_0(\mathcal{I}_k)\ominus\theta_1(\mathcal{I}_k)\oplus\sqrt{\theta_3(\mathcal{I}_k)}\right)^2}{4} \oplus \theta_2(\mathcal{I}_k)}} \tag{6.98}$$

for all $k \in \{1, \dots, M\}$.

On the unbounded part $[\xi_0, \infty)$ we use analytical methods to estimate θ from below by a constant $m_\infty \in \mathbb{R}$, i.e., we compute some $m_\infty \in \mathbb{R}$ such that

$$m_\infty \leq \theta(\xi) = \frac{\theta_0(\xi)(\theta_1(\xi) - \sqrt{\theta_3(\xi)}) - \theta_2(\xi)}{\frac{\theta_0(\xi)+\theta_1(\xi)-\sqrt{\theta_3(\xi)}}{2} + \sqrt{\frac{\left(\theta_0(\xi)-\theta_1(\xi)+\sqrt{\theta_3(\xi)}\right)^2}{4} + \theta_2(\xi)}} \quad \text{for all } |\xi| \geq \xi_0. \tag{6.99}$$

Finally, the procedure above implies that $\kappa := \min\{m_1, \dots, m_M, m_\infty\}$ is a suitable lower bound satisfying (6.82).

Remark 6.26. *As mentioned in Section 3.3 the evaluation of "wide" intervals, i.e., intervals of "large" diameter, results in an overestimation of the errors generated by the interval arithmetic computations. Hence, it is advisable to increase the number of subintervals M which shrinks the diameter of each single interval to obtain relatively "tight"* *lower bounds m_k. However, increasing the number of intervals massively enhances the computational effort since on each of the intervals $\mathcal{I}_k$ interval arithmetic calculations have to be performed. Nevertheless, since the lower bounds $m_1, \dots, m_M$ can be computed independently on each of the subintervals $\mathcal{I}_1, \dots, \mathcal{I}_M$ the procedure can easily be parallelized which finally reduces the computational effort again.*

To the end of this Section, we consider each of our assumptions (6.92) to (6.95) separately and provide a strategy how to obtain the desired functions $\theta_0, \dots, \theta_3$.

Computation of $\theta_0, \dots, \theta_3$

To compute the desired functions $\theta_0, \dots, \theta_3$ satisfying (6.92) to (6.95) we consider all assumptions separately in detail:

1. In view of (6.98) and (6.99) we see that it suffices to assume condition (6.92) piecewise on each of the subintervals $\mathcal{I}_1, \dots, \mathcal{I}_M$ and $[\xi_0, \infty)$ respectively, i.e., we can define θ_0 as a step function using constants $\theta_{0,1}, \dots, \theta_{0,M}, \theta_{0,\infty} \in \mathbb{R}$ (each corresponding to a single subinterval) such that

$$\lambda_{\min}(\tilde{A}_0(\xi)) \geq \theta_{0,k} \quad \text{for all } \xi \in \mathcal{I}_k,\ k \in \{1, \dots, M, \infty\}\,, \tag{6.100}$$

   where we set $\mathcal{I}_\infty := [\xi_0, \infty)$. Then, the definition of the lower bound m_k on a subinterval $\mathcal{I}_k$ for some $k \in \{1, \dots, M\}$ reads as

$$m_k := \min \frac{\theta_{0,k} \odot (\theta_1(\mathcal{I}_k) \ominus \sqrt{\theta_3(\mathcal{I}_k)}) \ominus \theta_2(\mathcal{I}_k)}{\frac{\theta_{0,k} \oplus \theta_1(\mathcal{I}_k) \ominus \sqrt{\theta_3(\mathcal{I}_k)}}{2} \oplus \sqrt{\frac{\left(\theta_{0,k} \ominus \theta_1(\mathcal{I}_k) \oplus \sqrt{\theta_3(\mathcal{I}_k)}\right)^2}{4} \oplus \theta_2(\mathcal{I}_k)}}$$

   and for the unbounded interval $\mathcal{I}_\infty$ we calculate m_∞ such that

$$m_\infty \leq \frac{\theta_{0,\infty}(\theta_1(\xi) - \sqrt{\theta_3(\xi)}) - \theta_2(\xi)}{\frac{\theta_{0,\infty} + \theta_1(\xi) - \sqrt{\theta_3(\xi)}}{2} + \sqrt{\frac{\left(\theta_{0,\infty} - \theta_1(\xi) + \sqrt{\theta_3(\xi)}\right)^2}{4} + \theta_2(\xi)}} \quad \text{for all } \xi \in \mathcal{I}_\infty. \tag{6.101}$$

   Thus, we are left with the computation of the constants $\theta_{0,1}, \dots, \theta_{0,M}$ and $\theta_{0,\infty}$ such that (6.100) holds true on each corresponding subinterval. Therefore, on each compact subinterval $\mathcal{I}_k$ (for some $k \in \{1, \dots, M\}$) we set up the interval matrix $\tilde{A}(\mathcal{I}_k)$ using the formulas in (6.89) and the definition of $\tilde{A}(\xi)$ (cf. (6.91)) together with interval arithmetic evaluations. Then, we apply the eigenvalue methods presented in Section 9.5.1 to enclose the eigenvalues of the interval matrix $\tilde{A}(\mathcal{I}_k)$ which finally yields the desired lower bound $\theta_{0,k}$ satisfying (6.100) for all $k \in \{1, \dots, M\}$. Again, we note that we expect the radii of the enclosing intervals to be "small" if the intervals $\mathcal{I}_1, \dots, \mathcal{I}_M$ are "small".

   Next, we have to treat the unbounded interval for which we cannot use interval arithmetic evaluations. Therefore, we use analytical methods which provide a lower bound $\theta_{0,\infty}$ satisfying $\lambda_{\min}(\tilde{A}_0(\xi)) \geq \theta_{0,\infty}$ for all $\xi \in \mathcal{I}_\infty = [\xi_0, \infty)$. Here, we use Gershgorin's Circle Theorem (cf. [84, Theorem 12.9]) which implies that the eigenvalues of $\tilde{A}_0(\xi)$ are contained in the union $\bigcup_{k=1}^{N} \mathcal{G}_k(\xi)$ of the Gershgorin circles

$$\mathcal{G}_k(\xi) := \left\{ z \in \mathbb{C}\colon |z - (\tilde{A}(\xi))_{k,k}| \leq \sum_{\substack{l=1\\ l\neq k}}^{N} |(\tilde{A}_0(\xi))_{k,l}| \right\} \quad \text{for all } \xi \in [\xi_0, \infty). \tag{6.102}$$

Moreover, since the matrix $\tilde{A}_0(\xi)$ is symmetric (and thus has real eigenvalues) we finally need to compute a lower bound for

$$\min_{k=1,\dots,N} \inf_{\xi\in[\xi_0,\infty)} \left((\tilde{A}(\xi))_{k,k} - \sum_{\substack{l=1\\ l\neq k}}^{N} |(\tilde{A}_0(\xi))_{k,l}| \right).$$

Then, this bound indeed is a lower bound for the eigenvalues of all matrices $\tilde{A}(\xi)$ for all $\xi \in [\xi_0, \infty)$ and thus, especially the desired lower bound on $\mathcal{I}_\infty$ satisfying (6.100).

To compute the desired lower bound, we first of all use the definition of $\tilde{A}$ (cf. (6.89)) to obtain the representation

$$(\tilde{A}(\xi))_{k,l} = \begin{cases} \frac{\xi^2-\sigma+k^2\pi^2}{\xi^2+\sigma+k^2\pi^2}, & k = l, \\ \frac{8klRe(1-(-1)^{k+l})}{\pi^2(k^2-l^2)^2\sqrt{\xi^2+\sigma+k^2\pi^2}\sqrt{\xi^2+\sigma+l^2\pi^2}}, & k \neq l \end{cases} \tag{6.103}$$

for all $k, l \in \mathbb{N}$ and $\xi \in \mathbb{R}$. Using this representation of $\tilde{A}$ together with the splitting given in (6.91) we calculate

$$\begin{aligned} |(\tilde{A}_0(\xi))_{k,l}| &= \frac{8klRe(1-(-1)^{k+l})}{\pi^2(k^2-l^2)^2\sqrt{\xi^2+\sigma+k^2\pi^2}\sqrt{\xi^2+\sigma+l^2\pi^2}} \\ &\leq \frac{8klRe(1-(-1)^{k+l})}{\pi^2(k^2-l^2)^2\sqrt{\xi_0^2+\sigma+k^2\pi^2}\sqrt{\xi_0^2+\sigma+l^2\pi^2}} \\ &= \frac{8Re(1-(-1)^{k+l})}{\pi^2(k^2-l^2)^2\sqrt{\pi^2+\frac{\xi_0^2+\sigma}{k^2}}\sqrt{\pi^2+\frac{\xi_0^2+\sigma}{l^2}}} \end{aligned}$$

for all $k, l \in \{1, \dots, N\}$ with $k \neq l$ and $\xi \in [\xi_0, \infty)$ which directly implies

$$\sum_{\substack{l=1\\ l\neq k}}^{N} |(\tilde{A}_0(\xi))_{k,l}| \leq \frac{8Re}{\pi^2} \sum_{\substack{l=1\\ l\neq k}}^{N} \frac{1-(-1)^{k+l}}{(k^2-l^2)^2\sqrt{\pi^2+\frac{\xi_0^2+\sigma}{k^2}}\sqrt{\pi^2+\frac{\xi_0^2+\sigma}{l^2}}}$$

for all $\xi \in [\xi_0, \infty)$ and $k \in \{1, \dots, N\}$. Furthermore, for all $k \in \{1, \dots, N\}$ we obtain

$$(\tilde{A}(\xi))_{k,k} = \frac{\xi^2-\sigma+k^2\pi^2}{\xi^2+\sigma+k^2\pi^2} \geq \frac{\xi_0^2-\sigma+k^2\pi^2}{\xi_0^2+\sigma+k^2\pi^2} \quad \text{for all } \xi \in [\xi_0, \infty).$$

Combining the previous results shows that the desired lower bound $\theta_{0,\infty}$ satisfying (6.100) can be defined as follows

$$\theta_{0,\infty} := \min_{k=1,\dots,N} \left(\frac{\xi_0^2-\sigma+k^2\pi^2}{\xi_0^2+\sigma+k^2\pi^2} - \frac{8Re}{\pi^2} \sum_{\substack{l=1\\ l\neq k}}^{N} \frac{1-(-1)^{k+l}}{(k^2-l^2)^2\sqrt{\pi^2+\frac{\xi_0^2+\sigma}{k^2}}\sqrt{\pi^2+\frac{\xi_0^2+\sigma}{l^2}}} \right).$$

We note that the terms on the right-hand side can easily be evaluated rigorously using interval arithmetic operations.

2. Since $\tilde{D}$ is diagonal we calculate

$$\lambda_{\min}(\tilde{D}(\xi)) = \inf_{k\in\mathbb{N}}(\tilde{D}(\xi))_{k,k} = \inf_{k\in\mathbb{N}} \frac{\xi^2 - \sigma + (N+k)^2\pi^2}{\xi^2 + \sigma + (N+k)^2\pi^2} = \frac{\xi^2 - \sigma + (N+1)^2\pi^2}{\xi^2 + \sigma + (N+1)^2\pi^2}$$

and

$$\inf_{\xi\in[\xi_0,\infty)} \lambda_{\min}(\tilde{D}(\xi)) = \inf_{\xi\in[\xi_0,\infty)} \frac{\xi^2 - \sigma + (N+1)^2\pi^2}{\xi^2 + \sigma + (N+1)^2\pi^2} = \frac{\xi_0^2 - \sigma + (N+1)^2\pi^2}{\xi_0^2 + \sigma + (N+1)^2\pi^2}.$$

Thus, we can define

$$\theta_1\colon \mathbb{R}\to\mathbb{R},\ \theta_1(\xi) := \begin{cases} \frac{\xi^2-\sigma+(N+1)^2\pi^2}{\xi^2+\sigma+(N+1)^2\pi^2}, & \xi\in[0,\xi_0), \\ \frac{\xi_0^2-\sigma+(N+1)^2\pi^2}{\xi_0^2+\sigma+(N+1)^2\pi^2}, & \xi\in[\xi_0,\infty) \end{cases}$$

which satisfies (6.93).

3. Using the identity

$$(1-(-1)^{k+l})^2 = 2(1-(-1)^{k+l}) \quad \text{for all } k,l\in\mathbb{N}$$

together with the representation formula for $\tilde{A}(\xi)$ (cf. (6.103)) we calculate

$$\begin{aligned} |(\tilde{A}(\xi))_{k,l}|^2 &= \frac{128k^2l^2Re^2(1-(-1)^{k+l})}{\pi^4(k^2-l^2)^4(\xi^2+\sigma+k^2\pi^2)(\xi^2+\sigma+l^2\pi^2)} \\ &\leq \frac{128k^2l^2Re^2(1-(-1)^{k+l})}{\pi^4(k^2-l^2)^4(\xi^2+\sigma+k^2\pi^2)(\xi^2+\sigma+(N+1)^2\pi^2)} \end{aligned} \tag{6.104}$$

for all $k,l\in\mathbb{N}$ with $k\neq l$, $l\geq N+1$ and $\xi\in[0,\infty)$ which directly implies

$$\begin{aligned} &\sum_{k=1}^{N}\sum_{l=N+1}^{\infty} |(\tilde{A}(\xi))_{k,l}|^2 \\ &\qquad \leq \frac{128Re^2}{\pi^4(\xi^2+\sigma+(N+1)^2\pi^2)} \sum_{k=1}^{N} \frac{k^2}{\xi^2+\sigma+k^2\pi^2} \sum_{l=N+1}^{\infty} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4}. \end{aligned}$$

Applying Lemma A.14 (v) yields

$$\begin{aligned} \sum_{l=N+1}^{\infty} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} &= \sum_{\substack{l=1\\ l\neq k}}^{\infty} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} - \sum_{\substack{l=1\\ l\neq k}}^{N} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} \\ &= \frac{\pi^4}{384k^2} - \frac{\pi^2}{64k^4} - \sum_{\substack{l=1\\ l\neq k}}^{N} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} \end{aligned} \tag{6.105}$$

which together with the previous estimate results in the definition

$$\begin{aligned} \sum_{k=1}^{N}\sum_{l=N+1}^{\infty} |(\tilde{A}(\xi))_{k,l}|^2 \leq \frac{Re^2}{\xi^2+\sigma+(N+1)^2\pi^2} \Bigg(&\sum_{k=1}^{N} \frac{\frac{1}{3}-\frac{2}{k^2\pi^2}}{\xi^2+\sigma+k^2\pi^2} \\ &- \sum_{k=1}^{N} \frac{128}{\pi^4\left(\pi^2+\frac{\xi^2+\sigma}{k^2}\right)} \sum_{\substack{l=1\\ l\neq k}}^{N} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} \Bigg) =: \theta_2(\xi) \end{aligned}$$

for all $\xi \in [0, \xi_0]$.

Based on (6.104) we additionally estimate

$$|(\tilde{A}(\xi))_{k,l}|^2 \leq \frac{128k^2l^2Re^2(1-(-1)^{k+l})}{\pi^4(k^2-l^2)^4(\xi_0^2+\sigma+k^2\pi^2)(\xi_0^2+\sigma+(N+1)^2\pi^2)}$$

for all $k, l \in \mathbb{N}$ with $k \neq l$, $l \geq N+1$ and $\xi \in [\xi_0, \infty)$.

Then, the same arguments as before yield

$$\sum_{k=1}^{N}\sum_{l=N+1}^{\infty}|(\tilde{A}(\xi))_{k,l}|^2 \leq \frac{Re^2}{\xi_0^2+\sigma+(N+1)^2\pi^2}\left(\sum_{k=1}^{N}\frac{\frac{1}{3}-\frac{2}{k^2\pi^2}}{\xi_0^2+\sigma+k^2\pi^2}\right.$$
$$\left.-\sum_{k=1}^{N}\frac{128}{\pi^4\left(\pi^2+\frac{\xi_0^2+\sigma}{k^2}\right)}\sum_{\substack{l=1\\l\neq k}}^{N}\frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4}\right) =: \theta_2(\xi)$$

for all $\xi \in [\xi_0, \infty)$. Hence, (6.94) holds true with this definition of θ_2.

4. For all $k \in \mathbb{N}$ we have the identity

$$\frac{k^2}{\xi^2+\sigma+k^2\pi^2} = \frac{1}{\pi^2}\left(1-\frac{\xi^2+\sigma}{\xi^2+\sigma+k^2\pi^2}\right) \leq \frac{1}{\pi^2} \quad \text{for all } \xi \in [0,\infty).$$

Then, using (6.104) we calculate

$$|(\tilde{A}(\xi))_{k,l}|^2 \leq \frac{128l^2Re^2(1-(-1)^{k+l})}{\pi^6(k^2-l^2)^4(\xi^2+\sigma+(N+1)^2\pi^2)} \tag{6.106}$$

for all $k, l \in \mathbb{N}$ with $k \neq l$, $l \geq N+1$ and $\xi \in [0, \infty)$ which implies

$$\sum_{k=N+1}^{\infty}\sum_{\substack{l=N+1\\l\neq k}}^{\infty}|(\tilde{A}(\xi))_{k,l}|^2 \leq \frac{128Re^2}{\pi^6(\xi^2+\sigma+(N+1)^2\pi^2)}\sum_{k=N+1}^{\infty}\sum_{\substack{l=N+1\\l\neq k}}^{\infty}\frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4}$$

for all $\xi \in [0, \infty)$. Similar as in part 3. (cf. (6.105)) applying Lemma A.14 (v) we obtain

$$\sum_{k=N+1}^{\infty}\sum_{\substack{l=N+1\\l\neq k}}^{\infty}|(\tilde{A}(\xi))_{k,l}|^2 \leq \frac{Re^2}{\xi^2+\sigma+(N+1)^2\pi^2}\left(\frac{1}{3\pi^2}\sum_{k=N+1}^{\infty}\frac{1}{k^2}-\frac{2}{\pi^4}\sum_{k=N+1}^{\infty}\frac{1}{k^4}\right.$$
$$\left.-\frac{128}{\pi^6}\sum_{l=1}^{N}l^2\sum_{k=N+1}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^4}\right).$$

Using the well-known identities $\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}$ and $\sum_{k=1}^{\infty}\frac{1}{k^4} = \frac{\pi^4}{90}$ (cf. [36, Section 0.233]) we calculate

$$\sum_{k=N+1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6} - \sum_{k=1}^{N}\frac{1}{k^2} \quad \text{and} \quad \sum_{k=N+1}^{\infty}\frac{1}{k^4} = \frac{\pi^4}{90} - \sum_{k=1}^{N}\frac{1}{k^4}.$$

Moreover, Lemma A.14 (iv) implies

$$\begin{aligned}
&\frac{128}{\pi^6}\sum_{l=1}^{N} l^2 \sum_{k=N+1}^{\infty} \frac{1-(-1)^{k+l}}{(k^2-l^2)^4}\\
&\quad = \sum_{l=1}^{N}\left(\frac{1}{3l^2\pi^2}+\frac{10}{l^4\pi^4}-\frac{64(1-(-1)^l)}{l^6\pi^6}\right)-\frac{128}{\pi^6}\sum_{l=1}^{N} l^2 \sum_{\substack{k=1\\k\neq l}}^{N}\frac{1-(-1)^{k+l}}{(k^2-l^2)^4}.
\end{aligned}$$

Hence, together with the estimate above we obtain the definition

$$\begin{aligned}
\sum_{k=N+1}^{\infty}\sum_{\substack{l=N+1\\l\neq k}}^{\infty}|(\tilde{A}(\xi))_{k,l}|^2 &\leq \frac{Re^2}{\xi^2+\sigma+(N+1)^2\pi^2}\left(\frac{1}{30}-\frac{2}{3\pi^2}\sum_{k=1}^{N}\frac{1}{k^2}-\frac{8}{\pi^4}\sum_{k=1}^{N}\frac{1}{k^4}\right.\\
&\qquad\left.+\frac{64}{\pi^6}\sum_{k=1}^{N}\frac{1-(-1)^k}{k^6}+\frac{128}{\pi^6}\sum_{l=1}^{N} l^2\sum_{\substack{k=1\\k\neq l}}^{N}\frac{1-(-1)^{k+l}}{(k^2-l^2)^4}\right)\\
&=: \theta_3(\xi) \quad \text{for all } \xi\in[0,\xi_0].
\end{aligned}$$

Similar as in part 3. using (6.106) we calculate

$$|(\tilde{A}(\xi))_{k,l}|^2 \leq \frac{128l^2Re^2(1-(-1)^{k+l})}{\pi^6(k^2-l^2)^4(\xi_0^2+\sigma+(N+1)^2\pi^2)}$$

for all $k,l\in\mathbb{N}$ with $l\geq N+1$ and $\xi\in[\xi_0,\infty)$ which together with the previous calculations justifies the definition

$$\begin{aligned}
\theta_3(\xi) &:= \frac{Re^2}{\xi_0^2+\sigma+(N+1)^2\pi^2}\left(\frac{1}{30}-\frac{2}{3\pi^2}\sum_{k=1}^{N}\frac{1}{k^2}-\frac{8}{\pi^4}\sum_{k=1}^{N}\frac{1}{k^4}\right.\\
&\qquad\left.+\frac{64}{\pi^6}\sum_{k=1}^{N}\frac{1-(-1)^k}{k^6}+\frac{128}{\pi^6}\sum_{l=1}^{N} l^2\sum_{\substack{k=1\\k\neq l}}^{N}\frac{1-(-1)^{k+l}}{(k^2-l^2)^4}\right)
\end{aligned}$$

for all $\xi\in[\xi_0,\infty)$.

Remark 6.27. (i) *We note that on the unbounded interval $\mathcal{I}_\infty=[\xi_0,\infty)$ the functions $\theta_1,\dots,\theta_3$ are constant and thus, independent of ξ. Hence, the desired lower bound m_∞ on $\mathcal{I}_\infty$ can be computed by rigorously evaluating the expressions $\theta_{0,\infty},\theta_1(\xi_0)$, $\theta_2(\xi_0)$ and $\theta_3(\xi_0)$ respectively in a single point (and not on an interval) using interval arithmetic calculations (with point intervals as input). Finally, with these data in hand we can compute m_∞ via the term* (6.101) *again using interval arithmetic evaluations.*

(ii) *The choice of $\xi_0>0$ separating the two different strategies for obtaining the desired lower bounds is somehow arbitrary. Nevertheless, in view of Remark 6.26 concerning the parallelization it makes sense to chose ξ as a power of two to be able to distribute the interval $[0,\xi_0]$ equally over all processes. Moreover, concerning the enclosure of eigenvalues via Gershgorin circles (cf.* (6.102)*) it is useful to fix ξ not "too close to zero" which results in "smaller" Gershgorin circles since the off-diagonal elements of $\tilde{A}_0(\xi)$ for $\xi\in[\xi_0,\infty)$ become "smaller".*

(iii) *Due to*

$$\langle \Phi^{-1}\,\hat{\mathrm{L}}_U\, u, \Phi^{-1}\,\hat{\mathrm{L}}_U\, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq \langle u, u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} + \int_\Omega u^T G_U u \,\mathrm{d}(x,y) \quad \textit{for all } u \in H(\Omega)$$

the desired lower bound $\hat{\kappa}$ coincides with the constant κ, i.e., we do not need an additional computation for $\hat{\kappa}$.

6.2.2 Bound for the Essential Spectrum

As mentioned above in this Section we will have a closer look at the computation of a lower bound for the essential spectrum of the eigenvalue problems (6.8) and (6.9) respectively. Again, we note that the essential spectrum of each of these problems is defined via the associated self-adjoint operator $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{L}_{U+\omega}$ and $\Phi^{-1}\,\mathrm{L}_{U+\omega}(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*$ respectively (cf. [74, Section 10.2.1] at the beginning of Section 6.2.1).

By Poincaré's min-max-principle the lower bound for the essential spectra is increasing with respect to the homotopy parameter (cf. [74, Section 10.2.4; p. 392]). Hence, the lower bound for the essential spectrum of the associated base problem is a lower bound for the "original" eigenvalue problem as well, i.e., $\sigma_0^{(0)} = \gamma_1$ and $\hat{\sigma}_0^{(0)} = \hat{\gamma}_1$ respectively can be used as the desired lower bounds for the essential spectrum of the eigenvalue problems (6.8) and (6.9) respectively (cf. Section 6.2.1.3).

Nevertheless, to the end of this Section we present a strategy (independent of the homotopy method) to compute the desired lower bounds for the essential spectra if the lower bounds κ and $\hat{\kappa}$ introduced in Section 6.2.1.4 (cf. (6.82) and (6.83) respectively) are in hand. Thus, in the further course we assume that such constants κ and $\hat{\kappa}$ are computed explicitly. For a strategy to obtain such lower bounds we refer the reader to Section 6.2.1.4.

Next, to obtain the desired lower bounds for the essential spectra we first prove the following compact perturbation result

Proposition 6.28. *The following assertions hold true:*

(i) $(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{L}_{U+\omega}$ *is a relative compact perturbation of* $(\Phi^{-1}\,\mathrm{L}_U)^*\Phi^{-1}\,\mathrm{L}_U$ *and*

$$\sigma_{ess}((\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{L}_{U+\omega}) = \sigma_{ess}((\Phi^{-1}\,\mathrm{L}_U)^*\Phi^{-1}\,\mathrm{L}_U).$$

(ii) $\Phi^{-1}\,\mathrm{L}_{U+\omega}(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*$ *is a relative compact perturbation of* $\Phi^{-1}\,\mathrm{L}_U(\Phi^{-1}\,\mathrm{L}_U)^*$ *and*

$$\sigma_{ess}(\Phi^{-1}\,\mathrm{L}_{U+\omega}(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*) = \sigma_{ess}(\Phi^{-1}\,\mathrm{L}_U(\Phi^{-1}\,\mathrm{L}_U)^*).$$

Proof. (i) To improve readability of the proof, we introduce the abbreviations

$$S := (\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{L}_{U+\omega} \quad \text{and} \quad S_0 := (\Phi^{-1}\,\mathrm{L}_U)^*\Phi^{-1}\,\mathrm{L}_U\,.$$

Thus, we need to show, that S is a relative compact perturbation of S_0. Since S_0 is linear and bounded it is closed. Thus, we need to show that $S - S_0$ is compact.

Now, let $u \in H(\Omega)$. Using the definition of $\hat{\mathrm{L}}_{U+\omega}$ (see (6.3)) and the equality (6.4) (which also holds true for U instead of $U+\omega$), we obtain

$$\begin{aligned}
&(S-S_0)u \\
&\quad = (\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{L}_{U+\omega}\,u - (\Phi^{-1}\,\mathrm{L}_U)^*\Phi^{-1}\,\mathrm{L}_U\,u \\
&\quad = (\Phi^{-1}\,\mathrm{L}_{U+\omega})^*[\Phi^{-1}\,\mathrm{L}_{U+\omega}\,u - \Phi^{-1}\,\mathrm{L}_U\,u] + [(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* - (\Phi^{-1}\,\mathrm{L}_U)^*]\Phi^{-1}\,\mathrm{L}_U\,u \\
&\quad = (\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}[\mathrm{L}_{U+\omega} - \mathrm{L}_U]u + \Phi^{-1}[\hat{\mathrm{L}}_{U+\omega} - \hat{\mathrm{L}}_U]\Phi^{-1}\,\mathrm{L}_U\,u.
\end{aligned}$$

Moreover, Proposition 3.1 (i) and (ii) together with the definition of L (see (3.10)) imply

$$\mathrm{L}_{U+\omega}\,u - \mathrm{L}_U\,u = -\Delta u + \mathrm{B}_{U+\omega}\,u - (-\Delta u + \mathrm{B}_U\,u) = \mathrm{B}_{U+\omega}\,u - \mathrm{B}_U\,u = \mathrm{B}_\omega\,u.$$

Equivalently, applying Proposition 6.2 (i) and (ii) as well as the definition of $\hat{\mathrm{L}}$ (see (6.3)) shows

$$\hat{\mathrm{L}}_{U+\omega}\,u - \hat{\mathrm{L}}_U\,u = -\Delta u + \hat{\mathrm{B}}_{U+\omega}\,u - (-\Delta u + \hat{\mathrm{B}}_U\,u) = \hat{\mathrm{B}}_{U+\omega}\,u - \hat{\mathrm{B}}_U\,u = \hat{\mathrm{B}}_\omega\,u.$$

Putting everything together, we obtain

$$(S-S_0)u = (\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{B}_\omega\,u + \Phi^{-1}\,\hat{\mathrm{B}}_\omega\,\Phi^{-1}\,\mathrm{L}_U\,u.$$

Additionally, we introduce the abbreviation

$$C := Re\max\left\{2\|(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\|_{\mathcal{B}}\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)},\ \|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)} + C_2\|\nabla\omega\|_{L^\infty(\Omega_R,\mathbb{R}^{2\times 2})}\right\},$$

where Ω_R is defined as in the beginning of Section 6.2.1.2 and $\|\cdot\|_{\mathcal{B}}$ denotes the corresponding operator norm.

Now, let $(u^{(n)})_{n\in\mathbb{N}}$ be a bounded sequence in $H(\Omega)$. Hence, $(u^{(n)})_{n\in\mathbb{N}}$ is bounded in $H_0^1(\Omega,\mathbb{R}^2)$ and also in $H^1(\Omega_R,\mathbb{R}^2)$. Since Ω_R is bounded, Sobolev-Kondrachev-Rellich's Embedding Theorem (cf. [26, Theorem 1; p. 272]) implies the existence of a subsequence $(u^{(n_k)})_{k\in\mathbb{N}}$ converging in $L^2(\Omega_R,\mathbb{R}^2)$. Thus, $(u^{(n_k)})_{k\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega_R,\mathbb{R}^2)$, i.e.,

$$\|u^{(n_k)} - u^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)} \to 0 \quad \text{as } k,l\to\infty.$$

Next, we define $v^{(n_k)} := \Phi^{-1}\,\mathrm{L}_U\,u^{(n_k)}$ for all $k\in\mathbb{N}$. Since $\Phi^{-1}\,\mathrm{L}_U$ is a bounded operator, $(v^{(n_k)})_{k\in\mathbb{N}}$ is a bounded sequence in $H^1(\Omega_R,\mathbb{R}^2)$. Thus, applying Sobolev-Kondrachev-Rellich's Embedding Theorem again yields the existence of a subsequence (again denoted by $(v^{(n_k)})_{k\in\mathbb{N}}$) such that $(v^{(n_k)})_{k\in\mathbb{N}}$ converges in $L^2(\Omega_R,\mathbb{R}^2)$. The same arguments as above yield

$$\|v^{(n_k)} - v^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)} \to 0 \quad \text{as } k,l\to\infty.$$

Thus, using the previous representation of $S-S_0$ together with the estimates provided by Lemma A.11 (i) and (ii) as well as the fact that $\operatorname{supp}(\omega)\subseteq\Omega_R$, we calculate

$$\begin{aligned}
&\|(S-S_0)(u^{(n_k)} - u^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad = \|(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\Phi^{-1}\,\mathrm{B}_\omega(u^{(n_k)} - u^{(n_l)}) + \Phi^{-1}\,\hat{\mathrm{B}}_\omega(v^{(n_k)} - v^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad \le \|(\Phi^{-1}\,\mathrm{L}_{U+\omega})^*\|_{\mathcal{B}}\|\Phi^{-1}\,\mathrm{B}_\omega(u^{(n_k)} - u^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\qquad\qquad + \|\Phi^{-1}\,\hat{\mathrm{B}}_\omega(v^{(n_k)} - v^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad \le C\left(\|u^{(n_k)} - u^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)} + \|v^{(n_k)} - v^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)}\right) \to 0 \quad \text{as } k,l\to\infty.
\end{aligned}$$

Hence, $((S - S_0)u^{(n_k)})_{k\in\mathbb{N}}$ is a Cauchy sequence in $H(\Omega)$. Since $H(\Omega)$ is a Hilbert space $((S - S_0)u^{(n_k)})_{k\in\mathbb{N}}$ converges in $H(\Omega)$ which directly implies the compactness of $S - S_0$.

Finally, the identity $\sigma_{\mathrm{ess}}(S) = \sigma_{\mathrm{ess}}(S_0)$ is a direct consequence of [50, Theorem 5.35] together with the fact that S is a relative compact perturbation of S_0.

(ii) To prove the second assertion, we use abbreviations again:

$$\hat{S} := \Phi^{-1}\,\mathrm{L}_{U+\omega}(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* \quad \text{and} \quad \hat{S}_0 := \Phi^{-1}\,\mathrm{L}_U(\Phi^{-1}\,\mathrm{L}_U)^*.$$

Similarly as in part (i), the definition of $\hat{\mathrm{L}}_{U+\omega}$ (see (6.3)) and the equality (6.4) (which also holds true for U instead of $U + \omega$), as well as Proposition 3.1 (i) and (ii) together with the definition of L (see (3.10)) and additionally Proposition 6.2 (i) and (ii) as well as the definition of $\hat{\mathrm{L}}$ (see (6.3)) yield

$$\begin{aligned}
&(\hat{S} - \hat{S}_0)u \\
&\quad = \Phi^{-1}\,\mathrm{L}_{U+\omega}(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* u - \Phi^{-1}\,\mathrm{L}_U(\Phi^{-1}\,\mathrm{L}_U)^* u \\
&\quad = \Phi^{-1}\,\mathrm{L}_{U+\omega}[(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* u - (\Phi^{-1}\,\mathrm{L}_U)^* u] + [\Phi^{-1}\,\mathrm{L}_{U+\omega} - \Phi^{-1}\,\mathrm{L}_U](\Phi^{-1}\,\mathrm{L}_U)^* u \\
&\quad = \Phi^{-1}\,\mathrm{L}_{U+\omega}\,\Phi^{-1}[\hat{\mathrm{L}}_{U+\omega} - \hat{\mathrm{L}}_U]u + \Phi^{-1}[\mathrm{L}_{U+\omega} - \mathrm{L}_U](\Phi^{-1}\,\mathrm{L}_U)^* u \\
&\quad = \Phi^{-1}\,\mathrm{L}_{U+\omega}\,\Phi^{-1}\,\hat{\mathrm{B}}_\omega\, u + \Phi^{-1}\,\mathrm{B}_\omega(\Phi^{-1}\,\mathrm{L}_U)^* u \quad \text{for all } u \in H(\Omega).
\end{aligned}$$

Now, let $(u^{(n)})_{n\in\mathbb{N}}$ be a bounded sequence in $H(\Omega)$. Again, Sobolev-Kondrachev-Rellich's Embedding Theorem implies the existence of a subsequence $(u^{(n_k)})_{k\in\mathbb{N}}$ converging in $L^2(\Omega_R, \mathbb{R}^2)$. Hence, $(u^{(n_k)})_{k\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega_R, \mathbb{R}^2)$ and thus, we obtain

$$\|u^{(n_k)} - u^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)} \to 0 \quad \text{as } k, l \to \infty.$$

Similarly as in the first part we set $w^{(n_k)} := (\Phi^{-1}\,\mathrm{L}_{U)})^* u^{(n_k)}$ for all $k \in \mathbb{N}$. Since the operator $(\Phi^{-1}\,\mathrm{L}_U)^* = \Phi^{-1}\,\hat{\mathrm{L}}$ is bounded again, $(w^{(n_k)})_{k\in\mathbb{N}}$ is a bounded sequence in $H^1(\Omega_R, \mathbb{R}^2)$ and thus, applying Sobolev-Kondrachev-Rellich's Embedding Theorem again yields the existence of a subsequence (again denoted by $(w^{(n_k)})_{k\in\mathbb{N}}$) such that $(w^{(n_k)})_{k\in\mathbb{N}}$ converges in $L^2(\Omega_R, \mathbb{R}^2)$. With the same arguments as above we get

$$\|w^{(n_k)} - w^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)} \to 0 \quad \text{as } k, l \to \infty.$$

Putting everything together and applying Lemma A.11 (i) and (ii) (again we note that $\mathrm{supp}(\omega) \subseteq \Omega_R$) as well as using the abbreviation

$$\hat{C} := Re \max\left\{\|\Phi^{-1}\,\mathrm{L}_{U+\omega}\|_{\mathcal{B}}\left(\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)} + C_2\|\nabla\omega\|_{L^\infty(\Omega_R,\mathbb{R}^{2\times 2})}\right), 2\|\omega\|_{L^\infty(\Omega_R,\mathbb{R}^2)}\right\}$$

instead of C the same arguments as in part (i) yield

$$\begin{aligned}
&\|(\hat{S} - \hat{S}_0)(u^{(n_k)} - u^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad = |||\Phi^{-1}\,\mathrm{L}_{U+\omega}\,\Phi^{-1}\,\hat{\mathrm{B}}_\omega(u^{(n_k)} - u^{(n_l)}) + \Phi^{-1}\,\mathrm{B}_\omega(w^{(n_k)} - w^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad \le \|\Phi^{-1}\,\mathrm{L}_{U+\omega}\|_{\mathcal{B}}\|\Phi^{-1}\,\hat{\mathrm{B}}_\omega(u^{(n_k)} - u^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\qquad\quad + \|\Phi^{-1}\,\mathrm{B}_\omega(w^{(n_k)} - w^{(n_l)})\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\quad \le \hat{C}\left(\|u^{(n_k)} - u^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)} + \|v^{(n_k)} - v^{(n_l)}\|_{L^2(\Omega_R,\mathbb{R}^2)}\right) \to 0 \quad \text{as } k, l \to \infty.
\end{aligned}$$

Hence, $((\hat{S}-\hat{S}_0)u^{(n_k)})_{k\in\mathbb{N}}$ is a Cauchy sequence in $H(\Omega)$ and thus $((\hat{S}-\hat{S}_0)u^{(n_k)})_{k\in\mathbb{N}}$ converges in $H(\Omega)$ implying the compactness of $\hat{S}-\hat{S}_0$ proving the first part of the assertion.

Again, [50, Theorem 5.35] proves the identity $\sigma_{\text{ess}}(\hat{S}) = \sigma_{\text{ess}}(\hat{S}_0)$ and thus the assertion.

□

Applying the compact perturbation results provided by Proposition 6.28 we end up with the computation of lower bounds for the essential spectra of the two eigenvalue problems

$$\langle \Phi^{-1}\,\mathrm{L}_U\,u, \Phi^{-1}\,\mathrm{L}_U\,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq \sigma_0\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u\in H(\Omega)$$

and

$$\langle (\Phi^{-1}\,\mathrm{L}_U)^*u, (\Phi^{-1}\,\mathrm{L}_U)^*u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq \hat{\sigma}_0\langle u,u\rangle_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u\in H(\Omega)$$

respectively.

Since our domain still contains the obstacle D we cannot apply well-known techniques (e.g. Fourier transform methods, etc.) to obtain the desired lower bounds for the essential spectrum directly. However, we can use the lower bounds κ and $\hat{\kappa}$ satisfying (6.82) and (6.83) as lower bounds, i.e., we set $\sigma_0 := \kappa$ and $\hat{\sigma}_0 := \hat{\kappa}$. We note that these lower bounds are probably worse than the actual infima of the essential spectra, however, they are computable (note that we assumed that we have such constants κ and $\hat{\kappa}$ in hand).

7 Reconstruction of the Pressure

Since a complete solution of the transformed Navier-Stokes equations (1.13) also contains the pressure (which is not provided by our approach using the divergence-free subspace $H(\Omega)$ and problem (1.15)), in this Chapter we present a strategy to "reconstruct" the pressure a posteriori if we already proved the existence of a velocity field satisfying our weak formulation (1.15). Therefore, we suppose that all assumptions in Theorem 3.4 (especially inequality (3.11)) are satisfied. In particular there exists an exact solution $u^* \in H(\Omega)$ of our Navier-Stokes equation (1.15) on $H(\Omega)$. Moreover, Theorem 3.4 provides the enclosure

$$\|u^* - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \frac{2K\delta}{1 + \sqrt{1 - 4K^2{C_4}^2 Re\,\delta}} =: \alpha,$$

where $\tilde{\omega} \in H(\Omega) \cap W(\Omega)$ denotes the approximate solution used to check the assumptions of Theorem 3.4. We want to emphasize that in the further course the error bound for the exact solution u^* (given by the right-hand side above) will be denoted by α.

Again, we note that by construction u^* is divergence-free and thus, the second equation of the transformed Navier-Stokes equations (1.13) is satisfied almost everywhere in Ω (note that $u \in H_0^1(\Omega, \mathbb{R}^2)$). To reconstruct the pressure associated to u^* we first need to fix a suitable solution space for the pressure. Therefore, for $R > 0$ we define the subdomains

$$\Omega_R \coloneqq ((-R, R) \times (0, 1)) \cap \Omega$$

and consider the local L^2-space

$$\mathcal{L}(\Omega) \coloneqq \Big\{u\colon \Omega \to \mathbb{R} \text{ measurable}\colon u\big|_{\Omega_R} \in L^2(\Omega_R) \text{ for all } R > 0\Big\} = L^2_{loc}(\overline{\Omega}).$$

Later on, we will see that $\mathcal{L}(\Omega)$ is the appropriate space for the pressure. Moreover, we require the following space of test functions:

$$\mathcal{H}(\Omega) \coloneqq \left\{u \in H_0^1(\Omega, \mathbb{R}^2)\colon \operatorname{supp}(u) \subseteq \overline{\Omega} \text{ is compact}\right\}.$$

Then, having a closer look at our transformed Navier-Stokes equation (1.13) introduced in Section 1.2, the ideas applied in [31, Lemmma XIII.1.1] (especially cf. [31, equation (XIII.1.6)]) suggest to prove the existence of a pressure associated to u^* satisfying the following equation:

Find $p \in \mathcal{L}(\Omega)$ such that

$$\begin{aligned} -\int_\Omega p \operatorname{div} \varphi \,\mathrm{d}(x,y) = &-\int_\Omega [(u^* \cdot \nabla)u^* + (u^* \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u^*] \cdot \varphi \,\mathrm{d}(x,y) \\ &+ \frac{1}{Re}\int_\Omega (g \cdot \varphi - \nabla u^* \bullet \nabla\varphi) \,\mathrm{d}(x,y) \quad \text{for all } \varphi \in \mathcal{H}(\Omega). \end{aligned} \tag{7.1}$$

In the sense of the previous Sections we can identify the equation for the pressure above as equation in the dual space of $H_0^1(G,\mathbb{R}^2)$ for any (sub) domain $G \subseteq \Omega$. Therefore, we shortly extend the techniques introduced in Section 2.2 with the space $H(\Omega)$ replaced by $H_0^1(G,\mathbb{R}^2)$. Hence, we can define the weak Laplacian for a function $u \in H_0^1(G,\mathbb{R}^2)$ as an element of $H^{-1}(G,\mathbb{R}^2)$ using the same formulas (cf. (2.9) and (2.11)), i.e., we define

$$(-\Delta u)[\varphi] \coloneqq \int_G \nabla u \bullet \nabla \varphi \, \mathrm{d}(x,y) \quad \text{for all } \varphi \in H_0^1(G,\mathbb{R}^2). \tag{7.2}$$

In view of (2.10) we obtain the corresponding estimate $\|-\Delta u\|_{H^{-1}(G,\mathbb{R}^2)} \leq \|\nabla u\|_{L^2(G,\mathbb{R}^{2\times 2})}$ for all $u \in H_0^1(G,\mathbb{R}^2)$ which shows a posteriori that $-\Delta u$ indeed defines a bounded linear functional on $H_0^1(G,\mathbb{R}^2)$.

Furthermore, the same arguments as in Section 2.2 show that each $u \in L^q(G,\mathbb{R}^2)$ with $q \in (1,2]$ defines a bounded linear functional on $H_0^1(G,\mathbb{R}^2)$ via

$$u[\varphi] \coloneqq \int_\Omega u \cdot \varphi \, \mathrm{d}(x,y) \quad \text{for all } \varphi \in H_0^1(G,\mathbb{R}^2)$$

(cf. (2.13)). Moreover, we have the estimate $\|u\|_{H^{-1}(G,\mathbb{R}^2)} \leq C_r \|u\|_{L^q(G,\mathbb{R}^2)}$ where $\frac{1}{r} + \frac{1}{q} = 1$ holds true.

Additionally, we define the weak gradient for functions $p \in L^2(G)$ as an element in $H^{-1}(G,\mathbb{R}^2)$ via

$$(\nabla p)[\varphi] \coloneqq -\int_G p \operatorname{div} \varphi \, \mathrm{d}(x,y) \quad \text{for all } \varphi \in H_0^1(G,\mathbb{R}^2). \tag{7.3}$$

By Cauchy-Schwarz' inequality and Lemma A.1 we calculate

$$|(\nabla p)[\varphi]| \leq \|p\|_{L^2(G)} \|\operatorname{div} \varphi\|_{L^2(G)} \leq \sqrt{2} \|p\|_{L^2(G)} \|\varphi\|_{H_0^1(G,\mathbb{R}^2)} \quad \text{for all } \varphi \in H_0^1(G,\mathbb{R}^2)$$

which a posteriori proves that ∇p defines a bounded linear functional on $H_0^1(G,\mathbb{R}^2)$.

7.1 Existence Theorem

To prove existence of a pressure $p^* \in \mathcal{L}(\Omega)$ (associated to u^*) on our unbounded domain we strongly exploit the following result for bounded (sub) domains.

Proposition 7.1. *Let $G \subseteq \Omega$ be a bounded subdomain of Ω with Lipschitz boundary. Then, for a given functional $f \in H^{-1}(G,\mathbb{R}^2)$ there exists*

$$p \in L_0^2(G) \coloneqq \left\{ q \in L^2(G) \colon \int_G q \, \mathrm{d}(x,y) = 0 \right\}$$

with $\nabla p = f$ (in $H^{-1}(G,\mathbb{R}^2)$) if and only if

$$f[\varphi] = 0 \quad \textit{for all } \varphi \in \left\{ u \in H_0^1(G,\mathbb{R}^2) \colon \operatorname{div} u = 0 \right\}.$$

A proof of Proposition 7.1 can be found for instance in [92, (proof of) Proposition 23.2, p. 444].

Successively applying Proposition 7.1 to our subdomains Ω_R we can prove the following result for our unbounded domain Ω.

Theorem 7.2. *Let $f \in H^{-1}(\Omega, \mathbb{R}^2)$ be a bounded functional such that $f[\varphi] = 0$ for all $\varphi \in H(\Omega)$. Then, there exists $p \in \mathcal{L}(\Omega)$ with*

$$f[\varphi] = \int_\Omega p \operatorname{div} \varphi \, \mathrm{d}(x,y) \quad \textit{for all } \varphi \in \mathcal{H}(\Omega).$$

Proof. Due to the embedding $H_0^1(\Omega_R, \mathbb{R}^2) \subseteq H_0^1(\Omega, \mathbb{R}^2)$ (by zero extension) for all $R > 0$ we have

$$f[\varphi] = 0 \quad \text{for all } \varphi \in \left\{ u \in H_0^1(\Omega_R, \mathbb{R}^2) \colon \operatorname{div} u = 0 \right\}.$$

Thus, for fixed $R \in \mathbb{N}$ Proposition 7.1 yields the existence of $p_R \in L^2(\Omega_R)$ such that

$$f[\varphi] = \int_{\Omega_R} p_R \operatorname{div} \varphi \, \mathrm{d}(x,y) \quad \text{for all } \varphi \in H_0^1(\Omega_R, \mathbb{R}^2). \tag{7.4}$$

Moreover, for all $\tilde{R} \in \mathbb{N}$ with $R \leq \tilde{R}$ we calculate

$$\int_{\Omega_R} (p_R - p_{\tilde{R}}) \operatorname{div} \varphi = 0 \quad \text{for all } \varphi \in H_0^1(\Omega_R, \mathbb{R}^2).$$

Hence, using definition (7.3) we obtain $\nabla(p_R - p_{\tilde{R}}) = 0$ in $H^{-1}(\Omega_R, \mathbb{R}^2)$ which implies that $p_R - p_{\tilde{R}}$ is constant almost everywhere on Ω_R. Thus, by successively choosing additive constants (for $R \in \mathbb{N}$) we can achieve

$$p_R = p_{\tilde{R}} \text{ almost everywhere on } \Omega_R \quad \text{for all } R \leq \tilde{R}.$$

Finally, $p \in \mathcal{L}(\Omega)$ defined by

$$p\big|_{\Omega_R} \coloneqq p_R \quad \text{for all } R \in \mathbb{N}$$

satisfies the desired equality since for a fixed test function $\varphi \in \mathcal{H}(\Omega)$ we find $R \in \mathbb{N}$ such that $\operatorname{supp}(\varphi) \subseteq \overline{\Omega_R}$, i.e., using the definition of p together with (7.4) we get

$$\int_\Omega p \operatorname{div} \varphi \, \mathrm{d}(x,y) = \int_{\Omega_R} p_R \operatorname{div} \varphi \, \mathrm{d}(x,y) = f[\varphi].$$

□

Remark 7.3. *We note that in general unbounded domains many results show that the pressure lies in $L^2_{loc}(\Omega)$ (cf. [64, Corollary 2.2]). Nevertheless, since our domain is bounded in y-direction in our setting we achieve a stronger result.*

Finally, to prove existence of the desired pressure p^* (corresponding to u^*) satisfying (7.1) we apply Theorem 7.2 to the functional

$$f \coloneqq \frac{1}{Re}(g + \Delta u^*) - [(u^* \cdot \nabla)u^* + (u^* \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u^*] \in H^{-1}(\Omega, \mathbb{R}^2)$$

which satisfies $f[\varphi] = 0$ for all $\varphi \in H(\Omega)$ due to the fact that u^* is a solution to our Navier-Stokes equations (1.15). Hence, we proved the following Corollary.

Corollary 7.4. *Let $u^* \in H(\Omega)$ be a (weak) solution of the Navier-Stokes equation* (1.15)*. Then, there exists a solution $p^* \in \mathcal{L}(\Omega)$ of* (7.1)*. Moreover, (u^*, p^*) is a weak solution of the transformed Navier-Stokes equations* (1.13)*, i.e.,*

$$\begin{aligned}
&\int_\Omega \big(\nabla u^* \bullet \nabla\varphi + Re([(u^* \cdot \nabla)u^* + (u^* \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u^*] \cdot \varphi - p^* \operatorname{div}\varphi)\big)\,\mathrm{d}(x,y)\\
&\qquad\qquad = \int_\Omega g \cdot \varphi\,\mathrm{d}(x,y) \quad \textit{for all } \varphi \in \mathcal{H}(\Omega),\\
&\int_\Omega q \operatorname{div} u^*\,\mathrm{d}(x,y) = 0 \quad \textit{for all } q \in L^2(\Omega).
\end{aligned}$$

7.2 Computation of a Numerical Approximation

In the context of computer-assisted proofs besides the pure existence result usually an enclosure result for the exact solution and a corresponding approximate solution is of interest. Since our existence proof for the pressure presented in the previous Section is independent of an approximate solution we do not obtain the desired enclosure result directly. However, for the computation of an error bound we first of all need an approximate solution for the pressure. Therefore, we use our computational domain Ω_0 again (cf. Chapter 4) and consider the system

$$\begin{aligned}
&\text{Find } (u,p) \in H_0^1(\Omega_0, \mathbb{R}^2) \times L^2(\Omega_0) \text{ such that}\\
&\int_{\Omega_0} \big(\nabla u \bullet \nabla\varphi + Re([(u \cdot \nabla)u + (u \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u] \cdot \varphi - p \operatorname{div}\varphi)\big)\,\mathrm{d}(x,y)\\
&\qquad\qquad = \int_{\Omega_0} g \cdot \varphi\,\mathrm{d}(x,y) \quad \text{for all } \varphi \in H_0^1(\Omega_0, \mathbb{R}^2),\\
&\int_{\Omega_0} q \operatorname{div} u\,\mathrm{d}(x,y) = 0 \quad \text{for all } q \in L^2(\Omega_0).
\end{aligned}$$

Since this problem has a saddle point structure, a lot of well-known numerical standard algorithms can be used to compute the desired approximation. In our application we use Taylor-Hood (mixed) finite elements where the idea of implementation can be found in many text books about finite elements (see e.g. [47, Chapter 6]), i.e., we are aiming at a solution of the following discrete problem

$$\begin{aligned}
&\text{Find } (u_h,p_h) \in P_{2,h,0}^2 \times P_{1,h} \subseteq H_0^1(\Omega_0, \mathbb{R}^2) \times H^1(\Omega_0) \text{ such that}\\
&\int_{\Omega_0} \big(\nabla u_h \bullet \nabla\varphi_h + Re([(u_h \cdot \nabla)u_h + (u_h \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u_h] \cdot \varphi_h - p_h \operatorname{div}\varphi_h)\big)\,\mathrm{d}(x,y)\\
&\qquad\qquad = \int_{\Omega_0} g \cdot \varphi_h\,\mathrm{d}(x,y) \quad \text{for all } \varphi_h \in P_{2,h,0}^2,\\
&\int_{\Omega_0} q_h \operatorname{div} u_h\,\mathrm{d}(x,y) = 0 \quad \text{for all } q_h \in P_{1,h},
\end{aligned}$$

where $P_{1,h}$ denotes the (discrete) linear Lagrangian finite element space and $P_{2,h,0}$ is the (discrete) quadratic Lagrangian finite element space with Dirichlet boundary conditions. Hence, we actually obtain an approximate solution $(\tilde{u}, \tilde{p}) \in H_0^1(\Omega_0, \mathbb{R}^2) \times H^1(\Omega_0)$ which yields $\nabla\tilde{p} \in L^2(\Omega_0, \mathbb{R}^2)$.

Having computed the approximation pair $(\tilde{u}, \tilde{p})$, we drop the approximation for the velocity field $\tilde{u}$ and just use the second part $\tilde{p}$ as the desired approximation for our pressure. At this stage, we want to emphasize that for the velocity we stick to the approximate solution $\tilde{\omega}$ used in Theorem 3.4, i.e., in the further course $(\tilde{\omega}, \tilde{p})$ is considered as approximate solution for our transformed Navier-Stokes equations (1.13).

7.3 Computation of an Error Bound

As already mentioned in the previous Sections, beside the pure existence results typically computer-assisted proofs also provide enclosures for the exact solution. Thus, in this Section we present a procedure to obtain an error bound for the approximation $\tilde{p}$ for the pressure computed by the means of the previous Section.

Therefore, in the sense of (7.3) we consider the weak gradient for the pressure $p^* \in \mathcal{L}(\Omega)$ as the linear functional

$$(\nabla p^*)[\varphi] = -\int_\Omega p^* \operatorname{div} \varphi \quad \text{for all } \varphi \in \mathcal{H}(\Omega) \tag{7.5}$$

defined on $\mathcal{H}(\Omega)$. Moreover, since p^* is a solution of our equation for the pressure (7.1) we obtain

$$\begin{aligned}(\nabla p^*)[\varphi] = &-\int_\Omega [(u^* \cdot \nabla)u^* + (u^* \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u^*] \cdot \varphi \,\mathrm{d}(x,y)\\ &+ \frac{1}{Re}\int_\Omega (g \cdot \varphi - \nabla u^* \bullet \nabla\varphi)\,\mathrm{d}(x,y) \quad \text{for all } \varphi \in \mathcal{H}(\Omega).\end{aligned}$$

Thus, using the fact that the right-hand side actually defines a bounded linear functional on $H_0^1(\Omega, \mathbb{R}^2)$ the functional $\nabla p^* \colon \mathcal{H}(\Omega) \to \mathbb{R}$ is bounded on $\mathcal{H}(\Omega)$ and, since $\mathcal{H}(\Omega)$ is dense in $H_0^1(\Omega, \mathbb{R}^2)$, it can be extended to a bounded linear functional on the entire space $H_0^1(\Omega, \mathbb{R}^2)$ (which will be denoted by ∇p^* again). Nevertheless, we note that the integral representation (7.5) only holds on the subspace $\mathcal{H}(\Omega)$.

Using the extension of ∇p^* and the definitions at the beginning of this Chapter we obtain the equality

$$\nabla p^* = \frac{1}{Re}\left(\Delta u^* - Re\,[(u^* \cdot \nabla)u^* + (u^* \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u^*] + g\right)$$

in $H^{-1}(\Omega, \mathbb{R}^2)$.

Extending $\tilde{p} \in L^2(\Omega_0)$ by zero we obtain a function in $L^2(\Omega)$ which will be denoted by $\tilde{p}$ again. Thus, by the definition of the weak gradient for L^2-functions (see (7.3)) we obtain $\nabla\tilde{p} \in H^{-1}(\Omega, \mathbb{R}^2)$. Hence, we are in a position to rewrite the difference of the weak gradients of the approximation and the solution p^* in $H^{-1}(\Omega, \mathbb{R}^2)$ as follows

$$\nabla\tilde{p} - \nabla p^* = \frac{1}{Re}\left(-\Delta u^* + Re\,[(u^* \cdot \nabla)u^* + (u^* \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u^* + \nabla\tilde{p}] - g\right). \tag{7.6}$$

Since the exact solution u^* (provided by Theorem 3.4) is not in hand explicitly, we use the approximation $\tilde{\omega}$ and write the difference above in terms of $\tilde{\omega}$ instead of u^*. We note that the error bound α in Theorem 3.4 is small, if the approximate solution is "sufficiently accurate".

Hence, using $\tilde{\omega}$ we equivalently rewrite the right-hand side of (7.6) to

$$\begin{aligned}\nabla\tilde{p} - \nabla p^* = {} & \frac{1}{Re}\left(-\Delta\tilde{\omega} + Re\left[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega} + \nabla\tilde{p}\right] - g\right)\\ & + (u^*\cdot\nabla)u^* - (\tilde{\omega}\cdot\nabla)\tilde{\omega} + ((u^*-\tilde{\omega})\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)(u^*-\tilde{\omega})\\ & + \frac{1}{Re}(\Delta\tilde{\omega} - \Delta u^*)\end{aligned}$$

which has to be understood as an equation in $H^{-1}(\Omega,\mathbb{R}^2)$ again.

Thus, to compute the desired error bound for $\|\nabla\tilde{p} - \nabla p^*\|_{H^{-1}(\Omega,\mathbb{R}^2)}$ we are left with the computation of upper bounds for the following terms:

- $\|-\Delta\tilde{\omega} + Re\left[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega} + \nabla\tilde{p}\right] - g\|_{H^{-1}(\Omega,\mathbb{R}^2)}$. (7.7)
- $\|(u^*\cdot\nabla)u^* - (\tilde{\omega}\cdot\nabla)\tilde{\omega} + ((u^*-\tilde{\omega})\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)(u^*-\tilde{\omega})\|_{H^{-1}(\Omega,\mathbb{R}^2)}$. (7.8)
- $\|\Delta\tilde{\omega} - \Delta u^*\|_{H^{-1}(\Omega,\mathbb{R}^2)}$. (7.9)

In the further course we treat each of these norms individually. Having a closer look at the first one (cf. (7.7)) we realize that it has the same structure as in the computation of the defect bound (cf. (5.4)). The only difference to the situation considered in Chapter 5 is the additional pressure term $\nabla\tilde{p}$. Since the approximation of the pressure $\tilde{p}$ computed with the algorithm described in Section 7.2 satisfies $\nabla\tilde{p} \in L^2(\Omega_0,\mathbb{R}^2)$ all considerations of Chapter 5, especially the calculations before (and in) (5.6) can be used mutatis mutandis to obtain

$$\begin{aligned}&\|-\Delta\tilde{\omega} + Re\left[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\Gamma + (\Gamma\cdot\nabla)\tilde{\omega} + \nabla\tilde{p}\right] - g\|_{H^{-1}(\Omega,\mathbb{R}^2)}\\ &\quad\le \|\tilde{\rho} - \nabla\tilde{\omega}\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})}\\ &\qquad + C_2\Big(\|-\operatorname{div}\tilde{\rho} + Re\left[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\tilde{\Gamma} + (\tilde{\Gamma}\cdot\nabla)\tilde{\omega} + \nabla\tilde{p}\right] - \tilde{g}\|_{L^2(\Omega_0,\mathbb{R}^2)}\\ &\qquad\quad + Re\Big[\|\nabla(\tilde{V} - V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})}\|U+\omega\|_{L^\infty(\Omega_0,\mathbb{R}^2)}\\ &\qquad\qquad + \|\tilde{V} - V\|_{L^2(\Omega_0,\mathbb{R}^2)}\|\nabla(U + \tilde{\omega} - \tilde{V})\|_{L^\infty(\Omega_0,\mathbb{R}^{2\times 2})}\Big]\Big),\end{aligned} \tag{7.10}$$

where $\tilde{g}$ and $\tilde{\Gamma}$ are defined in (5.5) and $\tilde{\rho} \in H(\operatorname{div},\Omega,\mathbb{R}^{2\times 2})$ denotes an approximation to $\nabla\tilde{\omega}$ again (cf. Section 5.2).

Remark 7.5. *Theoretically we can use the same approximation $\tilde{\rho}$ computed to obtain the defect bound δ (cf. Section 5.2). However, in practice to obtain a tighter defect bound we compute a new approximation via the functional $J\colon H(\operatorname{div},\Omega,\mathbb{R}^{2\times 2}) \to \mathbb{R}$ given by*

$$\begin{aligned}J(\tilde{\rho}) :={} & \frac{1}{2}\|\tilde{\rho} - \nabla\tilde{\omega}\|^2_{L^2(\Omega,\mathbb{R}^{2\times 2})}\\ & + \frac{1}{2}{C_2}^2\left\|-\operatorname{div}\tilde{\rho} + Re\left[(\tilde{\omega}\cdot\nabla)\tilde{\omega} + (\tilde{\omega}\cdot\nabla)\tilde{\Gamma} + (\tilde{\Gamma}\cdot\nabla)\tilde{\omega} + \nabla\tilde{p}\right] - \tilde{g}\right\|^2_{L^2(\Omega,\mathbb{R}^2)}.\end{aligned}$$

To deal with the second norm stated in (7.8), we use the equality $\Gamma + \tilde{\omega} = U + \omega$ (cf. definition (1.12) and (3.8)) to rewrite the term in the norm again. Hence, we obtain

$$\begin{aligned}
&(u^* \cdot \nabla)u^* - (\tilde{\omega} \cdot \nabla)\tilde{\omega} + ((u^* - \tilde{\omega}) \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)(u^* - \tilde{\omega}) \\
&\quad = (u^* \cdot \nabla)u^* - (\tilde{\omega} \cdot \nabla)\tilde{\omega} + ((u^* - \tilde{\omega}) \cdot \nabla)(U + \omega) + ((U + \omega) \cdot \nabla)(u^* - \tilde{\omega}) \\
&\qquad - ((u^* - \tilde{\omega}) \cdot \nabla)\tilde{\omega} - (\tilde{\omega} \cdot \nabla)(u^* - \tilde{\omega}) \\
&\quad = ((u^* - \tilde{\omega}) \cdot \nabla)(U + \omega) + ((U + \omega) \cdot \nabla)(u^* - \tilde{\omega}) + ((u^* - \tilde{\omega}) \cdot \nabla)(u^* - \tilde{\omega}).
\end{aligned}$$

Now, to estimate the H^{-1}-norm we first use the triangle inequality (to split the norm into three parts) and apply Lemma A.9 to the individual parts. Thus, we calculate

$$\begin{aligned}
&\|(u^* \cdot \nabla)u^* - (\tilde{\omega} \cdot \nabla)\tilde{\omega} + ((u^* - \tilde{\omega}) \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)(u^* - \tilde{\omega})\|_{H^{-1}(\Omega,\mathbb{R}^2)} \\
&\quad \leq C_2 \left(C_2 \|\nabla(U + \omega)\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})} + \|U + \omega\|_{L^\infty(\Omega,\mathbb{R}^2)} \right) \|\tilde{\omega} - u^*\|_{H_0^1(\Omega,\mathbb{R}^2)} \\
&\qquad + {C_4}^2 \|\tilde{\omega} - u^*\|^2_{H_0^1(\Omega,\mathbb{R}^2)}
\end{aligned}$$

Then, using the error bound α for u^* provided by Theorem 3.4 we get

$$\begin{aligned}
&\|(u^* \cdot \nabla)u^* - (\tilde{\omega} \cdot \nabla)\tilde{\omega} + ((u^* - \tilde{\omega}) \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)(u^* - \tilde{\omega})\|_{H^{-1}(\Omega,\mathbb{R}^2)} \\
&\quad \leq \left[C_2 \left(C_2 \|\nabla(U + \omega)\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})} + \|U + \omega\|_{L^\infty(\Omega,\mathbb{R}^2)} \right) + {C_4}^2 \alpha \right] \alpha.
\end{aligned} \tag{7.11}$$

An upper bound for the remaining norm $\|\Delta\tilde{\omega} - \Delta u^*\|_{H^{-1}(\Omega,\mathbb{R}^2)}$ (cf. (7.9)) can easily be obtained using the estimate for the weak divergence operator introduced in (7.2). Hence, we calculate

$$\|\Delta\tilde{\omega} - \Delta u^*\|_{H^{-1}(\Omega,\mathbb{R}^2)} \leq \|\nabla u^* - \nabla\tilde{\omega}\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \leq \|u^* - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \alpha. \tag{7.12}$$

Combining the three results in (7.10), (7.11) and (7.12) we obtain the desired error bound α_p for our pressure p^* and its corresponding approximate solution $\tilde{p}$:

$$\begin{aligned}
&\|\nabla\tilde{p} - \nabla p^*\|_{H^{-1}(\Omega,\mathbb{R}^2)} \\
&\quad \leq \left[C_2 \left(C_2 \|\nabla(U + \omega)\|_{L^\infty(\Omega,\mathbb{R}^{2\times 2})} + \|U + \omega\|_{L^\infty(\Omega,\mathbb{R}^2)} \right) + {C_4}^2 \alpha \right] \alpha \\
&\qquad + \frac{1}{Re}(\alpha + \|\tilde{\rho} - \nabla\tilde{\omega}\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})}) \\
&\qquad + C_2 \Big(\frac{1}{Re} (\| -\operatorname{div}\tilde{\rho} + Re \Big[(\tilde{\omega} \cdot \nabla)\tilde{\omega} + (\tilde{\omega} \cdot \nabla)\tilde{\Gamma} + (\tilde{\Gamma} \cdot \nabla)\tilde{\omega} + \nabla\tilde{p} \Big] - \tilde{g}\|_{L^2(\Omega_0,\mathbb{R}^2)} \\
&\qquad\quad + \Big[\|\nabla(\tilde{V} - V)\|_{L^2(\Omega_0,\mathbb{R}^{2\times 2})} \|U + \omega\|_{L^\infty(\Omega_0,\mathbb{R}^2)} \\
&\qquad\qquad + \|\tilde{V} - V\|_{L^2(\Omega_0,\mathbb{R}^2)} \|\nabla(U + \tilde{\omega} - \tilde{V})\|_{L^\infty(\Omega_0,\mathbb{R}^{2\times 2})} \Big] \Big).
\end{aligned}$$

8 Results and Summary

In the following, we present the verified results obtained by the application of Theorem 3.4 to several different domains and Reynolds numbers. Especially, the geometry presented in Figure 8.1 we use as an example domain to point out differences between our approaches for the computation of the norm bounds K and K^* satisfying (A2) and (A3) respectively. In the further course, this domain (as printed in Figure 8.1) will be referred as our "example domain". However, for our computations we use many more domains which will be presented in Section 8.3.

Our program uses the finite element software M++ (Meshes, Multigrid and More) developed by Wieners and his group (see [113] as well as [114] on gitLab). For more details about the software M++ we refer the reader to Chapter 9 and the references given in [114]. In the latest version, M++ also provides standard interval arithmetic operations and several routines for solving matrix eigenvalue problems as well as for the computation of eigenvalue bounds and the homotopy method. Again, for details we refer to Chapter 9. The results presented in the further course are based on computations running on the parallel ma-pde-cluster provided by Wieners and his working group as well as on the parallel high performance cluster HoreKa (see [96]). The programs and routines for obtaining our results consist of several thousand lines of code and thus, clearly cannot be presented here. However, among others the version used for the computations in this thesis can be found in the corresponding project on gitLab (see [116]).

Before going into the details concerning the comparison of our different approaches for the computation of the desired norm bounds we shortly give a rough overview about our algorithms used for our computer-assisted proof. In Algorithm 1 we describe our computations using the first approach (cf. Section 6.1) to obtain the desired norm bounds K and K^* respectively.

For the second approach together with the straightforward homotopy method (cf. Section 6.2) we have to adapt our algorithm slightly. In particular, lines 5-9 in Algorithm 1 have to be replaced by those presented in Algorithm 2. In the case of the extended coefficient homotopy method we have to replace lines 5-9 in Algorithm 1 by those presented in Algorithm 3.

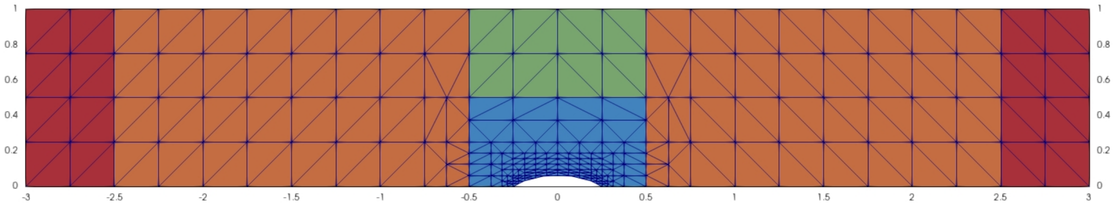


Figure 8.1: Example domain with its (coarsest) triangulation and subregions

Algorithm 1: Algorithm for our computer-assisted proof using the first approach for the computation of the norm bounds

```
$Re$: Prescribed Reynolds number
Compute approximate solution $\tilde{\omega}$ (cf. Section 4.2)
Compute $\tilde{V}$ and $\tilde{\rho}$ (cf. Section 5.2)
Evaluate $L^\infty$-norms via Bernstein polynomials (cf. Section 5.1)
Compute defect bound $\delta$ (cf. Chapter 5)
if $2C_2Re\|U+\omega\|_{L^\infty(\Omega,\mathbb{R}^2)} \geq 1$ then
    First approach failed
    break
end
Compute $K = K^*$ via (6.5) (cf. Section 6.1)
if assumption (3.11) in Theorem 3.4 holds true then
    Proof is successful: Theorem 3.4 yields the existence of an exact solution $u^*$
    Compute error bound $\alpha_{\min}$ and "radius of uniqueness" $\alpha_{\max}$
else
    Proof failed
    break
end
Compute approximate solution $\tilde{p}$ (cf. Section 7.2)
Compute error bound $\alpha_p$ (cf. Section 7.3)
```

Algorithm 2: Algorithm for the second approach with the straightforward homotopy method

```
Compute constants $\gamma_1 = \hat{\gamma}_1, \gamma_2 = \hat{\gamma}_2$ for the base problem (cf. p. 87 and p. 91)
if $\gamma_1 \leq 0$ then
    Second approach failed
    break
end
Enclose the smallest eigenvalues of the base problem (cf. Section 6.2.1.3)
Perform constraint, domain deformation and coefficient homotopy (cf. Section 6.2.1.2)
Perform final Lehmann-Goerisch computations (cf. Section 6.2.1.1) to define $K$ and
  $K^*$
```

Algorithm 3: Algorithm for the second approach with the extended homotopy method

```
Compute constants $\kappa = \hat{\kappa}$ (cf. Section 6.2.1.4)
Compute constants $\gamma_1, \gamma_2$ for the base problem (cf. (6.54) and (6.55))
Enclose the smallest eigenvalues of the base problem (cf. Section 6.2.1.3)
Perform constraint, domain deformation and coefficient homotopy (cf. Section 6.2.1.2)
Perform final Lehmann-Goerisch computation (cf. Section 6.2.1.1) to define $K$
Compute constants $\hat{\gamma}_1, \hat{\gamma}_2$ for the base problem (cf. (6.59) and (6.60))
Enclose the smallest eigenvalues of the "adjoint" base problem (cf. Section 6.2.1.3)
Perform constraint, domain deformation and coefficient homotopy (cf. Section 6.2.1.2)
Perform final Lehmann-Goerisch computation (cf. Section 6.2.1.1) to define $K^*$
```

8.1 Comparison of the Approaches for the Computation of the Norm Bounds

Figure 8.1 shows the example domain with its triangular mesh (on the coarsest level) used to compare our different approaches for the computation of the norm bounds K and K^* introduced in Chapter 6. For this example domain, the parameters d_1, d_2, d_3 which are required to describe the obstacle (cf. Chapter 1) as well as d_0, which additionally is needed for the definition of the function V (cf. Section 4.1), are chosen as follows:

$$d_0 := 2.5, \qquad d_1 := 0.5, \qquad d_2 := 0.5 \qquad \text{and} \qquad d_3 := 1.0.$$

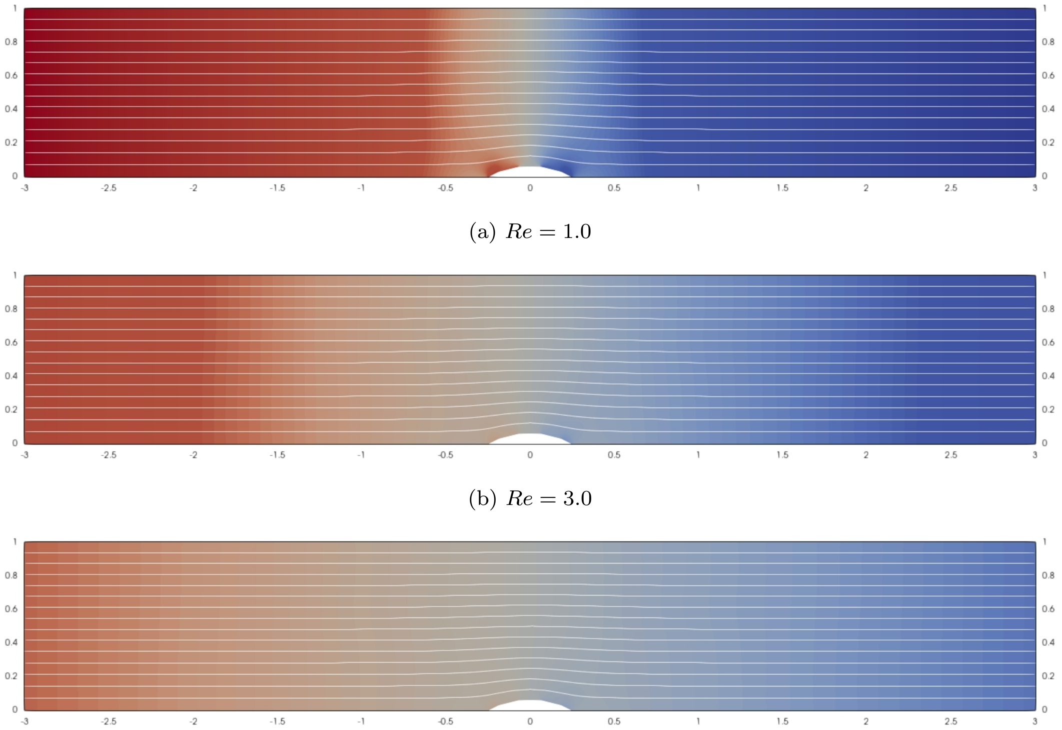
We note that the choice $d_3 := 1.0$ is somehow natural since the obstacle is only located at a single side of the strip (cf. corresponding Subsection in Section 4.1 and Remark 1.1 (i)).

Since our proofs heavily depend on the rigorous evaluation of integrals (see for instance Chapter 5) the finite element transformation $\Phi_{\mathcal{T}}$ introduced in Section 4.2 has to be evaluated rigorously using interval arithmetic operations. Since the refinement procedure in M++ is not yet implemented using interval arithmetic operations (i.e., the corners of the children cells are not computed using verified interval arithmetic algorithms) for all our (verified) computations we require the corners of the corresponding triangle $\mathcal{T}$ (see also Section 9.4.1) to be representable on the computer exactly. This fact ensures that the vertices of all children cells are exactly representable on the computer as well (at least up to a certain refinement level, which is not reached in our applications). We note that all meshes considered in this thesis are chosen such that all vertices of their cells are representable exactly on the computer.

Moreover, at this state we want to emphasize that by our choice of the parameters d_0, d_1, d_2 and d_3, the additional assumptions on the finite element mesh $\mathcal{M}$ required for the computation of the L^∞-norms (cf. (5.9) and (5.10) in Section 5.1) are satisfied for the triangulation presented in Figure 8.1. The four subregions (R_1) to (R_4) introduced in Section 5.1 to distinguish between the different definitions of Γ are printed with different colors (cf. Figure 8.1 and Figure 5.1).

As already mentioned in Section 4.2 (cf. Remark 4.3) in our approximation process we have to face the existence of reentrant corners in our domain Ω or computational domain Ω_0 respectively. Therefore, we do not choose cells of uniform diameter for the complete mesh but we actually add already refined cells in the neighborhood of the reentrant corners (cf. Figure 8.1). This strategy of dealing with reentrant corners is used in all our examples since the application of additional corner singular functions is difficult in the solenoidal case (cf. Remark 4.3).

Using the example domain introduced in Figure 8.1 (but with the mesh refined up to level 5 which means that each cell contained in our course mesh presented in Figure 8.1 is refined 5 times) for our computations, the algorithms introduced in Chapter 4 provide (exactly) divergence-free approximate solutions for several Reynolds numbers. Figure 8.2 shows some selected approximate solutions to the original Navier-Stokes equations, i.e., for some approximate solution $\tilde{\omega} \in H(\Omega)$ (to (1.15)), the associated approximate solution (to (1.7)) $U + \omega = \Gamma + \tilde{\omega}$ is plotted. We note that the approximate solutions for the velocity fields are represented by their corresponding stream lines and the scalar valued approximate pressures are plotted in the background. Moreover, in these plots, as well as in all plots of (scalar valued) approximate solutions appearing in the further course, positive values are represented by red colors and negative values are printed in shades of blue. Additionally, we note that larger (absolute) values of scalar functions are printed in darker shades, whereas smaller (absolute) values are represented by lighter colors.



(a) $Re = 1.0$

(b) $Re = 3.0$

(c) $Re = 5.0$

Figure 8.2: Different approximate solutions on example domain

Moreover, in Figure 8.3 we plotted the Euclidean norm of our approximate solution $U + \omega$ for $Re = 3.0$. It shows in which region of our domain the movement of the fluid actually can be observed. Due to our boundary conditions, on the boundary of our domain there is no movement, i.e., the velocity field vanishes (cf. dark green color in Figure 8.3). Whereas in the middle of the strip the red color indicates the fastest movement of the fluid.

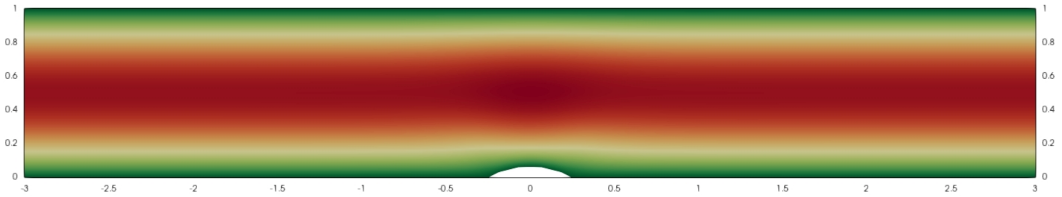


Figure 8.3: Euclidean norm of the approximate solution for $Re = 3.0$

With these approximate solutions in hand we first compute their defect bound δ using the procedure presented Chapter 5. Again, we note that for the computation of δ all integrals and L^∞-norms need to be evaluated using interval arithmetic operations. Therefore, as suggested in Chapter 5 we additionally compute the approximation $\tilde{V}$ to our function V

and use (5.6) to define the desired defect bound δ which satisfies the first assumption (A1) needed in Theorem 3.4.

Having computed the defect bound δ, we now require the norm bounds K and K^* introduced in assumptions (A2) and (A3) respectively. In the further course, we apply the different approaches introduced in Chapter 6 to compute the desired norm bounds and point out differences between the approaches.

First Approach

We start with the application of the first approach (see Section 6.1). Therefore, we use our approximate solutions (computed on the example domain) corresponding to different Reynolds numbers to compute the norm bounds K and K^* respectively. At this stage, we want to emphasize that in our first approach we made the choice $\sigma = 0$, where σ denotes the parameter of the inner product (cf. beginning of Section 6.1). Having K and K^* in hand (i.e., assumptions (A2) and (A3) of Theorem 3.4 are satisfied with K and K^* respectively), we check the crucial inequality (3.11) of Theorem 3.4, i.e., we check if the inequality $4K^2{C_4}^2 Re\,\delta < 1$ holds true.

In the affirmative case, Theorem 3.4 provides the existence of an exact solution u^* of our Navier-Stokes equations (1.15) and a corresponding error estimate for $\|u^* - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)}$. The results are presented in Table 8.1 where the successful application of Theorem 3.4 is indicated by writing down a corresponding value for the error estimate, whereas the missing of an error bound implies the failure of Theorem 3.4 for this choice of parameters, i.e., we could not verify the crucial inequality (3.11) in this case (using the first approach). We note that all scalar values in Table 8.1 as well as in every table in the further course are rounded upwards if the value represents an upper bound and rounded downwards otherwise.

Re	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla\tilde{p}\|_{H^{-1}}$	$2C_2 Re\|U + \omega\|_{L^\infty}$
	K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.04557009	0.0606225_{7}^{8}	0.05625958	0.01790798	0.15128232	0.17734018
	1.21556927		3.59872171	0.02669095		
1.5	0.04557040	0.1142377_{0}^{1}	0.06396965	0.02036218	0.12930864	0.26602739
	1.36244866		2.10999994	0.03034881		
2.0	0.04557483	0.1970889_{0}^{1}	0.07450074	0.02371433	0.12578686	0.35472626
	1.54972989		1.35893709	0.03534502		
2.5	0.04557102	0.3311274_{0}^{1}	0.09008390	0.02867460	0.13413203	0.44343698
	1.79674172		0.89901383	0.04273806		
3.0	0.04557140	0.5623587_{3}^{4}	0.11724996	0.03732182	0.15901863	0.53215972
	2.13748162		0.57560323	0.05562631		
3.5	0.04557230	0.9991811_{5}^{6}	0.23373186	0.07439916	0.29489009	0.62089467
	2.63778932		0.24750259	0.11088824		
4.0	0.04557234	1.9466539_{0}^{1}	-	-	-	0.70964199
	3.44402411		-	-		

Table 8.1: Results on example domain: First approach

Remark 8.1. *As already mentioned in Remark 6.5, both norm bounds $K > 0$ and $K^* > 0$ coincide provided both constants exist such that assumptions (A2) and (A3) hold true* *respectively. Therefore, in Table 8.1 as well as in the following tables we only present the norm bound K appearing in the crucial inequality of Theorem 3.4. However, we want to emphasize that in all our examples we computed both constants to guarantee their existence. Finally, in the affirmative case, we can choose the same constant for K and K^* respectively.*

Additionally, in view of Theorem 3.7 we computed the value $\alpha_{\max}$ corresponding to the radii of uniqueness introduced Section 3.2.

Moreover, in the case where the application of Theorem 3.4 was successful, Corollary 3.5 for instance yields the error bounds $\|u^* - \tilde{\omega}\|_{L^2(\Omega,\mathbb{R}^2)}$ and $\|u^* - \tilde{\omega}\|_{L^4(\Omega,\mathbb{R}^2)}$ which are also listed in Table 8.1. Finally, in the affirmative case we use the strategy introduced in Chapter 7 to compute the desired error bound $\|\nabla\tilde{p} - \nabla p^*\|_{H^{-1}(\Omega,\mathbb{R}^2)}$ for the pressure. These results are also listed in Table 8.1.

At this stage, we want to point out that our first approach fails for the Reynolds number $Re = 4.0$ (cf. Table 8.1). Formula (6.5) for computing the norm bound suggests that K increases with the Reynolds number (assuming that the other terms remain in the same magnitude, which we might expect if the same mesh is used in all the computations). Hence, on the basis of the results in Table 8.1 we expect that the first approach also fails for higher Reynolds numbers.

Considering the values in the last column of Table 8.1 we realize that our (first) approach actually does not fail because the crucial inequality $2C_2Re\|U + \omega\|_{L^\infty(\Omega,\mathbb{R}^2)} < 1$ (cf. Section 6.1) is not satisfied but the inequality (3.11) in Theorem 3.4 does not hold true. Thus, we might expect that the first approach is successful for slightly larger values of Re if we compute another approximate solution of higher accuracy (which would result in a lower defect bound δ). However, on the one hand, computing the approximate solutions on the next finer level, results in a significantly increased computational effort and on the other hand it makes the comparability between our different approaches difficult (or even impossible). Thus, for the reason of comparability in all our computations for each of the approaches we stick to the same approximate solutions (computed on the same mesh, i.e., on the same refinement level). Nevertheless, we want to point out that in general (if we do not only compare the different approaches) the additional computational effort (for the computations on a refined mesh) is justified if Theorem 3.4 can be applied successfully with this refined mesh.

Second Approach with Straightforward Coefficient Homotopy

Next, we consider our second approach (see Section 6.2) together with the simple straightforward coefficient homotopy presented in Section 6.2.1.1. As already mentioned in Section 6.2.1.1 we have to choose a positive parameter σ appearing in the inner product. Hence, in the case of the straightforward coefficient homotopy we fix $\sigma = 1.0$ for the most of our computations.

In Table 8.2 below we present the results for different values of the Reynolds number using our second approach. Again, we list the defect bound δ and the norm bounds K and K^* (which actually coincide in all our computations). Although, in each case we have used the same approximate solution (compared to the first approach), the defect bound slightly differs from that one presented in the previous Section concerning the first approach. This difference originates from the fact that for both approaches we have chosen a different parameter for the inner product, which (via the embedding constant C_2, cf. Lemma A.2) indirectly influences the value of δ (cf. (5.6)).

Moreover, (using the second approach) we compute the desired constants K and K^* satisfying (A2) and (A3) respectively. We note that in our applications we make use of the possibility to compute the norm bound with a second approximate solution (on a much coarser level) and then, transform it to the "fine setting" (cf. strategy introduced in the beginning of Section 6.2).

Re	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla\tilde{p}\|_{H^{-1}}$	γ_1	γ_2	n_0
	K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{n}_0$
1.0	0.04558526	0.0938195_{3}^{4}	0.07395372	0.02243124	0.17912429	0.72400_{099}^{100}	1.0072156_{2}^{3}	0
	1.58332845		3.00329150	0.03350376		0.72400_{099}^{100}	1.0072156_{2}^{3}	0
1.5	0.04558551	0.1409251_{7}^{8}	0.07496833	0.02273898	0.14073141	0.6780011_{6}^{7}	1.5101269_{7}^{8}	0
	1.58442553		1.97510798	0.03396341		0.6780011_{6}^{7}	1.5101269_{7}^{8}	0
2.0	0.04558578	0.18816_{199}^{200}	0.07604058	0.02306421	0.12207151	0.6320013_{2}^{3}	2.0125862_{4}^{5}	0
	1.58552412		1.46045129	0.03444918		0.6320013_{2}^{3}	2.0125862_{4}^{5}	0
2.5	0.04558609	0.3143781_{8}^{9}	0.09142339	0.02773004	0.12879221	0.586001_{49}^{50}	2.5146028_{0}^{1}	0
	1.83305964		0.97178039	0.04141816		0.586001_{49}^{50}	2.5146028_{0}^{1}	0
3.0	0.04558646	0.3818264_{6}^{7}	0.09412761	0.02855027	0.12284498	0.5400016_{5}^{6}	3.0161861_{2}^{3}	4
	1.84412788		0.78655785	0.04264327		0.5400016_{5}^{6}	3.0161861_{2}^{3}	4
3.5	0.04558730	0.4565164_{5}^{6}	0.09797822	0.02971821	0.12053771	0.4940018_{2}^{3}	3.5173457_{5}^{6}	10
	1.86684755		0.64770817	0.04438774		0.4940018_{2}^{3}	3.5173457_{5}^{6}	10
4.0	0.04558732	0.5542974_{3}^{4}	0.10520483	0.03191015	0.12329038	0.4480019_{8}^{9}	4.0180913_{6}^{7}	16
	1.92422558		0.52781474	0.04766166		0.4480019_{8}^{9}	4.0180913_{6}^{7}	16
4.5	0.04558781	0.696866_{79}^{80}	0.11960978	0.03627938	0.13443622	0.4020021_{5}^{6}	4.5184327_{0}^{1}	28
	2.03414005		0.41266972	0.05418763		0.4020021_{5}^{6}	4.5184327_{0}^{1}	28
5.0[1]	0.04558162	0.861454_{09}^{10}	0.13944852	0.04330448	0.15680348	0.4213858_{3}^{4}	5.0183795_{8}^{9}	20
	2.09902270		0.30480891	0.06458074		0.4213858_{3}^{4}	5.0183795_{8}^{9}	20

Table 8.2: Results on example domain: Second approach with straightforward coefficient homotopy

In the case where Theorem 3.4 was successfully applied we list the bounds for the error estimates for the velocity measured in the H_0^1-, L^2- and L^4-norm respectively as well as the error bound for the pressure measured in the H^{-1}-norm. Additionally, for each value of Re we present a lower bound for the corresponding "maximal" radius of uniqueness (in the sense introduced in the previous Subsection).

Furthermore, in Table 8.2 we give an overview about the "size" of the corresponding base problems which somehow gives an indication for the magnitude of the computational effort caused by the homotopy method. In view of the notations introduced in Section 6.2.1.1, n_0 denotes the number of eigenvalues (below some ρ_0) considered in our eigenvalue homotopy corresponding to our eigenvalue problem (6.8), whereas $\hat{n}_0$ represents the number of eigenvalues used in the "adjoint" homotopy (corresponding to our eigenvalue problem (6.9)). Looking at the number of eigenvalues listed in Table 8.2 we see that (beginning from the case $Re = 3.0$) the numbers n_0 and $\hat{n}_0$ respectively rapidly increase with the Reynolds number. Hence, in the homotopy process more eigenvalues have to be considered for higher values of Re, i.e., the homotopy method becomes more and more challenging since

[1] Computed with $\sigma = 0.5$ because otherwise too many eigenvalues have to be considered in the homotopy

the computational effort massively increases with the number of eigenvalues which have to be considered.

Using the explanations in Section 6.2.1.3, the essential spectra of the base problems consist of the single values γ_1 (for the eigenvalue problem (6.8)) and $\hat{\gamma}_1$ (for the "adjoint" eigenvalue problem (6.9)). Therefore, the procedure described in Section 6.2.2 about obtaining a lower bound for the essential spectrum via the homotopy method implies that the values for γ_1 and $\hat{\gamma}_1$ listed in Table 8.2 also provide the (required) lower bounds for the essential spectra of the eigenvalue problems (6.8) and (6.9) respectively.

Finally, we note that due to the structure of our simple coefficient homotopy (see definitions of γ_2 and $\hat{\gamma}_2$ in the case of the straightforward coefficient homotopy) the magnitudes of the constants γ_2 and $\hat{\gamma}_2$ increase "moderately" with respect to the Reynolds number which results in a relatively "moderate" (compared to the extended homotopy method) number of eigenvalues appearing in the base problems. Hence, for "moderate" values of the Reynolds number the computational effort for the homotopy method is reasonable in our considerations, but for larger values the number of eigenvalues which have to be considered in the homotopy becomes too large for our numerical computations, i.e., our existence proof fails due to memory issues and not because of the computer-assisted approach itself. However, at this stage we want to emphasize that we expect our method to be successful for larger values of the Reynolds number as well if enough memory is available for the numerical computations.

To get a better impression of this memory issue, we again have a closer look at the constraint homotopy introduced in Section 6.2.1.2. Especially, the computation of the functions w_2, w_3, w_5 (introduced in Section 6.2.1.2 after (6.67)) required in the Goerisch setting to compute the desired new eigenvalue bound via (6.21) in Corollary 6.9 is of interest. As mentioned at the end of Section 6.2.1.2 (see p. 102) we minimize b_t over a finite element space contained in $H^1(S, \mathbb{R}^{2\times 2}) \times L^2(S, \mathbb{R}^2) \times H^1(S)$ in order to obtain "good" lower bounds. Hence, using Lagrangian finite elements for the computation of the functions w_2, w_3, w_5 the matrices appearing in the approximation procedure become very large which results in a huge amount of memory. In this context one can also think of using other finite element spaces for the computation of the desired functions w_2, w_3, w_5.

Moreover, we would like to mention that the number of eigenvalues which have to be considered in our homotopy process heavily depends on the choice of the parameter σ appearing in the definition of the inner product. We do not go into further details at this stage, however, we will investigate this observation later.

We note that by construction of our straightforward coefficient homotopy method our second approach (together with this homotopy) fails for large Reynolds numbers which can be seen by definition of the constants γ_1 and $\hat{\gamma}_1$ (since for Re large enough they would become negative). To overcome this disadvantage we introduced the extended coefficient homotopy for which we present some results in the following Subsection.

Second Approach with Extended Coefficient Homotopy

Similar to the previous cases in Table 8.3 we list the results for our second approach using the extended coefficient homotopy. We note that all computations in this Subsection use the parameter $\sigma = 0.25$ for the inner product defined on our space $H(\Omega)$. We note that for the reason of comparability for all computations presented in Table 8.3 we fix the constant $\sigma = 0.25$. However, since the success of our eigenvalue homotopy method heavily depends on the choice of σ we drop this restriction to a single value of σ for all further computations. Especially, in Section 8.3 we exploit the freedom in the choice of the parameter σ to obtain more cases where our computer-assisted proof is successful (cf. for instance Table 8.16).

Moreover, we would like to emphasize that in contrast to the straightforward coefficient homotopy, now the constants γ_1 and $\hat{\gamma}_1$ as well as γ_2 and $\hat{\gamma}_2$ do not coincide anymore (see (6.54), (6.55) and (6.59), (6.60) respectively) implying that the base problems for the eigenvalue problems (6.8) and (6.9) differ in this approach. For this reason, the computational effort of this approach (compared to the simple coefficient homotopy) is much larger since in any case we have to perform the complete homotopy method (starting from the base problem) for both of our eigenvalue problems.

Re	δ	$4K^2{C_4}^2 Re\,\delta$	$\lVert u^* - \tilde{\omega}\rVert_{H_0^1}$	$\lVert u^* - \tilde{\omega}\rVert_{L^2}$	$\lVert \nabla p^* - \nabla\tilde{p}\rVert_{H^{-1}}$	γ_1	γ_2	n_0
	K		$\alpha_{\max}$	$\lVert u^* - \tilde{\omega}\rVert_{L^4}$	$4K^2{C_4}^2 Re\,\delta$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{n}_0$
1.0	0.04557395	0.1007673_0^1	0.07422755	0.02367565	0.18379895	0.9198516_2^3	3.1753815_4^5	0
	1.58660767		2.79606750	0.03478300		0.9120835_6^7	3.5252354_1^2	0
1.5	0.04557424	0.1513229_4^5	0.07531538	0.02273898	0.14549569	0.8855901_5^6	5.1775018_0^1	2
	1.58750513		1.83713287	0.03529276		0.8610172_5^6	5.5397852_1^2	0
2.0	0.04557453	0.2095956_0^1	0.07807149	0.02454205	0.12916831	0.8773136_4^5	7.196987_{29}^{30}	8
	1.61801702		1.32921658	0.03658428		0.7932801_6^7	7.2387341_4^5	4
2.5	0.04557489	0.3060956_5^6	0.08696706	0.02733840	0.12825000	0.8565686_6^7	9.3411618_0^1	12
	1.74889552		0.95461193	0.04075273		0.7093061_1^2	8.9408959_1^2	10
3.0	0.04557523	0.421580_{49}^{50}	0.09700553	0.03049403	0.13171053	0.8324558_6^7	11.6488803_8^9	16
	1.87362754		0.71319323	0.04545676		0.6093056_4^5	10.6425205_5^6	16

Table 8.3: Results on example domain: Second approach with extended coefficient homotopy

A More Detailed Analysis of the Approaches

In this Section we perform a more detailed analysis for each of our approaches and point out advantages and disadvantages of each of the methods. Moreover, we present significant differences between the strategies. To easily distinguish our two versions of the second approach, in the following the second approach combined with the straightforward (or simple) coefficient homotopy method will be referred as Approach 2 (a), whereas Approach 2 (b) denotes the second approach using the extended coefficient homotopy method. In the same sense, the first approach will be referred as Approach 1 in the further course.

First, we investigate our second approaches in view of the eigenvalues occurring in the different homotopy methods. Since in our formulation of the base problem the integral $\int_{S_R} u \cdot \varphi \,\mathrm{d}(x,y)$ directly following to the constant γ_2 in (6.71) is computed on the domain S_R (which is the domain $[-3,3] \times [0,1]$ in our examples) it makes sense to choose a slightly larger computational domain for our eigenvalue computations. As suggested in several works using these techniques (cf. [117]) in our examples we use a computational domain for the eigenvalue computations with radius twice as large as for our computational domain Ω_0 on which we have computed the approximate solution $\tilde{\omega}$, i.e., we consider the finite strip $[-6,6] \times [0,1]$ for our eigenvalue computations. Moreover, having a closer look at the symmetric and antisymmetric eigenfunctions of the base problem computed in Section 6.2.1.3 (which might have high oscillations in the compact set S_R) we do not use a mesh with cells of equal diameter. To obtain tight eigenvalue bounds it makes sense to add more cells (with smaller diameter) in the region $[-3,3] \times [0,1]$ (cf. red part in Figure 8.4).

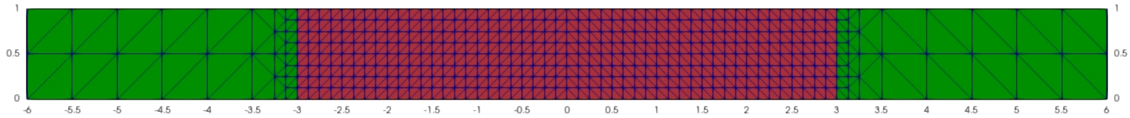


Figure 8.4: Domain for the eigenvalue computations with its (coarsest) triangulation

As already mentioned earlier, the number of eigenvalues of the base problem (and thus the eigenvalues needed to be considered in our homotopy method) heavily depends on the choice of the parameter σ of the inner product. To get an idea how the number of eigenvalues changes with respect to the magnitude of σ, we performed several computations for different values of σ but for the fixed Reynolds number $Re = 2.0$. For our test series we applied our second approach with the extended coefficient homotopy method.

Next, in Table 8.4 we list the lower bounds for the essential spectrum κ computed via the strategy introduced in Section 6.2.2 as well as Section 6.2.1.4. Based on the value of κ we can choose the lower bound ρ_0 (cf. Section 6.2.1.2) which somehow determines the number of base eigenvalues that have to be considered in our homotopy. We note that in our examples we choose the constant ρ_0 relatively "small" which results in an appropriate amount of eigenvalues. Moreover, Table 8.4 shows the crucial constants γ_0, γ_1 and γ_2 appearing in the course of the homotopy. Finally, we present the number of eigenvalues of each of the base problems below ρ_0.

σ	κ	γ_0	γ_1	γ_2	ρ_0	n_0
0.25	0.92731364	−3.41129180	0.8773136_4^5	7.196987_{29}^{30}	0.4	8
0.50	0.88790644	−3.41129180	0.8379064_4^5	7.2762225_2^3	0.4	10
1.00	0.80665580	−3.41129180	0.7601610_2^3	7.5030004_0^1	0.4	12
2.00	0.65756425	−3.41129180	0.617079_{09}^{10}	8.0324539_4^5	0.4	20

Table 8.4: Approach 2 (b): Comparison of the "size" of the base problem with respect to σ (and $Re = 2.0$)

Hence, it makes sense to choose σ "small" to avoid unreasonable computational effort concerning the homotopy method. On the other hand, our examples suggest to fix the parameter σ "large" enough since σ "small" has has a negative effect on the lower (Lehmann-Goerisch) bounds, i.e., they become worse if σ is chosen too "small". Thus, one has to find a suitable balance for the parameter of the inner product.

Next, we have a closer look at the number of eigenvalues of the base problem (below some constant ρ_0) with respect to the Reynolds number (i.e., we now fix the parameter $\sigma = 1.0$). From Table 8.5 we conclude that the number of eigenvalues below the constant ρ_0 increases with the value of the Reynolds number.

Moreover, our second approach compared to the first one provides the existence of solutions to the Navier-Stokes equations (1.15) for a larger range of Reynolds numbers if we use the same approximate solutions on the same level in both cases (compare Table 8.1 and Table 8.2 and to some extend Table 8.3). Thus, looking at the corresponding results again, we see that the first approach cannot be extended to larger values of Re, whereas in our second approach the proofs for higher Reynolds numbers do not fail because of the method itself but become more and more challenging in view of the computational effort.

Re	τ	γ_1	γ_2	$K = K^*$	ρ_0	n_0
1.0	2.00721563	0.72400_{099}^{100}	1.0072156_2^3	1.58332845	0.4	0
1.5	2.00675132	0.6780011_6^7	1.5101269_7^8	1.58442553	0.4	0
2.0	2.00629313	0.6320013_2^3	2.0125862_4^5	1.58552412	0.4	0
2.5	2.00584113	0.586001_{49}^{50}	2.5146028_0^1	1.83305964	0.3	0
3.0	2.00539538	0.5400016_5^6	3.0161861_2^3	1.84412788	0.3	4
3.5	2.00495593	0.4940018_2^3	3.5173457_5^6	1.86684755	0.3	10
4.0	2.00452285	0.4480019_8^9	4.0180913_6^7	1.92422558	0.3	16
4.5	2.00409616	0.4020021_5^6	4.5184327_0^1	2.03414005	0.3	28

Table 8.5: Approach 2 (a): Comparison of the "size" of the base problem with respect to Re (and $\sigma = 1.0$)

Finally, we compare our three approaches with respect to their computational time. For special values of Re in Table 8.6 we list the total computational times needed to achieve the results presented before for the different approaches. Additionally, we present the computational time required for the pure computation of the approximate solutions to the velocity and pressure, as well as the time needed to compute defect bound which coincides for all three approaches. In particular, the remaining computational time amounts to the computation of the norm bounds.

Re	Approach 1	Approach 2 (a)	Approach 2 (b)	Approximation	Defect computation
1.0	$1h : 51m : 42s$	$2h : 11m : 15s$	$3h : 00m : 40s$	$0h : 59m : 46s$	$0h : 49m : 05s$
2.0	$1h : 40m : 26s$	$1h : 51m : 01s$	$4h : 42m : 14s$	$0h : 49m : 17s$	$0h : 48m : 23s$
3.0	$1h : 40m : 24s$	$3h : 06m : 37s$	$18h : 55m : 37s$	$0h : 49m : 03s$	$0h : 48m : 34s$

Table 8.6: Example domain: Comparison of the computational time with respect to Re

Analysis of the Homotopy Method

In the further course, we describe the homotopy method using the extended coefficient homotopy for a specific set of parameters in detail. In our example we fix $\sigma = 0.25$ and $Re = 2.0$ and run our calculations on the example domain (cf. Figure 8.1) and the (extended) strip domain introduced in Figure 8.4. First, we consider the homotopy method corresponding to the eigenvalue problem (6.8) and afterwards, we shortly give some remarks on the "adjoint" eigenvalue problem(6.9).

Having a closer look at Table 8.3 again, we see that for our set of parameters we computed $\gamma_1 = 0.8773136_4^5$ (which yields a lower bound for the essential spectrum of the base problem). Hence, the lower bound $\rho_0 = 0.4 < \gamma_1$ is a possible choice to start our homotopy method.

Now, by the techniques presented in Section 6.2.1.3 we compute enclosures for all eigenvalues of the base problem (corresponding to the eigenvalue problem (6.8)) below $\rho_0 = 0.4$. Recall that for the computation of these eigenvalues we first have to consider the scalar valued eigenvalue problem (6.72) and count each eigenvalue with doubled multiplicity. The enclosure intervals (for the 8 eigenvalues below ρ_0) are listed in Table 8.7.

n	$\underline{\mu}_n$	$\overline{\mu}_n$
1	0.18158506	0.18158507
2	0.18158506	0.18158507
3	0.22440170	0.22440171
4	0.22440170	0.22440171
5	0.28567230	0.28567231
6	0.28567230	0.28567231
7	0.35512881	0.35512882
8	0.35512881	0.35512882

Table 8.7: Example domain: Eigenvalues of the base problem below $\rho_0 = 0.4$

On the basis of these eigenvalue enclosures, we conclude that our base problem has exactly $n_0 = 8$ eigenvalues (counted by multiplicity) below $\rho_0 = 0.4$.

Next, we apply our algorithm presented in Section 9.5.3 to compute the next homotopy parameter t_1. Moreover, in view of definition (9.7) given in Section 9.5.3 we denote the number of eigenvalues which have to be considered in the first homotopy step by N_1 which takes the value 8 in our example. Then, we check the crucial assumption needed in Corollary 6.9, i.e., we (try to) confirm the inequality

$$\frac{M_{t_1}(\tilde{u}_{N_1}^{(t_1)}, \tilde{u}_{N_1}^{(t_1)})}{\langle \tilde{u}_{N_1}^{(t_1)}, \tilde{u}_{N_1}^{(t_1)} \rangle_{H_0^1(S,\mathbb{R}^2)}} < \rho_0.$$

In the affirmative case, the application of Corollary 6.9 provides a lower bound ρ_1 for the 8th eigenvalue which can be computed by

$$\rho_1 := \frac{\rho_0 \langle \tilde{u}_{N_1}^{(t_1)}, \tilde{u}_{N_1}^{(t_1)} \rangle_{H_0^1(S,\mathbb{R}^2)} - M_{t_1}(\tilde{u}_{N_1}^{(t_1)}, \tilde{u}_{N_1}^{(t_1)})}{\rho_0 b(w_{N_1}^{(t_1)}, w_{N_1}^{(t_1)}) - \langle \tilde{u}_{N_1}^{(t_1)}, \tilde{u}_{N_1}^{(t_1)} \rangle_{H_0^1(S,\mathbb{R}^2)}} \leq \lambda_{N_1}^{(t_1)} < \rho_0.$$

Concerning our example, the results of the first homotopy step (as well as all these of the subsequent homotopy steps) are presented in Table 8.8.

Having computed the new lower bound ρ_1, we go on with this procedure (cf. description of the homotopy method in Section 6.2.1) and obtain new lower bounds ρ_i step by step (cf. Table 8.8). Finally, we obtain the lower bound $\rho_8 = 0.38421380$ for the first eigenvalue of the eigenvalue problem with parameter $t_8 = 0.96027$. Hence, ρ_8 is the desired eigenvalue bound which can be used for the computation of the norm bound K satisfying assumption (A2).

i	t_i	ρ_i	$\frac{M_{t_i}\left(\tilde{u}_{N_i}^{(t_i)},\tilde{u}_{N_i}^{(t_i)}\right)}{\left\langle \tilde{u}_{N_i}^{(t_i)},\tilde{u}_{N_i}^{(t_i)}\right\rangle_{H_0^1(S,\mathbb{R}^2)}}$	N_i
0	0.00000	0.40000000	-	-
1	0.05750	0.39911532	0.39939589	8
2	0.12200	0.39815305	0.39834289	7
3	0.15700	0.39725018	0.39751945	6
4	0.16150	0.39621090	0.39635284	5
5	0.18325	0.39526156	0.39540210	4
6	0.46425	0.39425466	0.39445505	3
7	0.76825	0.39332896	0.39371741	2
8	0.96027	0.38421380	0.39251742	1

Table 8.8: Example domain: Steps of the eigenvalue homotopy

Remark 8.2. *From Table 8.7 we gather that the first eigenvalue of the base problem is positive. Since eigenvalues are non-decreasing with respect to increasing homotopy parameter (cf. Section 6.2.1) one can think of using its lower bound for the computation of the desired norm bound K directly without any eigenvalue homotopy. However, our applications it show that for the norm bound computed via $K := \frac{1}{\sqrt{\underline{\mu}_1}}$ the crucial inequality* (3.11) *of Theorem 3.4 is not satisfied and hence, this choice of K does not lead to a successful application of Theorem 3.4. Thus, we have to perform a homotopy method anyway.*

Finally, we shortly give some remarks on the "adjoint" eigenvalue problem (6.9). Using the lower bound $\hat{\rho}_0 = 0.15$ for $\hat{\gamma}_1 = 0.79328801^7_6$ (cf. Table 8.3) we obtain the following eigenvalues of the "adjoint" eigenvalue problem below $\hat{\rho}_0$ (cf. Table 8.9).

n	$\hat{\underline{\mu}}_n$	$\hat{\overline{\mu}}_n$
1	0.09351593	0.09351594
2	0.09351593	0.09351594
3	0.13658093	0.13658094
4	0.13658093	0.13658094

Table 8.9: Example domain: Eigenvalues of the "adjoint" base problem below $\hat{\rho}_0 = 0.15$

At this stage we want to recall that for the "adjoint" constant K^* no crucial inequality has to be satisfied. The existence of a constant K^* satisfying (A3) is sufficient to prove the surjectivity of the operator $\mathrm{L}_{U+\omega}$ (cf. Proposition 3.3). As a consequence, it is not necessary to choose K^* as small as possible. Thus, for the "adjoint" problem it suffices to set $K^* := \frac{1}{\sqrt{\hat{\underline{\mu}}_1}}$ directly (without preforming any homotopy method) provided that the lower bound for the first eigenvalue of the "adjoint" base problem is positive (cf. Remark 8.2), i.e., we might not need to compute the homotopy method for the "adjoint" problem.

8.2 Existence and Enclosure Theorem for Continuous Values of the Reynolds Number

In each of the previous examples we applied Theorem 3.4 to a single discrete Reynolds number and therefore, we "only" obtained an existence and enclosure result for our Navier-Stokes equations for discrete values of Re. However, having a closer look at the proof of Theorem 3.4 again, we see that we can extend our Theorem to continuous values of the Reynolds number.

Therefore, we suppose that our assumptions (A1), (A2) and (A3) (see Chapter 3) do not only hold for a single discrete value Re, but for all Reynolds numbers in some compact interval $\left[\underline{Re}, \overline{Re}\right] \subseteq (0,\infty)$. Hence, in contrast to the assumptions introduced in Chapter 3, in the further course of this Subsection, we assume that the following assumptions hold true (uniformly in Re):

(A1b) Suppose on $\left[\underline{Re}, \overline{Re}\right]$ a uniform bound $\boldsymbol{\delta} \geq 0$ for the defect (residual) of $\tilde{\omega}$ has been computed, i.e., the following estimate holds true

$$\|\mathrm{F}\,\tilde{\omega}\|_{H(\Omega)'} \leq \boldsymbol{\delta} \quad \text{for all } Re \in \left[\underline{Re}, \overline{Re}\right].$$

(A2b) Assume a constant $\boldsymbol{K} > 0$ is in hand such that

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \boldsymbol{K} \|\Phi^{-1}\,\mathrm{L}_{U+\omega}\,u\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega),\ Re \in \left[\underline{Re}, \overline{Re}\right]$$

with $\mathrm{L}_{U+\omega}$ defined in (3.9).

(A3b) Let a constant $\boldsymbol{K}^* > 0$ be in hand such that

$$\|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \boldsymbol{K}^* \|(\Phi^{-1}\,\mathrm{L}_{U+\omega})^* u\|_{H_0^1(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega),\ Re \in \left[\underline{Re}, \overline{Re}\right].$$

Similar as before, assumptions (A2b) and (A3b) imply the bijectivity of $\mathrm{L}_{U+\omega}$ for all $Re \in \left[\underline{Re}, \overline{Re}\right]$. Then, we can mimic the proof of Theorem 3.4 and obtain the following existence and enclosure theorem for continuous values of the Reynolds number.

Theorem 8.3. *Let $\tilde{\omega} \in H(\Omega) \cap W(\Omega)$ be an approximate solution of* (1.15) *and constants $\boldsymbol{\delta} \geq 0$ and $\boldsymbol{K}, \boldsymbol{K}^* > 0$ be computed satisfying the assumptions (A1b), (A2b) and (A3b) uniformly on the compact interval $[\underline{Re}, \overline{Re}]$. If*

$$4\boldsymbol{K}^2 {C_4}^2 Re\,\boldsymbol{\delta} < 1 \quad \textit{for all } Re \in \left[\underline{Re}, \overline{Re}\right], \tag{8.1}$$

then for all $Re \in \left[\underline{Re}, \overline{Re}\right]$ there exists a locally unique solution $u^*_{Re} \in H(\Omega)$ of (1.15) *satisfying the error enclosure*

$$\|u^*_{Re} - \tilde{\omega}\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq \frac{2\boldsymbol{K}\boldsymbol{\delta}}{1 + \sqrt{1 - 4\boldsymbol{K}^2 {C_4}^2 Re\,\boldsymbol{\delta}}}.$$

Applying Theorem 8.3 several times to small compact intervals we proved the existence of a solution to our Navier-Stokes equations (1.15) (on our example domain; cf. Figure 8.1) for all $Re \in [1.0, 3.5]$. Each application of Theorem 8.3 on a suitable subinterval is listed in Table 8.10. For the most of our computations we used the first approach. However, similar to the previous observations, for larger values of Re (precisely for the interval

[3.375, 3.500]) the first approach failed. In this case, we used the second approach together with the straightforward coefficient homotopy.

Re	$\boldsymbol{\delta}$	$\boldsymbol{K}$	$4\boldsymbol{K}^2{C_4}^2 Re\,\boldsymbol{\delta}$	$\Vert u^*_{Re} - \tilde{\omega}\Vert_{H_0^1}$	$\Vert u^*_{Re} - \tilde{\omega}\Vert_{L^2}$	$\Vert u^*_{Re} - \tilde{\omega}\Vert_{L^4}$	$\Vert \nabla p^* - \nabla\tilde{p}\Vert_{H^{-1}}$
[1.000, 1.125]	0.04571812	1.24923125	$0.0_{6423461}^{7226395}$	0.05818335	0.01852034	0.02760364	1.61318551
[1.125, 1.250]	0.04571820	1.28481656	$0.0_{7643969}^{8493300}$	0.06004268	0.01911218	0.02848575	1.18336606
[1.250, 1.375]	0.04571829	1.32248992	$0.0_{8998700}^{9898572}$	0.06203717	0.01974705	0.02943199	0.90382092
[1.375, 1.500]	0.04571840	1.36244068	$0.1_{0505675}^{1460737}$	0.06418352	0.02043025	0.03045027	0.71450102
[1.500, 1.625]	0.04571848	1.40488178	$0.1_{2185902}^{3201395}$	0.06650149	0.02116809	0.03154997	0.58185646
[1.625, 1.750]	0.04571851	1.45005358	$0.1_{4063993}^{5145839}$	0.06901474	0.02196808	0.03274232	0.48614899
[1.750, 1.875]	0.04571861	1.49822840	$0.1_{6168963}^{7323890}$	0.07175217	0.02283943	0.03404102	0.41603555
[1.875, 2.000]	0.04571870	1.54971598	$0.1_{8535077}^{9770750}$	0.07474885	0.02379330	0.03546273	0.36431513
[2.000, 2.125]	0.04572294	1.60487021	$0.2_{1205035}^{2530351}$	0.07805617	0.02484605	0.03703180	0.32369057
[2.125, 2.250]	0.04571887	1.66409724	$0.2_{4221825}^{5646639}$	0.08170681	0.02600809	0.03876375	0.29357071
[2.250, 2.375]	0.04571896	1.72786549	$0.2_{7649914}^{9186021}$	0.08579502	0.02730940	0.04070330	0.27077029
[2.375, 2.500]	0.04571944	1.79671810	$0.3_{1558727}^{3219714}$	0.09040865	0.02877797	0.04289213	0.25370795
[2.500, 2.625]	0.04571914	1.87128839	$0.3_{6034173}^{7835883}$	0.09567397	0.03045398	0.04539013	0.24149996
[2.625, 2.750]	0.04571941	1.95231946	$0.4_{1183832}^{3144968}$	0.10177620	0.03239638	0.04828518	0.23323389
[2.750, 2.875]	0.04571934	2.04068896	$0.4_{7139104}^{9281792}$	0.10898345	0.03469051	0.05170448	0.22880396
[2.875, 3.000]	0.04571939	2.13744108	$0.5_{4065680}^{6416363}$	0.11772525	0.03747312	0.05585181	0.22814409
[3.000, 3.125]	0.04571962	2.24382799	$0.6_{2172465}^{4762985}$	0.12874807	0.04098179	0.06108131	0.23183904
[3.125, 3.250]	0.04571961	2.36136429	$0.7_{1725498}^{4594519}$	0.14356101	0.04569689	0.06810893	0.24134512
[3.250, 3.375]	0.04571973	2.49189959	$0.8_{3069798}^{6264791}$	0.16624558	0.05291762	0.07887106	0.26161113
[3.375, 3.500][2]	0.04572012	1.65008307	$0.3_{4492095}^{5769581}$	0.08375748	0.02540486	0.03794522	0.14526337

Table 8.10: Example domain: Approach 1 with continuous values of the Reynolds number

Remark 8.4. (i) *We note that for each compact subinterval we only pick one approximate solution $\tilde{\omega}$. Since we require a small defect bound $\boldsymbol{\delta}$ uniformly in Re it makes sense to choose the radius of the subintervals "small". In our applications it turned out that the radius* 0.125 *is sufficient to obtain the desired results. However, if the computation of $\boldsymbol{K}$ and $\boldsymbol{K}^*$ becomes more and more challenging one has to decrease the radius of the subintervals.*

(ii) *To check the crucial inequality* (8.1) *uniformly (for all $Re \in \left[\underline{Re}, \overline{Re}\right]$) we can use standard interval arithmetic operations (cf. Section 3.3) to evaluate the expression on the left-hand side of* (8.1) *for all values of the Reynolds number in the interval $\left[\underline{Re}, \overline{Re}\right]$ simultaneously. Finally, we check if the upper bound of the resulting interval value is strictly less than one.*

Finally, in the affirmative case of Theorem 8.3 we can apply Theorem 7.2 which provides the existence of a corresponding pressure for each $Re \in \left[\underline{Re}, \overline{Re}\right]$. Moreover, we can adapt the procedure introduced in Section 7.3 to obtain an error bound for the pressure as well (see Table 8.10).

[2] Computed with Approach 2 (a) since Approach 1 failed

In the subsequent section we present more results for different types of obstacles. In some cases we also proved the existence of solutions for a continuous interval of the Reynolds number (cf. Table 8.14).

8.3 More Results on Several Domains

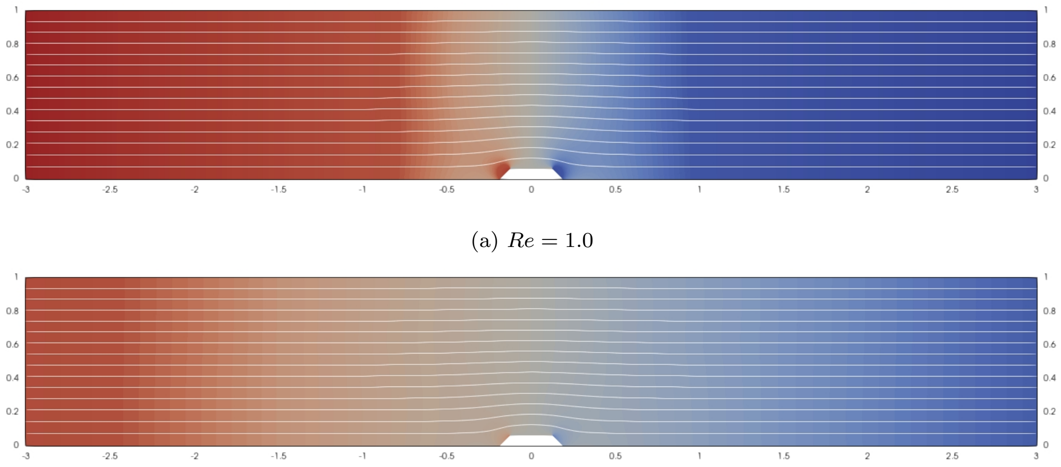
In this Section we list results for different domains and several Reynolds numbers. In view of the first example presented in Section 8.1 we continue with (symmetric) obstacles located at the boundary of the strip.

At this stage, we note that for each of the following example domains we use our first approach for the computation of the norm bounds whenever it is possible just to reduce the computational effort. However, we want to emphasize that in each parameter setting where the second approach is used (which can be gathered from the last column in the corresponding tables) our existence proof together with the first approach failed in these examples.

Parallelogram Obstacles Located at the Boundary of the Strip

We start the presentation of our results with a series of obstacles consisting of similar parallelograms with 45° angles at the bottom, varying in their height and width (cf. Figure 8.5, Figure 8.6 and Section 8.3).

The first (and smallest) obstacle considered in this series consists of a parallelogram with height 0.0625 and width 0.375 at the bottom. Some approximate solutions for this domain are plotted in Figure 8.5.



(a) $Re = 1.0$

(b) $Re = 3.0$

Figure 8.5: Some approximate solutions for 45°-parallelogram obstacle 1

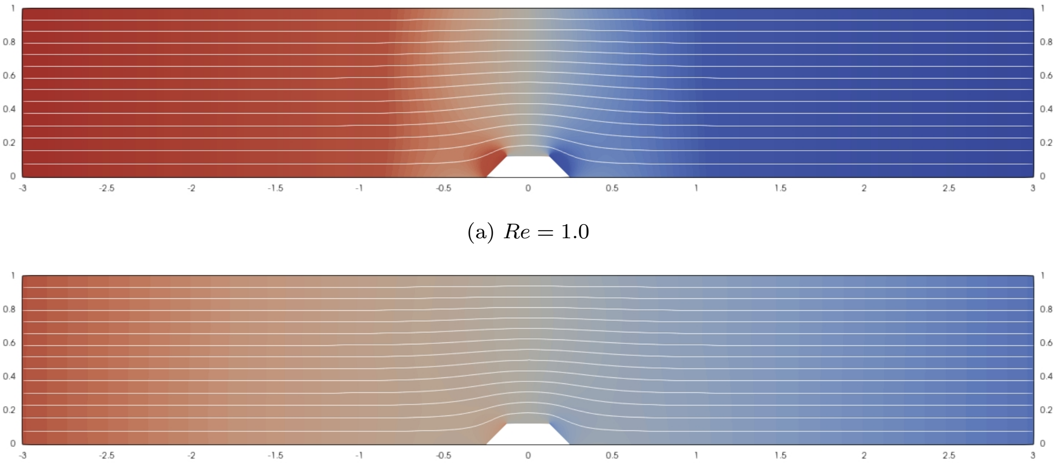
Similar to the example domain, for the entire series we fix the constants $d_0 := 2.5$, $d_1 := 0.5$, $d_2 := 0.5$ and $d_3 := 1.0$ (cf. Chapter 1 and Section 4.1). Again, we want to emphasize that each finite element mesh considered in our examples consists of triangles with corners which can be represented on the computer exactly. In particular, this is possible since the obstacles have 45° angles at the bottom and due to the fact that their corners are representable exactly on the computer.

Table 8.11 shows successful results for the smallest obstacle of the 45°-parallelogram type for different values of the Reynolds number. Similar to the first Section, we list the values for the defect bound δ, for the norm bounds K and K^* respectively, and for the error bound (measured in different norms), as well as for the radius of uniqueness.

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla\tilde{p}\|_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.00	0.04805920	0.0639661^{8}_{7}	0.05939963	0.01890749	0.23065094	Approach 1
		1.21587631		3.59465868	0.02818067		
1.5	0.00	0.04805956	0.1205806^{4}_{3}	0.06761023	0.02152101	0.21745318	Approach 1
		1.36303101		2.10543054	0.03207599		
2.0	0.00	0.04805993	0.2081070^{5}_{4}	0.07887099	0.02510542	0.22749530	Approach 1
		1.55074111		1.35363210	0.03741837		
2.5	0.00	0.04806037	0.3498807^{1}_{0}	0.09570313	0.03046326	0.25707329	Approach 1
		1.79845218		0.89245388	0.04540396		
3.0	0.00	0.04806088	0.5947035^{4}_{3}	0.12570938	0.04001454	0.32115032	Approach 1
		2.14040637		0.56619705	0.05963967		
3.5	0.50	0.04806939	0.4570278^{9}_{8}	0.09849545	0.03058688	0.23797859	Approach 2 (a)
		1.77944256		0.65013906	0.04561475		

Table 8.11: Results for 45°-parallelogram obstacle 1

Next, we consider a slightly larger obstacle (again a 45°-parallelogram) which now has height 0.125 and width 0.5 at the bottom. Again, we plotted some of the approximate solutions for this obstacle (cf. Figure 8.6).



(a) $Re = 1.0$

(b) $Re = 3.0$

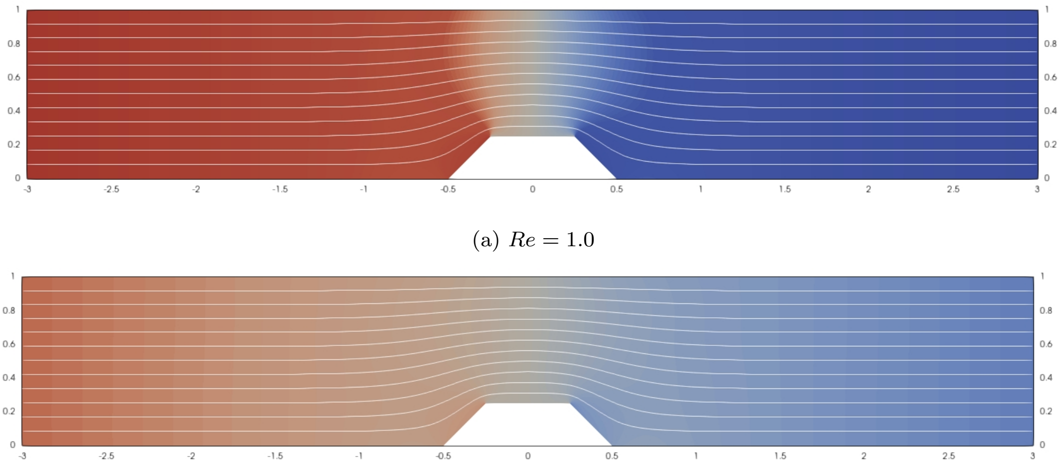
Figure 8.6: Some approximate solutions for 45°-parallelogram obstacle 2

Similar to the previous example, in the following Table we list the parameter settings where we successfully applied Theorem 3.4 in this second parallelogram domain.

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\lVert u^* - \tilde{\omega} \rVert_{H_0^1}$	$\lVert u^* - \tilde{\omega} \rVert_{L^2}$	$\lVert \nabla p^* - \nabla \tilde{p} \rVert_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\lVert u^* - \tilde{\omega} \rVert_{L^4}$		
1.0	0.00	0.04918807	0.066232^{40}_{39}	0.06118489	0.01947576	0.27538839	Approach 1
		1.22294751		3.57174530	0.02902764		
1.5	0.00	0.04918881	0.1258613^{4}_{3}	0.06998348	0.02227644	0.26841168	Approach 1
		1.37647985		2.08182567	0.03320191		
2.0	0.00	0.04918981	0.2194891^{1}_{0}	0.08222500	0.02617303	0.28821090	Approach 1
		1.57418723		1.32894223	0.03900960		
2.5	0.00	0.04919046	0.3741688^{8}_{7}	0.10097598	0.03214166	0.33410639	Approach 1
		1.83834010		0.86574021	0.04790553		
3.0	0.00	0.04919146	0.6484509^{2}_{1}	0.13644650	0.04343227	0.43379703	Approach 1
		2.20920253		0.53391347	0.06473363		
3.5	0.25	0.04919677	0.6364660^{7}_{6}	0.12592249	0.03958417	0.38512889	Approach 2 (a)
		2.05141426		0.50834825	0.05900724		

Table 8.12: Results for 45°-parallelogram obstacle 2

Finally, we close this series of 45°-parallelograms with the largest one which has height 0.25 and width 1.0 at the bottom. In Section 8.3 we present some approximate solutions for this type of obstacle.



(a) $Re = 1.0$

(b) $Re = 3.0$

Figure 8.7: Some approximate solutions for 45°-parallelogram obstacle 3

We note that the previous definitions of the constants d_0, d_1, d_2 and d_3 are appropriate in this setting as well.

Figure 8.8 shows the Euclidean norm of the approximate solution $U + \omega$ for $Re = 3.0$. The red color located in the center of the strip above the obstacle indicates that the speed of the fluid increases around the obstacle which coincides with the physical intuition. Moreover, we see that the fluid slows down in the corners to the right and left of the obstacle.

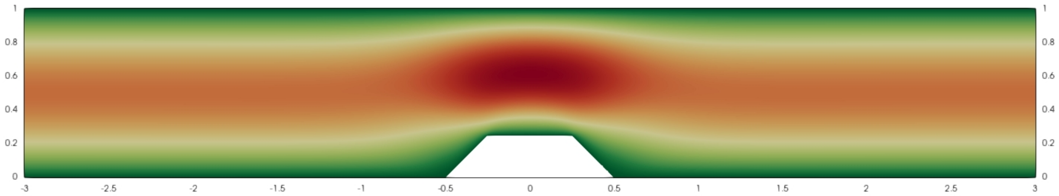


Figure 8.8: Euclidean norm of $U + \omega$ for 45°-parallelogram obstacle 3 and $Re = 3.0$

Again, we successfully applied Theorem 3.4 to a certain set of parameters which is presented in Table 8.13. Moreover, we want to point out that our second approach outscored the first one in this setting as well.

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla \tilde{p}\|_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.00	0.05471468	0.0802586^{8}_{7}	0.07129991	0.02269547	0.39771650	Approach 1
		1.27642863		3.40941402	0.03382646		
1.5	0.00	0.05471526	0.1621128^{2}_{1}	0.08462505	0.02693699	0.41569743	Approach 1
		1.48119165		1.91506348	0.04014824		
2.0	0.00	0.05471619	0.3066617^{7}_{6}	0.10534673	0.03353291	0.48278656	Approach 1
		1.76424680		1.15379766	0.04997913		
2.5	0.00	0.05471683	0.5858835^{5}_{4}	0.14522917	0.04622788	0.63723722	Approach 1
		2.18111035		0.66956368	0.06890035		
3.0	0.25	0.05471619	0.5374050^{7}_{6}	0.12575425	0.03953129	0.53116037	Approach 2 (a)
		1.93053462		0.66056196	0.05892840		

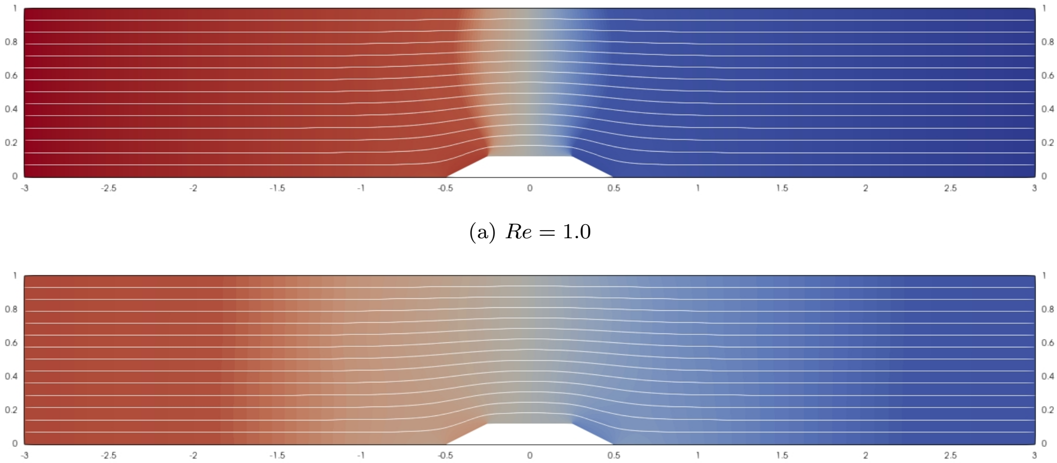
Table 8.13: Results for 45°-parallelogram obstacle 3

Moreover, using Theorem 8.3 on this domain we proved the existence of solutions to our Navier-Stokes equations for Reynolds numbers in the compact interval [1.0, 2.0]. The corresponding results are presented in Table 8.14.

Re	$\boldsymbol{\delta}$	$\boldsymbol{K}$	$4\boldsymbol{K}^2{C_4}^2 Re\,\boldsymbol{\delta}$	$\|u^*_{Re} - \tilde{\omega}\|_{H_0^1}$	$\|u^*_{Re} - \tilde{\omega}\|_{L^2}$	$\|u^*_{Re} - \tilde{\omega}\|_{L^4}$	$\|\nabla p^* - \nabla \tilde{p}\|_{H^{-1}}$
[1.000, 1.125]	0.05603943	1.32211234	$0.0^{9921516}_{8819125}$	0.07602539	0.02419964	0.03606835	2.21682022
[1.125, 1.250]	0.05603965	1.37119606	$0.1^{1857678}_{0671909}$	0.07926522	0.02523091	0.03760541	1.70923878
[1.250, 1.375]	0.05603983	1.42406659	$0.1^{4068740}_{2789763}$	0.08282804	0.02636499	0.03929570	1.37505353
[1.375, 1.500]	0.05604031	1.48117976	$0.1^{6603605}_{5219970}$	0.08677092	0.02762004	0.04116629	1.15164203
[1.500, 1.625]	0.05604022	1.54306758	$0.1^{9521726}_{8020054}$	0.09116440	0.02901853	0.04325067	1.00257891
[1.625, 1.750]	0.09197625	1.61035488	$0.3^{7579591}_{4895333}$	0.16548486	0.05267547	0.07851016	1.22605028
[1.750, 1.875]	0.05604051	1.68378047	$0.2^{6820655}_{5032610}$	0.10171112	0.03237566	0.04825430	0.83281779
[1.875, 2.000]	0.05604076	1.76422445	$0.^{31407744}_{29444759}$	0.10815909	0.03442811	0.05131338	0.79013492

Table 8.14: Continuous results for 45°-parallelogram obstacle 3

Next, we consider a parallelogram obstacle again but now with a flatter ramp to the left and right, i.e., the angle at the bottom is smaller than 45°. In particular, this parallelogram obstacle has height 0.125 and width 1.0 at the bottom. In this setting we compute several approximate solutions to our Navier-Stokes equations, where some of them are plotted in Figure 8.9.



(a) $Re = 1.0$

(b) $Re = 3.0$

Figure 8.9: Some approximate solutions for flatter parallelogram obstacle

Moreover, in Figure 8.10 we present the Euclidean norm of the velocity field.

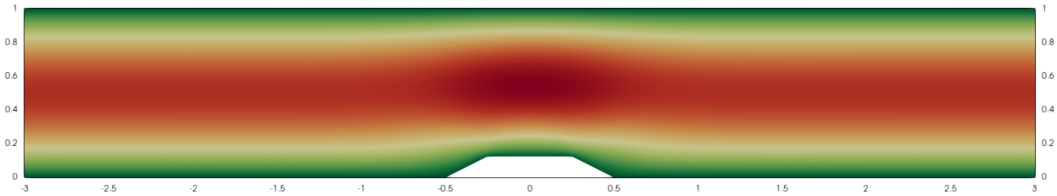


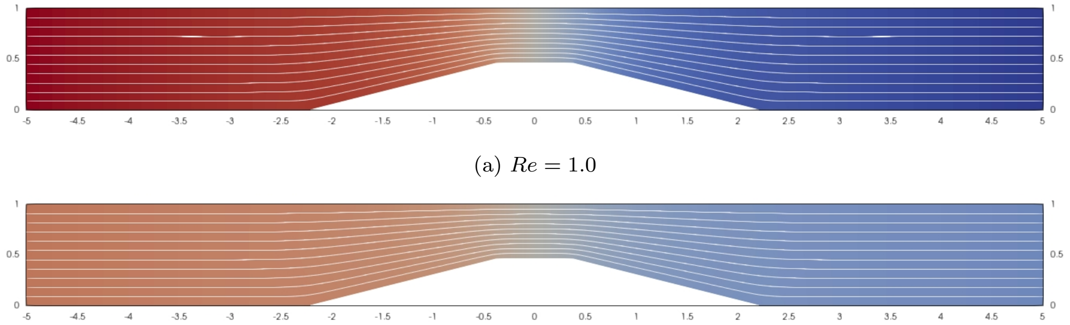
Figure 8.10: Euclidean norm of $U + \omega$ for flatter parallelogram obstacle and $Re = 3.0$

On the basis of our approximate solutions the application of Theorem 3.4 yields the existence of a solution to our Navier-Stokes equations in the parameter settings listed in Table 8.15.

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla\tilde{p}\|_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.00	0.04712400	0.063459^{20}_{19}	0.05857750	0.01864580	0.21033806	Approach 1
		1.22300621		3.57417834	0.02779063		
1.5	0.00	0.04712420	0.1205811^{2}_{1}	0.06694921	0.02131060	0.19575921	Approach 1
		1.37649461		2.08483688	0.03176238		
2.0	0.00	0.04712427	0.2102338^{2}_{1}	0.07854723	0.02500236	0.20360045	Approach 1
		1.57404235		1.33274990	0.03726477		
2.5	0.00	0.04712488	0.3582460^{7}_{6}	0.09617043	0.03061200	0.23021682	Approach 1
		1.83779906		0.87083036	0.04562566		
3.0	0.00	0.04712474	0.6203904^{5}_{4}	0.12875200	0.04098304	0.29128628	Approach 1
		2.20775007		0.54204899	0.06108317		

Table 8.15: Results for flatter parallelogram obstacle

Our next results are obtained on a strip perturbed by a large parallelogram. In this setting we consider a parallelogram of height 0.46875, width 4.5 at the bottom and width 0.75 at the top. Some approximate solutions are printed in Figure 8.11. We note that for the computation of approximate solutions, compared to the previous examples, in this setting we use a larger computational domain of width 10.0.



(a) $Re = 1.0$

(b) $Re = 2.0$

Figure 8.11: Some approximate solutions for larger parallelogram obstacle

Similar to the previous results, in Figure 8.12 we present the Euclidean norm of the velocity field, which again shows that the speed of the fluid increases in the narrowing caused by the obstacle.

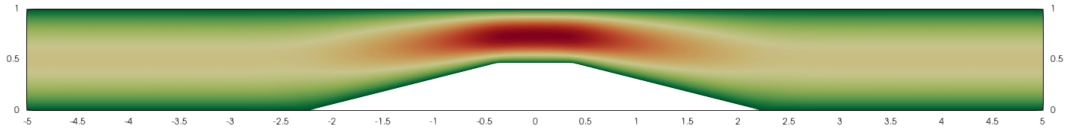


Figure 8.12: Euclidean norm of $U + \omega$ for larger parallelogram obstacle and $Re = 2.0$

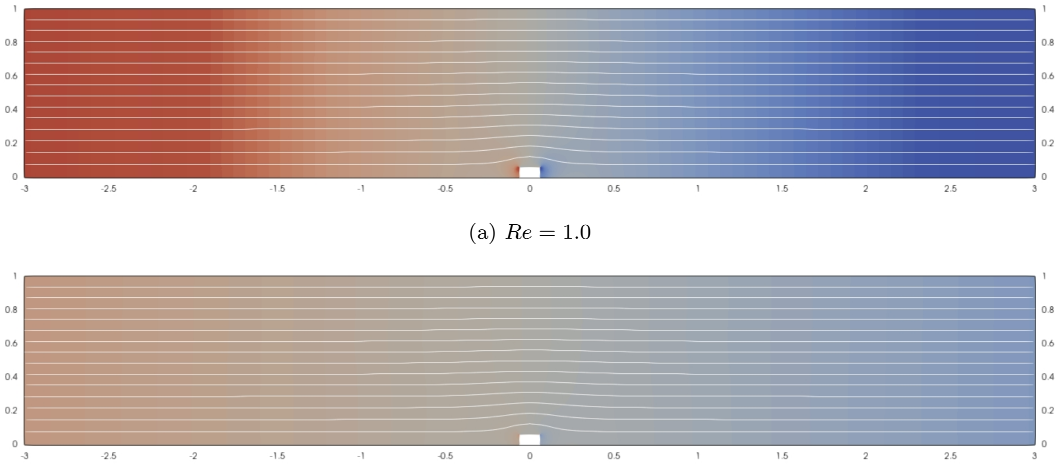
Using the approximate solutions together with Theorem 3.4 we proved the existence of exact solutions to or Navier-Stokes equations in several parameter sets where the results are listed in Table 8.16.

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\lVert u^* - \tilde{\omega} \rVert_{H_0^1}$	$\lVert u^* - \tilde{\omega} \rVert_{L^2}$	$\lVert \nabla p^* - \nabla \tilde{p} \rVert_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\lVert u^* - \tilde{\omega} \rVert_{L^4}$		
1.0	0.00	0.12425770	0.229716_{09}^{10}	0.18965892	0.06037031	0.72534203	Approach 1
		1.43296802		2.91081707	0.08997894		
1.5	1.00	0.12430478	0.395236_{59}^{60}	0.22472087	0.06816110	0.69129771	Approach 2 (a)
		1.60685102		1.79674422	0.10180682		
2.0	1.00	0.12430851	0.5778269_{0}^{1}	0.25356139	0.07690884	0.67713257	Approach 2 (a)
		1.68255797		1.19432033	0.11487263		
2.5	0.50	0.12429032	0.866366_{59}^{60}	0.32817082	0.10191049	0.86600720	Approach 2 (a)
		1.80278167		0.70634878	0.15198091		

Table 8.16: Results for larger parallelogram obstacle

Rectangular Obstacles Located at the Boundary of the Strip

In the further course we investigate a series of rectangular obstacles. Due to the large reentrant corners the numerical approximation procedure in this setting becomes more challenging compared to the previous examples. We start our considerations with small obstacle of height 0.0625 and width 0.125. Figure 8.13 shows some approximate solutions for this small rectangular obstacle.



(a) $Re = 1.0$

(b) $Re = 3.0$

Figure 8.13: Some approximate solutions for rectangular obstacle 1

In this setting, for the computations of our approximate solutions we choose the constants $d_0 := 2.5$, $d_1 := 0.5$, $d_2 := 0.5$ and $d_3 := 1.0$ (cf. Chapter 1 and Section 4.1) again.

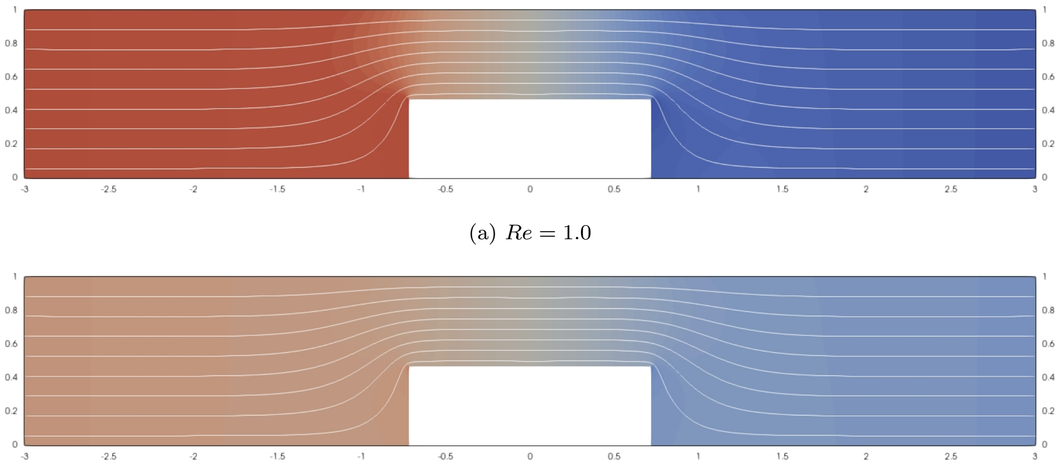
Applying Theorem 3.4 with the approximate solutions we proved the existence of exact solutions to our Navier-Stokes equations for the following parameter settings.

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla\tilde{p}\|_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.00	0.05699289	0.0785376^3_2	0.07195177	0.02290296	0.33751350	Approach 1
		1.23717467		3.51920076	0.03413572		
1.5	0.00	0.05699309	0.1516403^2_1	0.08328433	0.02651023	0.33501625	Approach 1
		1.40363139		2.02690071	0.03951217		
2.0	0.00	0.05699337	0.2699414^1_0	0.09969028	0.03173240	0.36792151	Approach 1
		1.62184763		1.27000771	0.04729556		
2.5	0.00	0.05699371	0.4730965^5_4	0.12683523	0.04037291	0.44303302	Approach 1
		1.92040964		0.79856792	0.06017381		
3.0	0.00	0.05699412	0.8528054^6_5	0.19390214	0.06172097	0.65193503	Approach 1
		2.35370375		0.43530233	0.09199202		

Table 8.17: Results for rectangular obstacle 1

The next obstacle of this series consists of a rectangle of height 0.46875 and width 1.4375. For this obstacle we changed the constants for the computation of V to $d_0 := 3.0$, $d_1 := 1.0$,

$d_2 := 0.5$ and $d_3 := 1.0$. Figure 8.14 shows some selected approximate solution for this domain.



(a) $Re = 1.0$

(b) $Re = 2.0$

Figure 8.14: Some approximate solutions for rectangular obstacle 2

Similar to the previous examples, Figure 8.15 shows the Euclidean norm of $U + \omega$ in the case $Re = 2.0$.

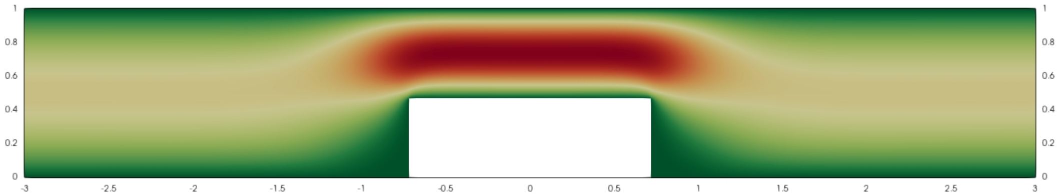


Figure 8.15: Euclidean norm of $U + \omega$ for rectangular obstacle 2 and $Re = 2.0$

Again, the computer-assisted techniques presented in this thesis yield the existence of a solution to our Navier-Stokes equations for this type of obstacle (cf. Table 8.18)

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla \tilde{p}\|_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.00	0.08451542	0.1632830_1^2	0.12931983	0.04116378	2.38358798	Approach 1
		1.46488983		2.90359284	0.06135256		
1.5	0.00	0.08452554	0.4160260_0^1	0.18293536	0.05823014	3.25545958	Approach 1
		1.90907544		1.36856013	0.08678911		
2.0	0.50	0.08453668	0.4952914_4^5	0.18266024	0.05672349	3.06620833	Approach 2 (a)
		1.84788008		1.07892931	0.08459274		

Table 8.18: Results for rectangular obstacle 2

In Figure 8.16 we print the solution for $Re = 2.0$ again with more stream lines. Especially, at each of the corners (at the bottom) on the left and right of the rectangular obstacle a vortex can be seen.

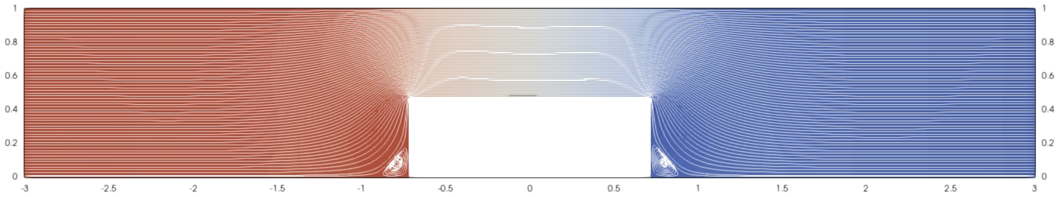
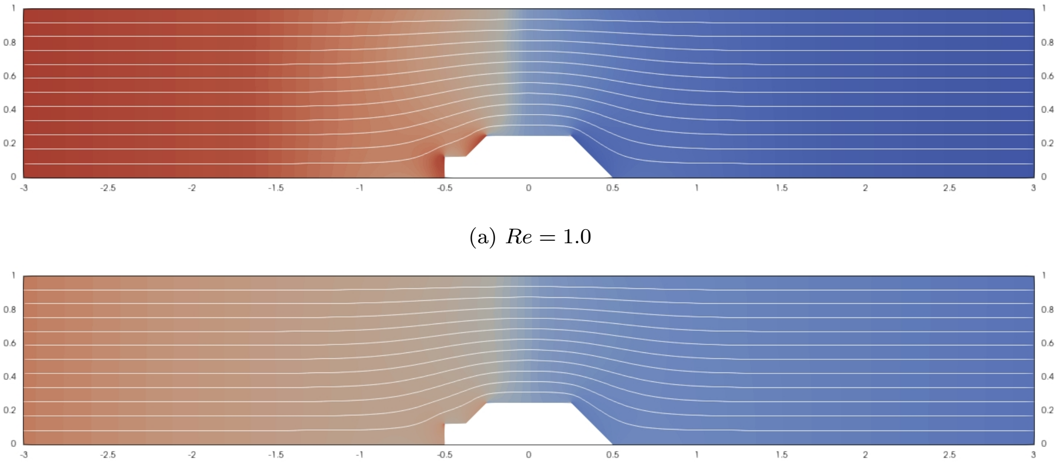


Figure 8.16: Approximate solution for rectangular obstacle 2 and $Re = 2.0$

Non-symmetric Obstacles Located at the Boundary of the Strip

In contrast to all previous examples, we now consider a non-symmetric obstacle located at the boundary of the strip.



(a) $Re = 1.0$

(b) $Re = 2.0$

Figure 8.17: Some approximate solutions for non-symmetric obstacle 1

To get an impression of the velocity field, in Figure 8.18 we plot the Euclidean norm of the approximate solution $U + \omega$ in the case $Re = 2.0$.

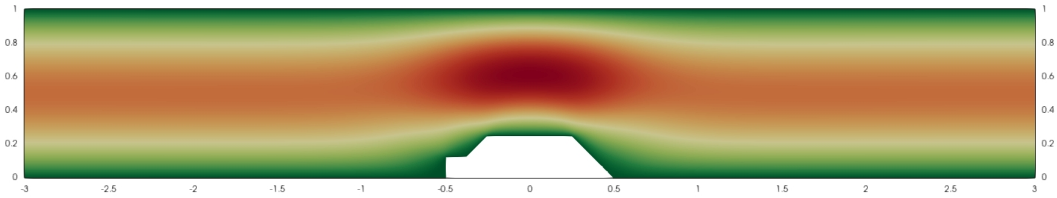


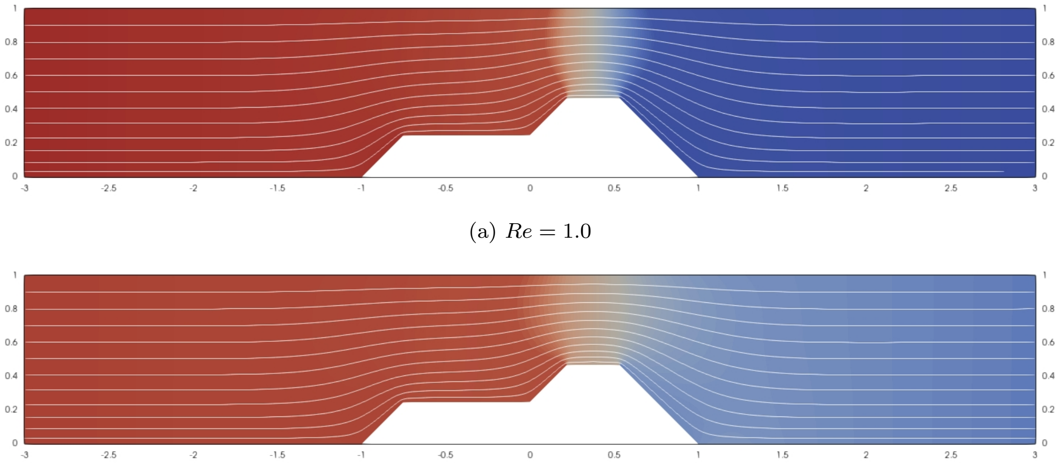
Figure 8.18: Euclidean norm of $U + \omega$ for non-symmetric obstacle 1 and $Re = 2.0$

We want to point out that our methods are applicable in the case of non-symmetric obstacles as well. In the present case, Theorem 3.4 provides the existence of an exact solution for several sets of parameters. The results are listed in Table 8.19.

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla\tilde{p}\|_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.00	0.06289941	0.0922482^{5}_{4}	0.08222160	0.02617195	0.52035283	Approach 1
		1.27631604		3.39879942	0.03900798		
1.5	0.00	0.06290065	0.1862825^{2}_{1}	0.09794348	0.03117638	0.55486615	Approach 1
		1.48086446		1.90218687	0.04646684		
2.0	0.00	0.06290399	0.3522285^{3}_{2}	0.12292196	0.03912728	0.65615897	Approach 1
		1.76343980		1.13679864	0.05831725		
2.5	0.00	0.06290061	0.6723382^{6}_{5}	0.17434757	0.05549656	0.89702908	Approach 1
		2.17920930		0.64115608	0.08271485		

Table 8.19: Results for non-symmetric obstacle 1

Next, we consider a second non-symmetric obstacle and plot some selected approximate solutions for this type of obstacle (cf. Figure 8.19).



(a) $Re = 1.0$

(b) $Re = 2.0$

Figure 8.19: Some approximate solutions for non-symmetric obstacle 2

In the following the Euclidean norm of the approximate solution $U + \omega$ in the case $Re = 2.0$ is visualized.

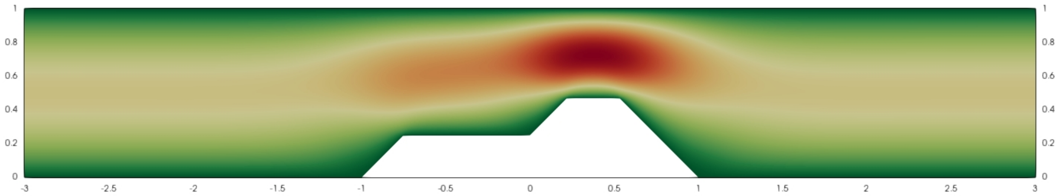


Figure 8.20: Euclidean norm of $U + \omega$ for non-symmetric obstacle 2 and $Re = 2.0$

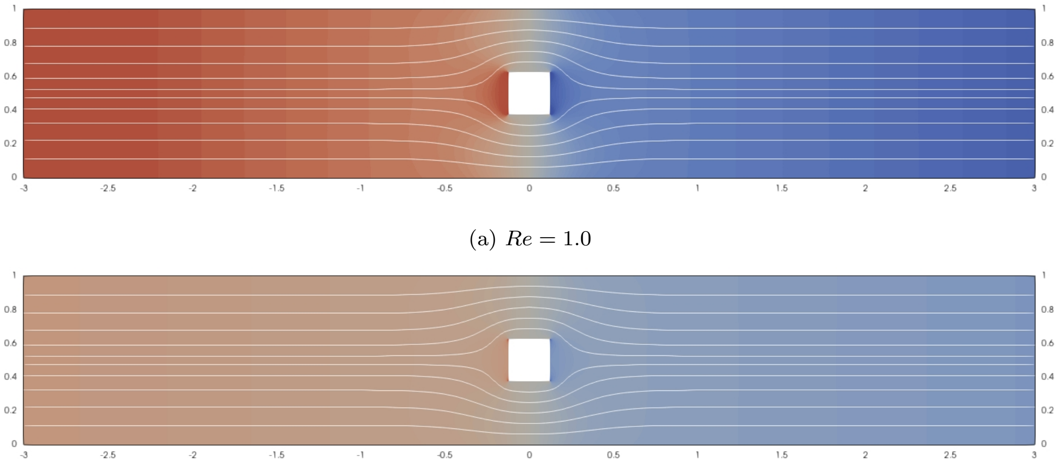
Applying our computer-assisted techniques we obtain the existence of an exact solution for different values of the Reynolds number (cf Table 8.20).

Re	σ	δ K	$4K^2{C_4}^2 Re\,\delta$	$\Vert u^* - \tilde{\omega} \Vert_{H_0^1}$ $\alpha_{\max}$	$\Vert u^* - \tilde{\omega} \Vert_{L^2}$ $\Vert u^* - \tilde{\omega} \Vert_{L^4}$	$\Vert \nabla p^* - \nabla \tilde{p} \Vert_{H^{-1}}$	Approach
1.0	0.00	0.07780904 1.44994737	0.147275_{19}^{20}	0.11731017 2.94685825	0.03734099 0.05565488	1.00082994	Approach 1
1.5	0.00	0.07781151 1.87066029	0.3677223_{7}^{8}	0.16216826 1.42118813	0.05161976 0.07693668	1.27269739	Approach 1
2.0	0.00	0.07782548 2.63507766	0.9730468_{3}^{4}	0.35231185 0.49071509	0.11214435 0.16714555	2.64285233	Approach 1

Table 8.20: Results for non-symmetric obstacle 2

Rectangular Obstacle Detached from the Boundary of the Strip

Finally, we consider an obstacle which is detached from the boundary of the strip. We want to emphasize that in this setting our domain is not simply connected. As already mentioned earlier, our methods can also treat such obstacles as well (Recall that our computer-assisted approach does not require a stream function formulation of the Navier-Stokes equations). Figure 8.21 shows some approximate solutions computed on the strip perturbed by a square of length 0.5 in the center.



(a) $Re = 1.0$

(b) $Re = 2.0$

Figure 8.21: Some approximate solutions for centered square obstacle

Moreover, for this type of obstacle, in Figure 8.22 we present the Euclidean norm of the approximate solution $U + \omega$ for $Re = 2.0$.

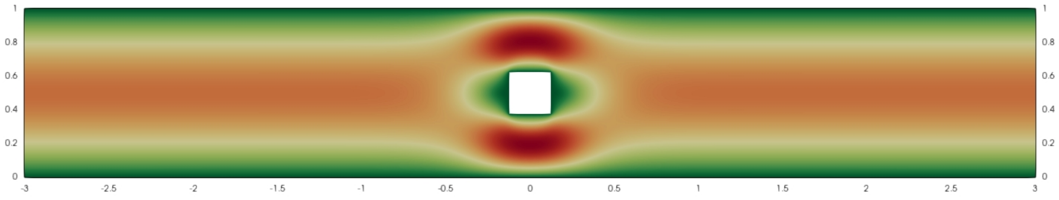


Figure 8.22: Euclidean norm of $U + \omega$ for centered square obstacle and $Re = 2.0$

Having a closer look at Figure 8.22 again, we see that the flow actually splits up at the obstacle into a lower and upper part. From the red color between the obstacle and the boundary we conclude that the speed of the flow increases near the obstacle which again fits into our physical intuition.

As in the previous examples, in this setting our existence Theorem 3.4 applied to our approximate solutions provides the existence of an exact solution for several parameter settings (cf. Table 8.21).

Re	σ	δ	$4K^2{C_4}^2 Re\,\delta$	$\|u^* - \tilde{\omega}\|_{H_0^1}$	$\|u^* - \tilde{\omega}\|_{L^2}$	$\|\nabla p^* - \nabla\tilde{p}\|_{H^{-1}}$	Approach
		K		$\alpha_{\max}$	$\|u^* - \tilde{\omega}\|_{L^4}$		
1.0	0.00	0.10202764	0.1583439^6_5	0.13972537	0.04447597	2.28109593	Approach 1
		1.31293815		3.24419864	0.06628921		
1.5	0.00	0.10202473	0.3338582^7_6	0.17488916	0.05566895	2.74146059	Approach 1
		1.55662915		1.72789030	0.08297179		
2.0	0.00	0.10204330	0.6713998^6_5	0.24797407	0.07893260	3.80963606	Approach 1
		1.91155077		0.91414074	0.11764510		

Table 8.21: Results for centered square obstacle

Finally, we want to mention that in this setting (since the obstacle is symmetric and located in the center of the strip) one might exploit the symmetry $u(x, y) = u(x, 1 - y)$ for all $(x, y) \in \Omega$ to reduce the computational effort. Especially, in the context of our eigenvalue computations using symmetric sub spaces could result in less eigenvalues of the base problem. However, we do not make use of this symmetry since we do not want to exclude symmetry breaking solutions a piori. However, later we did not find any symmetry breaking solution.

8.4 Conclusion

In this thesis we introduced an appropriate computer-assisted setting to prove the existence of exact solutions to the Navier-Stokes equations

$$\left.\begin{aligned} -\Delta u + Re\,[(u \cdot \nabla)u + (u \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u + \nabla p] &= g \\ \operatorname{div} u &= 0 \end{aligned}\right\} \text{ in } \Omega$$
$$u = 0 \quad \text{on } \partial\Omega$$

in a weak setting. Especially, the weak formulation

$$\begin{aligned} &\text{Find } u \in H(\Omega) \text{ such that} \\ &\int_\Omega (\nabla u \bullet \nabla\varphi + Re\,[(u \cdot \nabla)u + (u \cdot \nabla)\Gamma + (\Gamma \cdot \nabla)u] \cdot \varphi)\,\mathrm{d}(x, y) \\ &\qquad = \int_\Omega g \cdot \varphi\,\mathrm{d}(x, y) \quad \text{for all } \varphi \in H(\Omega) \end{aligned}$$

on the divergence-free subspace $H(\Omega) \subseteq H_0^1(\Omega, \mathbb{R}^2)$ played a crucial role in our computer-assisted proof.

At this stage, we want to emphasize again that for the Navier-Stokes equations the linearized operator is not self-adjoint and therefore, the established computer-assisted techniques providing the surjectivity of $\Phi^{-1}\,\mathrm{L}$ are not applicable in this setting. Nevertheless, in this thesis we showed that computer-assisted techniques can successfully be applied to the Navier-Stokes equations as well.

Therefore, in Chapter 6 we suggested two approaches for the computation of the crucial norm bounds K and K^* respectively. As already expected in advance our applications showed that by construction our first approach fails for higher Reynolds numbers. To overcome this disadvantage and to make our computer-assisted techniques applicable also for higher Reynolds numbers we introduced our second approach based on eigenvalue bounds for suitable eigenvalue problems and a homotopy method (cf. Section 6.2.1).

In this setting, we introduced a straightforward eigenvalue homotopy which opens up the opportunity to deal with larger Reynolds numbers as well. We want to point out that using this approach massively increases the computational effort (and time) for the computation of the desired norm bounds compared to the first approach. However, applying this approach we were able to prove the existence of exact solutions to our Navier-Stokes equations for parameter settings where the application of the first approach was unsuccessful (cf. for instance Table 8.16). Nevertheless, by construction also this approach is expected to fail for higher Reynolds numbers (cf. definition of γ_1 in this homotopy setting).

Therefore, we finally introduced a more complex homotopy method which has no theoretical restriction to small Reynolds numbers. However, our examples showed that due to the complex homotopy structure many eigenvalues have to be considered in the homotopy methods and thus, the computational effort compared to the straightforward approach again increases. Nevertheless, we want to emphasize that using the computer-assisted methods presented in this thesis together with our second approach and the extended homotopy method, at least theoretically allow us to prove the existence of a solution of the Navier-Stokes equations for arbitrarily high Reynolds numbers provided it exists and enough computational power is available.

Concerning the computational effort we note that the computation of an approximate solution to our Navier-Stokes equations is possible without any problems also for higher Reynolds numbers. The problems appear in the computations of the homotopy method, especially, in the computations of accurate functions w_i, which are required for the Lehmann Goerisch method, we were running out of memory (cf. remarks on p. 140). Hence, it makes sense to use the first approach whenever it is possible. If the first approach fails one tries the second approach together with the straightforward homotopy method and finally, in the non-affirmative case the second approach with the extended homotopy method has to be applied.

8.5 Future Projects and Outlook

In the following, we would like to give some concluding remarks and an outlook on several interesting questions concerning computer-assisted techniques for Navier-Stokes equations.

Our applications showed that most of the eigenvalues which have to be considered in our eigenvalue homotopy emerge in the (final) constraint homotopy (cf. Section 6.2.1.2). In this context one can think of considering the base problem on the space $H(S)$ instead of $H_0^1(S, \mathbb{R}^2)$, i.e., using the eigenvalue problem (6.61) as our final base problem. However, solving this eigenvalue problem rigorously on the space $H(S)$ requires new ideas to enclose all eigenvalue of the problem below the prescribed bound ρ_0. If such enclosures are in hand, our constraint homotopy is obsolete and we expect that the number of eigenvalues, which have to be considered in the remaining homotopies, is much smaller compared to the situation considered in our examples.

Concerning the increasing memory usage for the computation of the homotopy method, especially, for the computation of the functions w_i needed for the Lehmann Goerisch computations, one can think of using symmetric matrices (and `MatrixGraph`) in M++ to reduce the memory usage required in the approximation procedure. However, to exploit this symmetry major changes in basic routines of M++ (like solvers, matrix access, et cetera) are required.

Moreover, in the case of a symmetric obstacle (with respect to the y-axis) we can use this symmetry in the definitions of each of our spaces. Thus, we expect the number of eigenvalues of the base problems to be smaller than in case without using the symmetry. Additionally, in this setting it seems to make sense to use half of the strip (or half of the computational domain, respectively) for our numerical computations. Therefore, suitable boundary conditions have to be imposed on $(\{0\} \times [0,1]) \cap \overline{\Omega}$.

Since in our second approach together with the extended homotopy method the number of eigenvalues which have to be considered in the homotopy method depends on the values γ_1 and γ_2 (see (6.54) and (6.55)) we are interested in a constant $\kappa > 0$ (satisfying (6.82)) as large as possible. Therefore, the estimate $\langle \Phi^{-1}(\mathrm{B}_U - \sigma)u, \Phi^{-1}(\mathrm{B}_U - \sigma)u \rangle_{H_0^1(\Omega,\mathbb{R}^2)} \geq 0$ (cf. p. 110) might be too “rough”. In this context, we tried to improve this estimate by using bilinear forms with appropriate constraint conditions which however did not lead to the desired success.

Finally, we would like to emphasize that the methods presented in this thesis also apply to the 3-dimensional case, especially, Theorem 3.4 remains valid for the $3D$ case. However, compared to the 2-dimensional case at several stages adaptions are necessary. For instance, the definition of the function V introduced to obtain Dirichlet boundary conditions on the complete boundary (cf. Chapter 1) needs to be adapted properly to transform the Navier-Stokes equations in the $3D$ case correctly. Moreover, the divergence-free finite element using Argyris elements introduced in Section 4.2 is not applicable in the 3-dimensional case. Therefore, for the computation of the required approximate solution $\tilde{\omega} \in H(\Omega)$ we have to use different divergence-free finite elements like the Scott-Vogelius finite element (cf. [17] and [37]) which are not part of the M++ finite element software package yet. However, several other finite element software packages, for instance in the Verified Finite Element Method (VFEM) libary (written in MATLAB language) by Liu (see [60]), already provide the Scott-Vogelius finite element.

9 Extensions for the FEM-Software M++ (Meshes, Multigrid and More)

The programs realizing the numerical parts needed for our computer-assisted proof described in the previous Sections are implemented in C++. For the computation of the approximate solution (cf. Section 4.2) and other finite element functions, we use the Finite Element Software M++ (Meshes, Multigrid and More) developed by Wieners and his group (see [113]). The software M++ is written in C++ and uses the MPI (Message Passing Interface) which allows parallel computations to distribute computational load and to reduce computational time. Furthermore, the basic routines of M++ provide several solvers for linear problems, a realization of the Newton method described in Section 4.2 as well as iterative solvers for eigenvalue problems which can be used to compute approximate eigenpairs. The latest version of the Finite Element Software M++, a tutorial project as well as several applications using M++ can be downloaded form gitLab (see [114]). For all our computations we used the Release 2.6.1 of M++.

To treat all our numerical computations appearing in our computer-assisted proof, especially the interval arithmetic parts, the software M++ needs to be extended by several new classes and routines. We add an efficient implementation of the Argyris element for the approximation procedure as well as the corresponding shape functions (cf. Section 4.2 and Section 9.4.1). For the interval arithmetic basic data types we use the external library C-XSC (see [38] and [43]) and wrote wrapper classes to use the interval arithmetic data types of C-XSC in M++. Additionally, the new wrapper classes open up the opportunity to easily change the basic interval arithmetic types by multiple-precision types if necessary.

Furthermore, we use modern C++ standard techniques to extend several established M++ classes to improve their efficiency. Additionally, we exploit templates to implement interval arithmetic versions of the basic M++ classes like `Quadrature`, `Transformation`, `Shape`, `Discretization`, `Element` and many more. Hence, these established classes can now be used with interval arithmetic operations just by changing the corresponding template parameter.

Finally, we added several routines for the computation and verification of eigenvalues as described in Section 6.2.1 like the Rayleigh-Ritz method and the Lehmann-Goerisch method. Moreover, we added a semi-automatic homotopy procedure which performs the computations needed for the homotopy method described in Section 6.2.1. Especially, the "next" homotopy parameter t_i is computed automatically in this homotopy process. We want to emphasize that our implementation of the eigenvalue methods is based on an assemble class which can easily be adapted to many other problems which guarantees a flexible use of the new features.

Throughout this chapter we give an overview about the developments concerning the established classes of M++ and describe the main ideas of the new routines, especially their implementations. All addressed changes are already part of the latest release of the Finite Element Software M++ to be found on gitLab (see [110]).

9.1 New Interval Arithmetic Classes

As already mentioned, at several stages in this thesis verified computations are required (cf. Section 3.3). Therefore, we implemented a wrapper class that provides the basic data types `IAInterval` and `IACInterval`, the basic interval arithmetic operations ($\oplus, \ominus, \odot, \oslash$) as well as several standard functions (cf. Section 3.3). In principle each external interval arithmetic library can now be used with M++ if a proper implementation of the wrapper class is written. Moreover, each interval arithmetic class contained in M++ can easily be identified by its prefix `IA`.

At the moment, the interval arithmetic library C-XSC (see [43]) is integrated into M++. Since the latest release of C-XSC is written using the C++14 standard (the latest release is form 2014) we had to adapt the code at several points to make C-XSC run with a today's C++ compiler that uses the C++17 standard (or newer). However, using our wrapper class we were successfully able to integrate the C-XSC library to M++.

Moreover, we integrated the interval arithmetic library MPFI (based on the MPFR library [21]) by Nathalie Revol and Fabrice Rouillier which supports the latest C++ standard (see [87]). MPFI is a multi-precision interval arithmetic library that allows interval arithmetic computations with arbitrary large mantissa. However, the usage of MPFI lead to a strong increase of computational time in all our examples.

9.2 Developments for the Established Classes of M++

Extensions Using Templates

Up to the present day, none of the former interval arithmetic extensions for M++ (see for instance [89]) were integrated into its source files. For successful interval arithmetic computations with M++ several classes, starting from `Point`, `VectorField`, `Tensor` over `Transformation`, `Quadrature`, `Discretization` and the shape functions up to the elements (to name just a few of them), need their interval arithmetic counterpart. Since the modern C++ standard provides templates, we restructured the existing classes by adding a template parameter that play the role of the underlying data type, i.e., a class can either be compiled with the type `double` or with one of the interval arithmetic data types `IAInterval` and `IACInterval`. This approach combines the most possible flexibility on the one hand and a very efficient and compact implementation of the new interval arithmetic classes on the other hand.

Nevertheless, since some of the implementations strongly differ depending on the template parameter we need template specialization for some classes. For instance, the implementation of the `Quadrature` class, which provides several quadrature rules for the different cell types and for multiple polynomial degrees, obviously depends on the used data type. While in the standard case (for `double`) approximate quadrature points and weights are sufficient for the computation of integrals, in the interval arithmetic case verified enclosures for the (exact) quadrature points and weights are required to compute integrals rigorously. In Section 9.3 we shortly describe the computation of verified interval arithmetic quadrature rules.

As mentioned in Chapter 8, in our examples we are restricted to meshes where the corners of the triangles are exactly representable on the computer. We want to point out that for the setup of the transformation $\Phi_{\mathcal{T}}$ (see Section 9.4.1) the corners or the corresponding cell need to known rigorously to guarantee the verified evaluation of the transformation $\Phi_{\mathcal{T}}$. Hence, the classes representing the different cell types and the mesh itself can also be extended by the additional template parameter to provide a verified cell refinement procedure. If this is the case, the (verified) methods developed in this thesis can be applied to meshes with arbitrary triangles.

A More Flexible Class `dof`

To determine the nodal points and the correct size of memory corresponding to a certain finite element discretization, M++ uses the interface `DoF`. In particular, a class inheriting from `DoF` implements functions for the computation of the nodal points and the number of degrees of freedom at each of these points.

For Lagrangian finite elements we introduced the flexible class `LagrangeDoF` which provides all nodal points and degrees of freedom depending on the cell type and the desired order of the polynomials. Thus, the new `LagrangeDoF` opens up the opportunity to use higher order Lagrangian finite elements assuming that appropriate shape functions are implemented.

Moreover, we extended the existing class `dof`, which could handle only a single `DoF` class so far, to an arbitrary number of `DoF` classes $D_1, \dots, D_N$. In particular, the new implementation of the class dof allows the combination of all existing `DoF` classes among themselves. For that purpose, the new class automatically combines all nodal points of the single classes $D_1, \dots, D_N$ (corresponding to shapes $S_1, \dots, S_N$) to a collection of nodal points. In this process it is checked whether a new nodal point needs to be added to the existing list or the nodal point already exists and thus, a doubling of nodal points is impossible. Additionally, the number of degrees of freedom at each nodal point is accumulated for all $D_1, \dots, D_N$. Besides the access to the complete collections of nodal points one can also access the list of nodal point of each class D_i for $i = 1, \dots, N$ individually. This enables the implementation of Taylor Hood elements of arbitrary order by adding two `LagrangeDoF` objects of appropriate order to the container `dof`.

A class of the type `Vector` contains the coefficients of a function lying in a finite element space, i.e., a `Vector` represents a finite element solution by storing its coefficients corresponding to the finite element basis. In this context, we also needed to adapt the access to the entries of a `Vector`. If multiple `DoF` classes are added to the container `dof`, the new classes `MixedRowValues`, `MixedRowBndValues` and `MixedRowEntries` have to be used in order to distinguish between the different shapes S_i for $i = 1, \dots, N$. For that purpose, an additional integer component in the case of `MixedRowValues` and `MixedRowBndValues` and two additional integer components in `MixedRowEntries` to identify the corresponding shape S_i have been added to all operators dealing with the access to the `Vector`. For an example of usage we refer to the tutorial project [111] on gitLab where the new features are used for some of the computations.

9.3 Quadrature Rules

Since the Argyris element uses polynomial shape functions with relatively high degree, quadrature rules of higher order are required to evaluate the occurring integrals at least accurately or even exactly (cf. Section 4.2 and explanations in 5). Hence, we extended the existing `Quadrature` class of M++ by several new quadrature rules for instance higher order quadrature rules on the reference interval $[0,1]$. Furthermore, we added a set of symmetric quadrature rules for triangles which allow the integration of polynomials up to a specific order with relatively low number of quadrature points. The quadrature points and corresponding weights for such symmetric quadrature rules on triangles are given for instance in [23].

Since we are interested in a rigorous analytical proof (cf. Theorem 3.4) several integrals have to be evaluated using verified quadrature rules, i.e., verified enclosures of the quadrature points and its corresponding weights have to be in hand explicitly. The computation of such enclosures is presented in the following Subsection.

Verified Quadrature Rules

In our examples, we only need verified quadrature rules on triangles, however, for the reader's convenience we begin our explanations in the one-dimensional case since the argumentation becomes more understandable and remains almost the same for the most parts or can be extended to the triangular case. Hence, we try to compute quadrature points $\hat{x}_1, \ldots, \hat{x}_d \in [-1,1]$ and their corresponding weights $w_1, \ldots, w_d$ (or at least verified enclosures of these values) such that $\int_{-1}^{1} p(x,y)\,\mathrm{d}x = \sum_{i=1}^{d} w_i p(\hat{x}_i)$ for all polynomials $p\colon [-1,1] \to \mathbb{R}$ up to a certain degree. It is well-known (cf. [84, Corollary 6.38]) that the Gaussian quadrature rule of order d integrates all monomials up do degree $2d-1$ exactly, i.e., the following equalities hold true:

$$\int_{-1}^{1} x^q \,\mathrm{d}(x,y) = \sum_{i=1}^{d} w_i \hat{x}_i^q \quad \text{for all } q = 0, \ldots, 2d-1. \tag{9.1}$$

Thus, to obtain the desired quadrature points and weights a non-linear system with $2d$ equations and $2d$ unknowns has to be solved rigorously. Since the computational effort for rigorously solving a non-linear system strongly increases with the number of unknowns, we are interested in reducing the size of our system. Therefore, we exploit the symmetry $x \mapsto -x$ to obtain a system which has almost half the size of the original system, more precisely, we obtain a non-linear system with $\left\lfloor \frac{d}{2} \right\rfloor$ equations as well as unknowns. Moreover, we have the relations $\hat{x}_i = -\hat{x}_{d-i}$ and $w_i = w_{d-i}$ for all $i = 1, \ldots, d$ for the quadrature points and their weights (Note that if d is odd we obtain $\hat{x}_{\left\lfloor \frac{d}{2} \right\rfloor} = 0$).

For solving the non-linear system we apply an algorithm provided by the toolbox of the interval library C-XSC which uses an interval Gauss-Seidel method and an interval Newton method to enclose the solutions of a non-linear system in a prescribed interval region. For more details about the implementation we refer the reader to [38, Chapter 13]. As initial intervals for our verification procedure we use the non-verified quadrature points and weights (which can be found in several numeric books, for instance [1, Table 25.4]) and inflate them by a “small” ε. Thus, we expect that our algorithm successfully encloses a solution of our non-linear system, if the approximate (and non-verified) quadrature points and weights are “sufficiently accurate”. We note that for our applications in M++ we need quadrature rules on the reference interval $[0, 1]$ and thus, we need to perform a final verified transformation of the quadrature points and weights from $[-1, 1]$ to $[0, 1]$.

Remark 9.1. *To obtain the desired quadrature points and weights in the one-dimensional case we actually do not need to solve a non-linear system. In this case, we can compute (or enclose) $\hat{x}_1, \ldots, \hat{x}_d$ as the roots of a Legendre polynomial of appropriate order (cf. [84, Theorem 6.35]. Inserting these roots into* (9.1) *yields a linear system for the quadrature weights $w_1, \ldots, w_d$. However, since this strategy is not applicable for the computation of quadrature rules for triangles we do not use this fact in our explanations here.*

For integration on triangles, the reduction of the size plays an even more crucial role for a successful verification procedure. To the end of this Section we will have a closer look at the verification procedure of quadrature points and their corresponding weights for an quadrature rule that integrates polynomials exact up to degree d on triangles. Especially, we are interested in quadrature rules on the reference triangle $\hat{\mathcal{T}}$ which first was defined in Section 4.2. Similar to the one-dimensional case considering a suitable transformed triangle and polar coordinates, we can reduce the size of our non-linear system. For detailed information about the reduction of the system we refer the reader to a paper by Dunavant (cf. [23]). Applying the methods described by Dunavant, we end up with a non-linear system which can be solved rigorously using the same algorithms as in the one-dimensional case.

After transforming the output of our algorithm back to the reference triangle, we obtain the enclosures $(\hat{X}_1, \hat{Y}_1), \ldots, (\hat{X}_{N_d}, \hat{Y}_{N_d})$ for the quadrature points $(\hat{x}_1, \hat{y}_1), \ldots, (\hat{x}_{N_d}, \hat{y}_{N_d})$ and enclosures $W_1 \ldots, W_{N_d}$ for the weights $w_1, \ldots, w_{N_d}$, i.e., the following enclosing result holds true for all $p \in \mathbb{P}^d(\hat{\mathcal{T}}) := \operatorname{span}\left\{x^i y^j : 0 \leq i, j,\ i + j \leq d\right\}$:

$$\int_{\hat{\mathcal{T}}} p(x, y)\, \mathrm{d}(x, y) \in \sum_{i=1}^{N_d} W_i \odot p(\hat{X}_i, \hat{Y}_i) \quad \text{for all } p \in \mathbb{P}^d(\hat{\mathcal{T}}),$$

where all operations (including the sum and the evaluation of the polynomial p) on the right-hand side have to be performed using interval arithmetic calculations.

As described before the successful enclosure of the quadrature points and their weights strongly depends on the size of the non-linear system which has to be solved. To compute the 73 quadrature points and weights corresponding to the quadrature rule which integrates polynomials exact up to degree 19, the non-linear system that has to be considered consists of 40 equations as well as unknowns. It turned out that the interval arithmetic with `double` precision is not sufficient to obtain the desired accuracy. Therefore, we extended the single-precision algorithms contained in C-XSC (cf. [38, Chapter 13]) to multiple-precision versions which finally yields enclosures with sufficient accuracy.

All programs used in our verification procedure including the extensions for the C-XSC toolbox are part of the source code of the Finite Element Software M++ (see gitLab [110]) and can easily be extended to higher order quadrature rules if necessary (cf. Dunavant [23]).

9.4 Implementation of Additional Finite Elements

In addition to the verified quadrature rules for our applications, we added several new discretizations and their corresponding finite elements to the M++. In the following we explain the implementation of the Argyris element which is used to construct an exactly divergence-free element (cf. 4.2). Afterwards, a possible construction of higher order Raviart Thomas elements for triangles based on a description by Ervin (cf. [25, Section 3.4]) is presented in Section 9.4.2. All elements introduced in the further course are already implemented in the latest version of M++ which can be found on gitLab (see [110]).

9.4.1 Argyris Element

In this Section we present a development of the well-known Argyris element based on ideas of Dominguez and Sayas described in [22]. This implementation uses the transformation $\Phi_{\mathcal{T}} \colon \hat{\mathcal{T}} \to \mathcal{T}$ (cf. Section 4.2) to "move" the evaluation of the shape functions to the reference cell $\hat{\mathcal{T}} \coloneqq \{(0,0), (1,0), (0,1)\}$. At this stage we note, that in the following in the context of finite elements for any object $\mathcal{O}$ (in the context of a triangle $\mathcal{T}$) we denote the corresponding object in the reference setting by $\hat{\mathcal{O}}$.

In the further course we will shortly describe some of the ideas by Dominguez and Sayas in detail to get an impression how our implementation in M++ is constructed. Therefore, we first fix an arbitrary triangle $\mathcal{T} = \operatorname{conv}\{(x_0,y_0),(x_1,y_1),(x_2,y_2)\}$ (where conv denotes the convex hull) contained in our finite element mesh $\mathcal{M}$. Then, the transformation $\Phi_{\mathcal{T}}$ (cf. Figure 4.7) is explicitly given by

$$\Phi_{\mathcal{T}}(\hat{x},\hat{y}) = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \begin{pmatrix} x_1 - x_0 & x_2 - x_0 \\ z_1 - z_0 & y_2 - y_0 \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}.$$

Since we assumed our cell $\mathcal{T}$ to be non-degenerate, the matrix in the definition of $\Phi_{\mathcal{T}}$ is clearly invertible and thus, the transformation is bijective (which was claimed in Section 4.2).

For the reference triangle $\hat{\mathcal{T}}$ we consider the functionals $\hat{\mathcal{L}}_1, \dots, \hat{\mathcal{L}}_{21} \colon \mathbb{P}^5(\hat{\mathcal{T}}) \to \mathbb{R}$ representing the degrees of freedom introduced in Section 4.2 for $\hat{\mathcal{T}}$ (cf. Figure 4.6). To obtain the desired local shape functions $\hat{\zeta}_1, \dots, \hat{\zeta}_{21} \in \mathbb{P}^5(\hat{\mathcal{T}})$ on the reference triangle we have to determine the dual basis corresponding to the degrees of freedom $\hat{\mathcal{L}}_1, \dots, \hat{\mathcal{L}}_{21}$, i.e., we have to compute reference shape functions $\hat{\zeta}_1, \dots, \hat{\zeta}_{21}$ such that $\hat{\mathcal{L}}_i(\hat{\zeta}_j) = \delta_{i,j}$ holds true for all $i, j = 1, \dots, 21$. These equations lead to a linear system which can easily be solved using a computer algebra software (cf. Appendix A.4). To the end of this Section we suppose that the reference shape functions $\hat{\zeta}_1, \dots, \hat{\zeta}_{21} \in \mathbb{P}^5(\hat{\mathcal{T}})$ have been computed already. We will not state the actual definition of these local shape functions here and refer the reader to Appendix A.4 or the implementation in the class `ArgyrisShapes` in M++ (see [110]).

Applying the transformation $\Phi_{\mathcal{T}}$ we define linear functionals $\tilde{\mathcal{L}}_1^{\mathcal{T}}, \dots, \tilde{\mathcal{L}}_{21}^{\mathcal{T}}$ on $\mathbb{P}^5(\mathcal{T})$ via

$$\tilde{\mathcal{L}}_i(\zeta) \coloneqq \hat{\mathcal{L}}_i(\zeta \circ \Phi_{\mathcal{T}}) \quad \text{for all } \zeta \in \mathbb{P}^5(\mathcal{T}),\ i = 1, \dots, 21 \tag{9.2}$$

We note that, in contrast to the Lagrangian case, the functionals $\tilde{\mathcal{L}}_1^{\mathcal{T}}, \dots, \tilde{\mathcal{L}}_{21}^{\mathcal{T}} \colon \mathbb{P}^5(\mathcal{T}) \to \mathbb{R}$ do not necessarily coincide with the functionals $\mathcal{L}_1^{\mathcal{T}}, \dots, \mathcal{L}_{21}^{\mathcal{T}} \colon \mathbb{P}^5(\mathcal{T}) \to \mathbb{R}$ representing the degrees of freedom on $\mathcal{T}$ as described in Figure 4.6. Recall that $\mathcal{L}_1^{\mathcal{T}}, \dots, \mathcal{L}_{21}^{\mathcal{T}}$ form a dual basis corresponding to the (unknown) local shape functions $\zeta_1^{\mathcal{T}}, \dots, \zeta_{21}^{\mathcal{T}} \in \mathbb{P}^5(\mathcal{T})$ defined on $\mathcal{T}$, i.e., we want to construct $\zeta_1^{\mathcal{T}}, \dots, \zeta_{21}^{\mathcal{T}}$ such that the duality condition $\mathcal{L}_i^{\mathcal{T}}(\zeta_j^{\mathcal{T}}) = \delta_{i,j}$ holds true for all $i, j = 1, \dots, 21$. Since both sets of functionals $\tilde{\mathcal{L}}_1^{\mathcal{T}}, \dots, \tilde{\mathcal{L}}_{21}^{\mathcal{T}}$ and $\mathcal{L}_1^{\mathcal{T}}, \dots, \mathcal{L}_{21}^{\mathcal{T}}$ respectively form a basis of the dual space of $\mathbb{P}^5(\mathcal{T})$, we can compute a transformation matrix $C^{\mathcal{T}} \in \mathbb{R}^{21\times 21}$ such that

$$\tilde{\mathcal{L}}_i^{\mathcal{T}} = \sum_{j=1}^{21} C_{i,j}^{\mathcal{T}} \mathcal{L}_j^{\mathcal{T}} \quad \text{for all } i = 1, \dots, 21. \tag{9.3}$$

We do not want to go into further details about the computation of the transformation matrix $C^{\mathcal{T}}$ and refer to the paper [22] by Dominguez and Sayas for more information about the construction process. In the following we suppose that we have the matrix $C^{\mathcal{T}}$ in hand and thus, we are in a position to show that the local shape functions $\zeta_1^{\mathcal{T}}, \dots, \zeta_{21}^{\mathcal{T}}$ defined by

$$\zeta_j^{\mathcal{T}} : \mathcal{T} \to \mathbb{R},\ \zeta_j^{\mathcal{T}}(x, y) := \sum_{k=1}^{21} C_{k,j}^{\mathcal{T}}\ \hat{\zeta}_k(\Phi_{\mathcal{T}}^{-1}(x, y)) \quad \text{for all } j = 1, \dots, 21$$

(cf. (4.10)) actually form a dual basis associated to the degrees of freedom $\mathcal{L}_1^{\mathcal{T}}, \dots, \mathcal{L}_{21}^{\mathcal{T}}$ defined via Figure 4.6 in Section 4.2.

With $D^{\mathcal{T}}$ denoting the inverse matrix of $C^{\mathcal{T}}$ the transformation formula (9.3) implies $\mathcal{L}_i^{\mathcal{T}} = \sum_{k=1}^{21} D_{i,k}^{\mathcal{T}} \tilde{\mathcal{L}}_k^{\mathcal{T}}$ for all $i = 1, \dots, 21$. Hence, using the local shape functions from above (see (4.10)), we calculate

$$\mathcal{L}_i^{\mathcal{T}}(\zeta_j^{\mathcal{T}}) = \sum_{k=1}^{21} D_{i,k}^{\mathcal{T}} \tilde{\mathcal{L}}_k^{\mathcal{T}}\Bigg(\sum_{l=1}^{21} C_{l,j}^{\mathcal{T}}\Big(\hat{\zeta}_l \circ \Phi_{\mathcal{T}}^{-1}\Big)\Bigg) = \sum_{k,l=1}^{21} D_{i,k}^{\mathcal{T}} C_{l,j}^{\mathcal{T}} \tilde{\mathcal{L}}_k^{\mathcal{T}}(\hat{\zeta}_l \circ \Phi_{\mathcal{T}}^{-1})$$

for all $i, j = 1, \dots, 21$. Furthermore, the definition of the functionals (see (9.2)) and the duality condition on the reference cell $\hat{\mathcal{T}}$ yield $\tilde{\mathcal{L}}_k^{\mathcal{T}}(\hat{\zeta}_l \circ \Phi_{\mathcal{T}}^{-1}) = \hat{\mathcal{L}}_k(\hat{\zeta}_l) = \delta_{k,l}$ for all $k, l = 1, \dots, 21$. Combining the arguments above and the fact that $C^{\mathcal{T}}$ and $D^{\mathcal{T}}$ are inverse to each other, we obtain

$$\mathcal{L}_i^{\mathcal{T}}(\zeta_j^{\mathcal{T}}) = \delta_{i,j} \quad \text{for all } i, j = 1, \dots, 21$$

which proves the desired duality property for the functionals $\mathcal{L}_1^{\mathcal{T}}, \dots, \mathcal{L}_{21}^{\mathcal{T}}$ and the local shape functions $\zeta_1^{\mathcal{T}}, \dots, \zeta_{21}^{\mathcal{T}}$.

Applying the chain rule and using the fact that the derivative of the transformation $\Phi_{\mathcal{T}}$ is a constant matrix, we obtain representation formulas for the gradients and Hessian matrices of the local shape functions on $\mathcal{T}$ using the corresponding counterparts on the reference cell $\hat{\mathcal{T}}$. To obtain these formulas we first denote the inverse of the Jacobian of the transformation $\Phi_{\mathcal{T}}$ by $F^{\mathcal{T}}$. Hence, we calculate

$$\begin{aligned} \frac{\partial \zeta_j^{\mathcal{T}}}{\partial x}(x, y) &= \sum_{k=1}^{21} C_{k,j}^{\mathcal{T}} \left[\frac{\partial \hat{\zeta}_k}{\partial \hat{x}}(\Phi_{\mathcal{T}}^{-1}(x, y)) F_{1,1}^{\mathcal{T}} + \frac{\partial \hat{\zeta}_k}{\partial \hat{y}}(\Phi_{\mathcal{T}}^{-1}(x, y)) F_{2,1}^{\mathcal{T}} \right], \\ \frac{\partial \zeta_j^{\mathcal{T}}}{\partial y}(x, y) &= \sum_{k=1}^{21} C_{k,j}^{\mathcal{T}} \left[\frac{\partial \hat{\zeta}_k}{\partial \hat{x}}(\Phi_{\mathcal{T}}^{-1}(x, y)) F_{1,2}^{\mathcal{T}} + \frac{\partial \hat{\zeta}_k}{\partial \hat{y}}(\Phi_{\mathcal{T}}^{-1}(x, y)) F_{2,2}^{\mathcal{T}} \right] \end{aligned} \tag{9.4}$$

for all $j = 1, \dots, 21$ and $(x, y) \in \mathcal{T}$.

Similar arguments yield the following representation for the entries of the Hessian matrices

$$
\begin{aligned}
\frac{\partial^2 \zeta_j^{\mathcal{T}}}{\partial x^2}(x,y) &= \sum_{k=1}^{21} C_{k,j}^{\mathcal{T}} \Bigg[\frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x}^2}(\Phi_{\mathcal{T}}^{-1}(x,y))(F_{1,1}^{\mathcal{T}})^2 + \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x} \partial \hat{y}}(\Phi_{\mathcal{T}}^{-1}(x,y)) 2 F_{2,1}^{\mathcal{T}} F_{2,1}^{\mathcal{T}} \\
&\qquad + \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{y}^2}(\Phi_{\mathcal{T}}^{-1}(x,y))(F_{2,1}^{\mathcal{T}})^2 \Bigg], \\
\frac{\partial^2 \zeta_j^{\mathcal{T}}}{\partial x \partial y}(x,y) &= \sum_{k=1}^{21} C_{k,j}^{\mathcal{T}} \Bigg[\frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x}^2}(\Phi_{\mathcal{T}}^{-1}(x,y)) F_{1,1}^{\mathcal{T}} F_{1,2}^{\mathcal{T}} + \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x} \partial \hat{y}}(\Phi_{\mathcal{T}}^{-1}(x,y))(F_{1,1}^{\mathcal{T}} F_{2,2}^{\mathcal{T}} + F_{1,2}^{\mathcal{T}} F_{2,1}^{\mathcal{T}}) \\
&\qquad + \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{y}^2}(\Phi_{\mathcal{T}}^{-1}(x,y)) F_{2,1}^{\mathcal{T}} F_{2,2}^{\mathcal{T}} \Bigg], \qquad (9.5) \\
\frac{\partial^2 \zeta_j^{\mathcal{T}}}{\partial y^2}(x,y) &= \sum_{k=1}^{21} C_{k,j}^{\mathcal{T}} \Bigg[\frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x}^2}(\Phi_{\mathcal{T}}^{-1}(x,y))(F_{1,2}^{\mathcal{T}})^2 + \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x} \partial \hat{y}}(\Phi_{\mathcal{T}}^{-1}(x,y)) 2 F_{1,2}^{\mathcal{T}} F_{2,2}^{\mathcal{T}} \\
&\qquad + \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{y}^2}(\Phi_{\mathcal{T}}^{-1}(x,y))(F_{2,2}^{\mathcal{T}})^2 \Bigg]
\end{aligned}
$$

for all $j = 1, \dots, 21$ and $(x,y) \in \mathcal{T}$. Finally, these representation formulas are used for the implementation of the Argyris element in M++ (see [110, `ArgyrisElement`]).

9.4.2 Raviart Thomas Element

In Section 5.2 and Section 7.2 we rely on finite elements that provide an approximation in $H(\mathrm{div}, \Omega, \mathbb{R}^{2\times 2})$. In this Section we present an implementation of higher order Raviart Thomas elements using the ideas described by Ervin in [25, Section 3.4], i.e., we are interested in the explicit construction of a basis for the Raviart Thomas space RT_k. Once more, for the computation of the Raviart Thomas basis functions we consider the reference triangle $\hat{\mathcal{T}}$ (see Section 4.2) and consider a counterclockwise numbering of the edges starting at the bottom edge with the number zero.

Following the lines in [25] we first define the auxiliary functions $\hat{\mathbf{e}}_0, \dots, \hat{\mathbf{e}}_4$ by

$$
\hat{\mathbf{e}}_0(\hat{x},\hat{y}) := \begin{pmatrix} \hat{x} \\ \hat{y}-1 \end{pmatrix}, \quad \hat{\mathbf{e}}_1(\hat{x},\hat{y}) := \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \quad \hat{\mathbf{e}}_2(\hat{x},\hat{y}) := \begin{pmatrix} \hat{x}-1 \\ \hat{y} \end{pmatrix},
$$

$$
\hat{\mathbf{e}}_3(\hat{x},\hat{y}) := \hat{y} \begin{pmatrix} \hat{x} \\ \hat{y}-1 \end{pmatrix}, \quad \hat{\mathbf{e}}_4(\hat{x},\hat{y}) := \hat{x} \begin{pmatrix} \hat{x}-1 \\ \hat{y} \end{pmatrix}
$$

for all $(\hat{x},\hat{y}) \in \hat{\mathcal{T}}$. Note that the functions $\hat{\mathbf{e}}_0, \hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$ are the well-known basis functions for the lowest order Raviart Thomas space $RT_0(\hat{\mathcal{T}})$.

In contrast to Ervin for the construction of the Lagrangian polynomials $l_0, \dots, l_k$ (cf. [25, Section 3.4]), we do not use the Gaussian quadrature points on $[0,1]$. Rather, we chose points such that $[0,1]$ is divided into $k+2$ equidistant subintervals which allows us to use the Lagrangian basis functions of order k (together with an easy transformation) as the desired polynomials $l_0, \dots, l_k$. We note that only the $k+1$ points in the interior of $[0,1]$ have to be considered which provides the correct number of Raviart Thomas basis functions (cf. Figure 9.1 and [25, Section 3.4]). We do not want to go into further details here and refer the reader to the implementation in the class `RTShapesTriangular` where especially the formulas for the transformation mentioned above can be found. Furthermore, let $\{b_i \colon i = 1, \dots, \frac{k(k+1)}{2}\}$ denote a basis for $\mathbb{P}^{k-1}(\hat{\mathcal{T}})$, i.e., we can use Lagrangian basis functions of order $k-1$ for triangles to form the desired basis of $\mathbb{P}^{k-1}(\hat{\mathcal{T}})$. Thus, in the implementation of the Raviart Thomas shape functions we can mostly use functions that are already part of the Finite Element Software M++.

Similar to Ervin we define the functions

$$\hat{\mathbf{r}}_{0,i}(\hat{x},\hat{y}) := l_i(\hat{x})\hat{\mathbf{e}}_0(\hat{x},\hat{y}), \quad \hat{\mathbf{r}}_{1,i}(\hat{x},\hat{y}) := l_i(\hat{y})\hat{\mathbf{e}}_1(\hat{x},\hat{y}), \quad \hat{\mathbf{r}}_{2,i}(\hat{x},\hat{y}) := l_{k-i}(\hat{y})\hat{\mathbf{e}}_2(\hat{x},\hat{y})$$

for all $(\hat{x},\hat{y}) \in \hat{\mathcal{T}}$ and $i = 0, \dots, k$ as well as

$$\hat{\mathbf{r}}_{3,i}(\hat{x},\hat{y}) := b_i(\hat{x},\hat{y})\hat{\mathbf{e}}_3(\hat{x},\hat{y}), \quad \hat{\mathbf{r}}_{4,i}(\hat{x},\hat{y}) := b_i(\hat{x},\hat{y})\hat{\mathbf{e}}_4(\hat{x},\hat{y})$$

for all $(\hat{x},\hat{y}) \in \hat{\mathcal{T}}$ and $i = 1, \dots, \frac{k(k+1)}{2}$. Then, from [25, Section 3.4] it follows that these functions $\hat{\mathbf{r}}_{k,i}$ form the desired basis for $RT_k(\hat{\mathcal{T}})$.

9.5 Methods for Approximating and Enclosing Eigenvalues

For a successful computation of the norm bounds K, and K^* respectively, described in Section 6.2 we heavily rely on rigorous bounds for the isolated eigenvalues of an eigenvalue problem of the form

$$M(u,\varphi) = \lambda N(u,\varphi) \quad \text{for all } \varphi \in H, \tag{9.6}$$

where H denotes a separable (complex) Hilbert space endowed with the inner product $N \colon H \times H \to \mathbb{R}$ and $M \colon H \times H \to \mathbb{R}$ is a bounded, positive definite symmetric bilinear form on H. We note that the unconventional notation N for the inner product coincides with the naming in M++ and is only used for the reader's convenience when looking at the source code.

All eigenvalue methods described in Section 6.2.1 require the computation of approximate eigenpairs or at least approximate eigenvalues of our eigenvalue problem (9.6). To obtain such approximations, we use an implementation of the Locally Optimal Block Preconditioned Conjugate Gradient method (LOBPCG) which is an iterative procedure to compute approximate eigenpairs of a symmetric generalized eigenvalue problem (cf. [20]). Note that several versions of LOBPCG are already implemented in M++ (see for instance

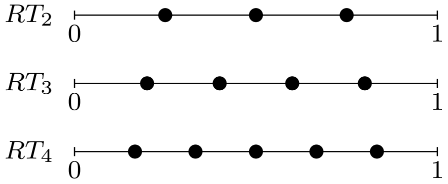


Figure 9.1: Points used for the Lagrangian polynomials $l_0, \dots, l_k$ for $k = 2, 3, 4$

[62]). However, at several stages of the code we made small changes in the implementation to increase the efficiency of the algorithms.

The implementation of the Rayleigh Ritz method and Lehmann method (together with its Goerisch extension) to compute upper and lower eigenvalue bounds respectively are straightforward (cf. Section 6.2.1). In contrast to that the realization of an "automated" homotopy method needs several new ideas. Hence, in Section 9.5.3 we explain our implementation which allows the application of the homotopy method in an almost automatic framework.

Moreover, to guarantee the most possible flexibility for all eigenvalue methods presented in Section 6.2.1 we make use of an adaptable assemble class which can easily be adjusted to several other problems. In Section 9.5.2 we shortly explain the usage of our new assemble class `IAEigenvalueAssemble`.

9.5.1 Verified Matrix Eigenvalue Computations

At several stages in this thesis, especially in the application of the eigenvalue methods presented earlier (cf. Theorem 6.7 or Theorem 6.8), we have to compute verified enclosures for eigenvalues of the generalized matrix eigenvalue problem $Ax = \lambda Bx$ with a hermitian matrix A and a positive definite, hermitian matrix B.

First, regarding the definitions in Section 3.3 we denote the space of complex $d \times d$ interval matrices with $[\mathbb{C}]^{d\times d}$. Then, the following Lemma together with the interval arithmetic operations described in Section 3.3 yields verified eigenvalue enclosures on the basis of approximate eigenpairs. For a proof we refer to [42] and references therein.

Lemma 9.2. *Let $[A],[B] \in [\mathbb{C}]^{d\times d}$ be hermitian matrices with interval entries such that $[B]$ is positive definite for all hermitian $B \in [B]$. Moreover, for fixed hermitian $A_0 \in [A]$ and $B_0 \in [B]$ let $(\tilde{\lambda}_i, \tilde{x}_i)$ for all $i = 1, \dots, d$ denote approximate eigenpairs of $A_0 x = \lambda B_0 x$ with $\overline{\tilde{x}_i}^T B_0 \tilde{x}_j \approx \delta_{ij}$ for all $i, j = 1, \dots, d$. Additionally, suppose that some $r_0, r_1 > 0$ exist such that*

$$\left\|\overline{X}^T AX - \overline{X}^T BXK\right\|_\infty \leq r_0 \quad \text{and} \quad \left\|\overline{X}^T BX - \mathrm{id}\right\|_\infty \leq r_1 \quad \text{for all } A \in [A], B \in [B]$$

with $X := (\tilde{x}_1, \dots, \tilde{x}_d)$, $K := \operatorname{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_d)$.

If $r_1 < 1$, we have for all $A \in [A]$, $B \in [B]$ and all eigenvalues λ of $Ax = \lambda Bx$:

$$\lambda \in \bigcup_{i=1}^{d} B(\tilde{\lambda}_i, r) \quad \text{where } r = \frac{r_0}{1 - r_1}, \text{ and } B(\lambda, r) = \{z \in \mathbb{C} \colon |z - \lambda| \leq r\}$$

Additionally, each connected component of this union contains as many eigenvalues as midpoints $\tilde{\lambda}_i$.

Remark 9.3. *The MATLAB package INTLAB by Rump (cf. [90]) provides highly accurate algorithms for enclosing the eigenvalues of the generalized matrix eigenvalue problem $Ax = \lambda Bx$. However, we do not make use of these algorithms since our programs are written in C++, i.e., each time a generalized matrix eigenvalue problem has to be solved we would have to transform the matrices to MATLAB, perform the calculations and transform back the results which is simply not practical.*

We note that the fixed hermitian matrices A_0 and B_0 required in Lemma 9.2 to compute the approximate eigenpairs can be chosen arbitrarily in $[A]$ and $[B]$ respectively. Hence, in our implementations we use the midpoint matrix of $[A]$ and $[B]$ respectively.

To compute the required approximate eigenpairs we use routines implemented in the standard BLAS library (see for instance [80]). The verification process described above is implemented in the latest version of M++ (see `IASpectrum`).

Since all calculations (except of the computation of the approximate eigenpairs) require interval arithmetic operations, the enclosing procedure needs relatively high computational effort. But in all (our) applications the dimension d is small, which makes the computational effort acceptable.

9.5.2 Assemble Class `IAEigenvalueAssemble`

As mentioned in the beginning of this Section, all implementations of the eigenvalue approximation and enclosure methods introduced in Section 6.2.1 are based on the assemble class `IAEigenvalueAssemble` which provides an interface with all routines necessary for the computations. This assemble class can easily be adopted to several other problems guaranteeing the maximal flexibility in the application of the eigenvalue methods. Moreover, we added the data type `Eigenpair` which contains a `Vector` to store the approximate eigenfunction and a `double` data type for the approximate eigenvalue. Note that there exists also the corresponding “vector version” `Eigenpairs` which consists of a set of finitely many eigenpairs.

First of all, the assemble can be initialized with an additional shift parameter which is accessible in each function and thus, for instance, can be used as required in the Lehmann Goerisch method (cf. Section 6.2.1).

Furthermore, since all implemented eigenvalue methods require approximate eigenpairs the assemble also provides an interface to the non-verified iterative eigenvalue solvers LOBPCG mentioned above. In this procedure the local finite element basis functions are used to compute approximate eigenpairs to our eigenvalue problem (9.6). Therefore, the user can implement the boundary conditions cell-wise by overriding the function `BoundaryConditionsEigenSolver`. Moreover, the mass matrix as well as the stiffness matrix corresponding to our eigenvalue problem (9.6) are assembled cell-wise by overriding the function `MatricesEigenSolver`. Note that besides the shift parameter mentioned above a second parameter `t` is passed in this function which can be used as the homotopy parameter introduced in Section 6.2.1. Finally, the computation of approximate eigenpairs is realized in the function `InitEigenpairs` which is contained in all the eigenvalue methods of M++. Note that `InitEigenpairs` requires an instance of `IAEigenvalueAssemble` where the functions for the approximate computation of eigenpairs are overridden.

In the further course let $(\tilde{u}_1, \tilde{\lambda}_1), \dots, (\tilde{u}_d, \tilde{\lambda}_d)$ denote approximate eigenpairs which can be computed by the algorithms presented above. In the following, we shortly describe which functions have to be implemented for an application of the eigenvalue methods of M++.

For the application of the Rayleigh Ritz method the user has to override the function `MatrixentriesRayleighRitz`, which provides the matrix entries $M(\tilde{u}_i, \tilde{u}_j)$ and $N(\tilde{u}_i, \tilde{u}_j)$ for all $i, j = 1, \dots, d$ (cf. Section 6.2.1). We want to point out that the parameter `t` can be ignored if no homotopy is involved and only upper bounds for a single eigenvalue problem are of interest. Finally, the Rayleigh Ritz computation can be started by calling the `operator()` of the class `RayleighRitzMethod`.

Recall that in the Lehmann Goerisch method the computation of additional functions w_i (for all $i = 1, \ldots, d$) is a crucial step (cf. Section 6.2.1). Thus, in our assemble class one has to implement the functions `BoundaryConditionsGoerisch`, `EnergyGoerisch`, `ResidualGoerisch` and `JacobiGoerisch` which provide all the data needed for the computation of the functions w_i (for all $i = 1, \ldots, d$) via Newton's methods. Additionally, the user has to override the function `MatrixentriesGoerisch` which assembles the cell-wise computation of the matrix entries $b(w_i, w_j)$ for all $i, j = 1, \ldots, d$ needed for the Lehmann Goerisch method. Again, the step parameter `t` can be ignored if only a single eigenvalue problem is considered.

Remark 9.4. (i) *Besides a verified version of the functions to set up the Rayleigh Ritz and Lehmann Goerisch matrices, our assemble class* `IAEigenvalueAssemble` *provides a non-verified version which can be used to run our eigenvalue methods without verified computations which massively decreases the computational time. In particular, this feature can be used to get a rough idea if the eigenvalue might be successful for this setting of parameters.*

(ii) *From Section 6.2.1 it follows that the homotopy method needs a Lehmann Goerisch computation (of "small" size) in each homotopy step. Thus, we can use the functions of the Lehmann Goerisch method described above for the homotopy method as well, where the additional homotopy parameter* `t` *comes into play.*

(iii) *In some applications for the assembling of the Rayleigh Ritz matrices additional Newton computations are required. Therefore, similar to the Lehmann Goerisch method our assemble provides the additional functions* `BoundaryConditionsRayleighRitz`*,* `EnergyRayleighRitz`*,* `ResidualRayleighRitz` *and* `JacobiRayleighRitz`*.*

9.5.3 Implementation of the Homotopy Method

In the following we present an approach for the implementation of the homotopy method described in Section 6.2.1 which allows an almost automatic usage of the algorithm. Thereby, the main difficulty lies in the correct computation of the next homotopy step parameter, i.e., the new parameter value for which we have to compute new approximate eigenpairs (cf. Section 6.2.1).

Before starting the homotopy algorithm the user has to fix the range in which the homotopy parameter can be varied, i.e., `stepMin` denotes the starting value for the homotopy parameter for which the exact number of eigenvalues below the lower bound of the essential spectrum ρ_0 is known. In addition to that, `stepMax` contains the maximal value for our homotopy parameter which will be denoted by $t_{\max}$ in the further course.

In the following, with respect to Section 6.2.1, we denote the number of eigenpairs considered in the i-th homotopy step by N_i and obtain

$$N_i = \begin{cases} n_0, & i = 1, \\ n_0 - n_{i-1}, & i > 1, \end{cases} \tag{9.7}$$

where the values n_i are defined as in the description of the homotopy method in Section 6.2.1.

For the reader's convenience, in the following description of our approach for the computation of a new homotopy parameter, we first assume that all eigenvalues of the considered eigenvalue problem are well separated, i.e., no clustered eigenvalues do exist. If there are such eigenvalue clusters we slightly have to change our strategy which will be presented at the end of this Section.

Now, assume that for some $i \in \mathbb{N}$ we already computed $i-1$ steps of the homotopy method, i.e., the next step under investigation is the i-th one (in the case $i = 1$ this is the first step). In other words the currently taken homotopy parameter is t_{i-1} and we are interested in a parameter $t_i > t_{i-1}$ such that assumption (6.20) in Corollary 6.9 is satisfied (and almost an equality) for the approximate eigenfunction with v replaced by $\tilde{u}_{N_i}^{(t_i)}$ and ρ replaced by ρ_{i-1} (cf. blue bracket in Figure 9.2).

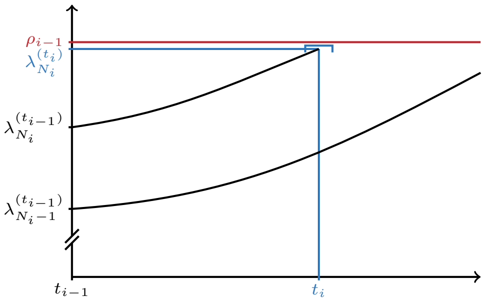


Figure 9.2: Possible course of the largest eigenvalues in the i-th homotopy step

Since up to the present day there exists no analytical theory on the computation of an optimal homotopy parameter value t_i, we developed an algorithm that exploits the fact that the computation of approximate eigenpairs is of lower computational effort and thus relatively fast compared to the computation of the lower bounds via the Lehmann-Goerisch method. Our algorithm is divided into two parts, whereas the first part deals with the computation of two step parameters $t_{i-1} < \underline{t}_i < \bar{t}_i$ such that $\tilde{\lambda}_{N_i}^{(\underline{t}_i)} \leq \rho_{i-1} < \tilde{\lambda}_{N_i}^{(\bar{t}_i)}$. In the second part we perform several bisection steps based on both values $\underline{t}_i$ and $\bar{t}_i$ to obtain the desired candidate for t_i. Algorithm 4 gives an overview of the entire algorithm in pseudo code. Afterwards we present the steps of our algorithm in detail.

Remark 9.5. *We note that, compared to a simple "guess" of the new homotopy parameter, in our algorithm several approximate eigenpairs have to be computed (cf. Algorithm 4). However, our examples showed that the quality of the lower eigenvalue bounds required in the homotopy steps heavily depends on the choice of the homotopy parameter value t_i. Therefore, the extra effort for the additional computation of approximate eigenpairs is justified. Moreover, in Remark 9.6* (iii) *we present a strategy to reduce the computational effort by first using a coarser finite element mesh for a rough computation of the desired homotopy parameter.*

Algorithm 4: Computation of the next homotopy parameter t_i

```
$N_i$          : Number of eigenpairs to be considered
$\rho_{i-1}$   : Lower bound for the (probably existing) $(N_i+1)$st eigenvalue
$t_{i-1}$      : Previous homotopy parameter
$t_{\max}$     : Prescribed maximal homotopy parameter (stepMax)
$\tau$         : Prescribed step size (stepSize)
$\varepsilon$  : Prescribed tolerance (nearRho)
$S$            : Prescribed maximal number of bisection steps (numBisectionSteps)

$\overline{t}_i := t_{i-1}$
$r := 0$
while true do
    $\underline{t}_i := \overline{t}_i$
    $\overline{t}_i := \min\{\underline{t}_i + 2^r\tau, t_{\max}\}$
    Compute approximate eigenvalues $\{\tilde{\lambda}_j^{(\overline{t}_i)} : 1, \ldots, N_i\}$ for step parameter $\overline{t}_i$
    if $\rho_i < \tilde{\lambda}_{N_i}^{(\overline{t}_i)}$ then break
    if $\overline{t}_i == t_{\max}$ then return $t_{\max}$
    $r := r + 1$
end
if $\rho_{i-1} - \varepsilon \leq \tilde{\lambda}_{N_i}^{(\underline{t}_i)}$ then return $\underline{t}_i$
for $k := 0$ to $S$ do
    $m_i := \frac{1}{2}(\underline{t}_i + \overline{t}_i)$
    Compute approximate eigenvalues $\{\tilde{\lambda}_j^{(m_i)} : 1, \ldots, N_i\}$ for step parameter $m_i$
    if $\rho_{i-1} \leq \tilde{\lambda}_{N_i}^{(m_i)}$ then
        $\overline{t}_i := m_i$
    else if $\rho_{i-1} - \varepsilon \leq \tilde{\lambda}_{N_i}^{(\underline{t}_i)}$ then
        return $m_i$
    else
        $\underline{t}_i := m_i$
    end
end
return $\underline{t}_i$
```

Now, we describe the ideas used in our algorithm in detail starting with the first part. Therefore, we first fix some step size $\tau > 0$ (which in our software is stored in the variable `stepSize`). Starting from the previous homotopy parameter t_{i-1}, we first consider the new homotopy parameter $\overline{t}_{i,0} := \min\{t_i + \tau, t_{\max}\}$ and compute approximate eigenpairs $(\tilde{\lambda}_j^{(\overline{t}_{i,0})}, \tilde{u}_j^{(\overline{t}_{i,0})})$ for $j = 1, \ldots, N_i$.

Then, on the basis of the approximate eigenvalues we check the inequality $\rho_{i-1} < \tilde{\lambda}_{N_i}^{(\overline{t}_{i,0})}$. In the affirmative case we set $\underline{t}_i := t_{i-1}$ and $\overline{t}_i := \overline{t}_{i,0}$. Otherwise, we have to distinguish two cases: First, if $\overline{t}_{i,0}$ equals $t_{\max}$ our algorithm for the computation of the desired homotopy parameter stops and we obtain $t_i = t_{\max}$.

Otherwise, we have $\overline{t}_{i,0} < t_{\max}$ and thus, we continue our algorithm by repeating the first step but now with the new homotopy parameter $\overline{t}_{i,1} := \min\{t_i + 2\tau, t_{\max}\}$, i.e., we compute approximate eigenpairs $(\tilde{\lambda}_j^{(\overline{t}_{i,1})}, \tilde{u}_j^{(\overline{t}_{i,1})})$ for $j = 1, \ldots, N_i$ and check the condition $\rho_{i-1} < \tilde{\lambda}_{N_i}^{(\overline{t}_{i,1})}$.

Now, in the affirmative case we set $\underline{t}_i := \overline{t}_{i,0}$ and $\overline{t}_i := \overline{t}_{i,1}$. If the inequality is not satisfied for the approximate eigenvalue $\tilde{\lambda}_{N_i}^{(\bar{t}_{i,1})}$ we check if $\bar{t}_{i,1} = t_{\max}$ which (in the affirmative case) stops our algorithm with $t_i = t_{\max}$.

Otherwise, we can repeat this procedure until we either end up with $t_i = t_{\max}$ (which finishes our algorithm) or after finitely many iterations of the procedure presented above, we obtain $\underline{t}_i := t_{i-1} + 2^{r-1}\tau$ and $\bar{t}_i := t_{i-1} + 2^r\tau$ for some integer $r \in \mathbb{N}$ such that $\tilde{\lambda}_{N_i}^{(t_{i-1}+2^{r-1}\tau)} \le \rho_{i-1} < \tilde{\lambda}_{N_i}^{(t_{i-1}+2^r\tau)}$ (cf. Figure 9.3).

We note that if the approximate eigenpairs are computed with sufficient accuracy we might expect that the same inequalities hold true for the exact eigenvalues instead of the approximations as well, i.e., we expect $\lambda_{N_i}^{(\underline{t}_i)} \le \rho_{i-1} < \lambda_{N_i}^{(\bar{t}_i)}$ to be satisfied too.

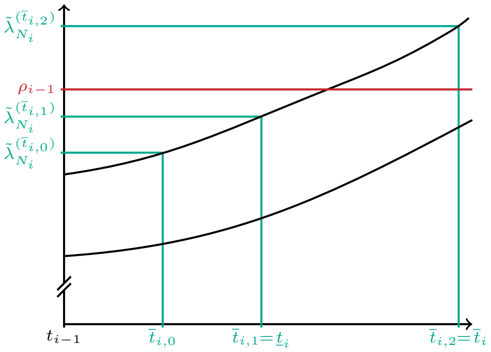

$\tilde{\lambda}_{N_i}^{(\bar{t}_{i,2})}$
ρ_{i-1}
$\tilde{\lambda}_{N_i}^{(\bar{t}_{i,1})}$
$\tilde{\lambda}_{N_i}^{(\bar{t}_{i,0})}$
t_{i-1}
$\bar{t}_{i,0}$
$\bar{t}_{i,1}=\underline{t}_i$
$\bar{t}_{i,2}=\bar{t}_i$

Figure 9.3: First part of the algorithm in the i-th homotopy step

Now, the second part of our algorithm uses several bisection steps to "optimize" the step parameter $\underline{t}_i$ such that $\tilde{\lambda}_{N_i}^{(\underline{t}_i)} < \rho_{i-1}$ is almost an equality. Thus, for the explanations in the following, we set $\underline{t}_{i,0} := \underline{t}_i$ and $\bar{t}_{i,0} := \bar{t}_i$.

To specify the "quality" of a parameter $\tilde{t}$ we first fix a tolerance $\varepsilon > 0$ which in our implementation is denoted by `nearRho` and state that our present homotopy parameter $\tilde{t}$ is "sufficiently good" if $\rho_{i-1} - \varepsilon \le \tilde{\lambda}_{N_i}^{(\tilde{t})} < \rho_{i-1}$. Thus, we check this condition for our final parameter $\underline{t}_{i,0}$ (below ρ_{i-1}) from the previous step, i.e., if $\rho_{i-1} - \varepsilon \le \tilde{\lambda}_{N_i}^{(\underline{t}_{i,0})} (< \rho_{i-1})$ holds true $t_i := \underline{t}_{i,0}$ is our desired homotopy parameter.

If this is not the case, we start our bisection procedure for the "improvement" of $\underline{t}_{i,0}$, i.e., in the first refinement step we compute approximate eigenpairs $(\tilde{\lambda}_j^{(m_{i,1})}, \tilde{u}_j^{(m_{i,1})})$ for $j = 1, \ldots, N_i$ with the midpoint parameter $m_{i,1} := \frac{1}{2}(\underline{t}_{i,0} + \bar{t}_{i,0})$. Then, if $\lambda_{N_i}^{(m_{i,1})} > \rho_{i-1}$ holds true we set $\bar{t}_{i,0} := m_{i,1}$ and $\underline{t}_{i,1} := \underline{t}_{i,0}$. Otherwise, we define $\underline{t}_{i,0} = m_{i,1}$ and $\bar{t}_{i,1} := \bar{t}_{i,0}$. Moreover, we check again if $\rho_{i-1} - \varepsilon \le \tilde{\lambda}_{N_i}^{(\underline{t}_{i,1})} (< \rho_{i-1})$ holds true. In the affirmative case we obtain $t_i := \underline{t}_{i,0}$ which is the desired final homotopy parameter (cf. Figure 9.4).

Otherwise, we repeat this step until either we have found the desired homotopy parameter t_i with $\rho_{i-1} - \varepsilon \le \tilde{\lambda}_{N_i}^{(t_i)} < \rho_{i-1}$ or a predefined maximal number of bisection steps is reached (in our algorithm denoted by `numBisectionSteps`). Note that in the second case our algorithm stops without finding an "optimal" homotopy parameter, nevertheless, the final parameter can be used in the further homotopy steps but the bounds might be a bit smaller than possibly could be achieved.

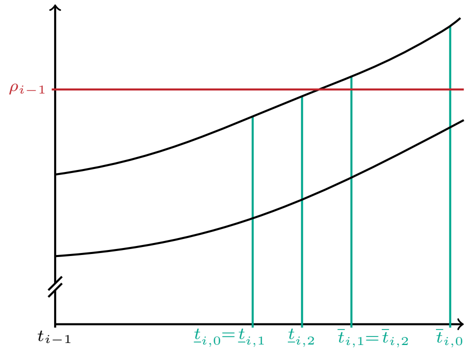

ρ_{i-1}
t_{i-1}
$\underline{t}_{i,0}=\underline{t}_{i,1}$
$\underline{t}_{i,2}$
$\overline{t}_{i,1}=\overline{t}_{i,2}$
$\overline{t}_{i,0}$

Figure 9.4: Second part of the algorithm in the i-th homotopy step

Having computed the next homotopy step parameter t_i we use the Temple-Lehmann method together with its Goerisch extension in form of Corollary 6.9 to compute the new bound ρ_i. The algorithm used for the homotopy method is listed in Algorithm 5.

Algorithm 5: Computation of the homotopy

```
$\rho_0$ : Prescribed bound (cf. Section 6.2.1)
$n_0$ : Number of eigenvalues of the base problem below $\rho_0$
$t_{\mathbf{min}}$ : Prescribed minimal homotopy parameter (stepMin)
$t_{\mathbf{max}}$ : Prescribed maximal homotopy parameter (stepMax)

$t := t_{\min}$, $\rho := \rho_0$, $n := n_0$
while true do
    $t \leftarrow$ Compute next parameter (cf. Algorithm 4) starting from previous step $t$
    if $t == t_{\max}$ then break
    Compute approximate eigenfunctions $\{\tilde{u}_j^{(t)} : 1, \dots, n\}$ for step parameter $t$
    Check condition (6.20) in Corollary 6.9 with $\tilde{u}_n^{(t)}$ and $\rho$
    Compute $w$ corresponding to $\tilde{u}_n^{(t)}$ satisfying (6.16) in Theorem 6.8
    $\rho \leftarrow$ Compute new lower bound by (6.21) in Corollary 6.9
    $n := n - 1$ // eigenvalues are assumed to be well separated
    if $n == 0$ then return $\rho$
end
Final Rayleigh Ritz computation for the $n$ remaining eigenpairs
return $\rho$
```

Since the implementation of the Lehmann-Goerisch algorithm (or its Corollary) is straightforward we omit an description of the implementation in this thesis, however, it is part of our M++ package for (verified) eigenvalue methods. Note that for the computation of the function w needed in Corollary 6.9 in the `IAEigenvalueAssemble` the functions `BoundaryConditionsGoerisch`, `ResidualGoerisch` and `JacobiGoerisch` need to be implemented. Of course, for the computation of the lower bound given by (6.21) in our eigenvalue assemble the functions `MatrixentriesGoerisch` and `MatrixentriesRayleighRitz` have to be overridden as well.

Remark 9.6. (i) *In our applications it turned out that for the computation of the new bounds ρ_i using Corollary 6.9, it makes sense to reduce the bound ρ_{i-1} (which is the bound from the previous step) by some small constant $\hat{\rho} > 0$ which is denoted by* `separationRho` *in our implementation. Hence, compared to the explanations above our algorithm in fact uses the bound* $\rho_{i-1} - \hat{\rho}$ *instead of* ρ_{i-1} *to check the crucial inequalities like* $\tilde{\lambda}^{(\underline{t}_i)}_{N_i} \leq \rho_{i-1} - \hat{\rho} < \tilde{\lambda}^{(\bar{t}_i)}_{N_i}$ *mentioned above. Nevertheless, for the reader's convenience we used the non-shifted bounds in our explanations above.*

(ii) *In all our examples choosing the step size in the range* 0.001 *to* 0.05 *turned out to be a "good" choice. However, this choice heavily depends on the problem which is under investigation.*

(iii) *In view of Remark 9.5, to speed up the computation of the new homotopy parameter and to reduce the computational effort of our procedure, we first perform our algorithm on a coarser mesh to obtain a "rough" homotopy parameter which afterwards (in a second step) can be improved by additional bisection steps computed on the fine mesh. In all our applications it turned our that this additional improvement part only requires a few (in most of the examples only one) steps.*

At the end of this Section, we shortly describe our algorithm if eigenvalue clusters appear during the application. Therefore, we assume that starting from step $i-1$ (and t_{i-1}) we have computed a new homotopy step parameter t_i using the algorithm presented above. Moreover, we fix a parameter $\varepsilon > 0$ which represents the minimal distance of two consecutive eigenvalues (in our software denoted by `separationCluster`). Now, before computing the desired new bound ρ_i, on the basis of the approximate eigenvalues $\tilde{\lambda}^{(t_i)}_j$ for $j = 1, \ldots, N_i$, we check if the distance between $\tilde{\lambda}^{(t_i)}_{N_i}$ and $\tilde{\lambda}^{(t_i)}_{N_i-1}$ is greater or equal to a prescribed tolerance ε. In the affirmative case, we expect the eigenvalues $\lambda^{(t_i)}_{N_i}$ and $\lambda^{(t_i)}_{N_i-1}$ to be "well separated" and we are in the situation described previously.

Otherwise, we bunch together all eigenvalues (counted from above) which have a distance smaller than our prescribed tolerance ε. Hence, we compute some $m_i \leq N_i$ such that the distance $\tilde{\lambda}^{(t_i)}_{k+1} - \tilde{\lambda}^{(t_i)}_k < \varepsilon$ for all $m_i \leq k < N_i$. Then, instead of Corollary 6.9 we apply the Lehmann-Goerisch method (cf. Theorem 6.8) to the complete eigenvalue cluster and obtain the desired bound $\rho_i < \lambda^{(t_i)}_{m_i}$.

A Appendix

A.1 Basic Inequalities and Embedding Constants

In this section we present proofs of several basic inequalities and embedding constants. We start with the investigation of the divergence operator, after that we give a short proof of Poincaré's inequality on the unbounded strip $S := \mathbb{R} \times (0,1)$. Moreover, we present a method to obtain computable upper bounds for Sobolev's embedding constants. Finally, we close this Section with a proof providing that the operator Φ (see (2.14) and (2.15)) actually defines an isometric isomorphism.

Lemma A.1. *Let $G \subseteq \Omega$ be a (sub) domain of Ω. Then, the divergence operator*

$$\operatorname{div}\colon H_0^1(G,\mathbb{R}^2) \to L^2(G)$$

is bounded with $\|\operatorname{div} u\|_{L^2(G)} \leq \sqrt{2}\|u\|_{H_0^1(G,\mathbb{R}^2)}$ for all $u \in H_0^1(G,\mathbb{R}^2)$.

Proof. For all $u \in H_0^1(G,\mathbb{R}^2)$ we calculate

$$\begin{aligned}\|\operatorname{div} u\|_{L^2(G)}^2 &= \int_G \left|\frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y}\right|^2 \,\mathrm{d}(x,y)\\ &\leq 2\int_G \left[\left(\frac{\partial u_1}{\partial x}\right)^2 + \left(\frac{\partial u_2}{\partial y}\right)^2\right] \mathrm{d}(x,y)\\ &\leq 2\|\nabla u\|_{L^2(G,\mathbb{R}^{2\times 2})}^2 \leq 2\|u\|_{H_0^1(G,\mathbb{R}^2)}^2.\end{aligned}$$

Thus, we obtain $\|\operatorname{div} u\|_{L^2(G)} \leq \sqrt{2}\|u\|_{H_0^1(G,\mathbb{R}^2)}$ for all $u \in H_0^1(G,\mathbb{R}^2)$ which also implies the boundedness of div. □

All over this thesis we trust on the existence of computable bounds for the Sobolev embeddings $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$ for $p \in [2,\infty)$. The following Lemma yields possibly not optimal, but easily computable constants C_p satisfying

$$\|u\|_{L^p(\Omega)} \leq C_p\|u\|_{H_0^1(\Omega)} \quad \text{for all } u \in H_0^1(\Omega) \tag{A.1}$$

with the H_0^1-norm defined in [85, p. 34], i.e., the identity $\|u\|_{H_0^1(\Omega)}^2 = \|\nabla u\|_{L^2(\Omega,\mathbb{R}^2)}^2 + \sigma\|u\|_{L^2(\Omega)}^2$ holds true for all $u \in H_0^1(\Omega)$.

Lemma A.2. *Let $G \subseteq S$ be a (sub) domain of S, $p \in [2,\infty)$, $\nu := \lceil \frac{p}{2} \rceil$ and*

$$\gamma := \begin{cases} \left(\frac{2}{p\sigma}\right)^{\frac{2}{p}} \left(1-\frac{2}{p}\right)^{1-\frac{2}{p}}, & 2(\pi^2+\sigma) \leq p\sigma, \\ \frac{\pi^{2(1-\frac{2}{p})}}{\pi^2+\sigma}, & \textit{otherwise.} \end{cases} \tag{A.2}$$

Then

$$\|u\|_{L^p(G)} \leq C_p \|u\|_{H_0^1(G)} \quad \textit{for all } u \in H_0^1(G) \tag{A.3}$$

holds true with

$$\left(\frac{1}{2}\right)^{\frac{1}{2}+\frac{2\nu-3}{p}} \left[\frac{p}{2}\left(\frac{p}{2}-1\right)\cdots\left(\frac{p}{2}-\nu+2\right)\right]^{\frac{2}{p}} \sqrt{\gamma},$$

where the bracket-term is set equal to 1 *if* $\nu = 1$.

For a proof we refer the reader to [74, Lemma 7.10]. We note, that the constants given in [74] depend on a minimum ρ^* of the spectrum of $-\Delta$ on $H_0^1(G)$. Since $G \subseteq S$, [74, Remark 7.11 (a)] gives $\rho^* = \pi^2$. Thus, [74, Lemma 7.10] directly yields the assertion for $\sigma > 0$. Nevertheless, for $\rho^* > 0$, the proof can be extended to the case $\sigma = 0$ as well.

Hence, Lemma A.2 applied with $G := \Omega$ yields the desired embedding constants C_p satisfying (A.1). Especially, Lemma A.2 provides explicit bounds for the cases $p = 2$ and $p = 4$, respectively, i.e., we obtain

$$C_2 = \frac{1}{\sqrt{\pi^2+\sigma}} \quad \text{and} \quad C_4 = \begin{cases} \frac{1}{\sqrt[4]{8\sigma}}, & \pi^2 \leq \sigma, \\ \frac{1}{\sqrt[4]{2}}\sqrt{\frac{\pi}{\pi^2+\sigma}}, & \text{otherwise.} \end{cases}$$

At this stage, we want to point out that for $p = 2$ the first case in (A.2) cannot occur which results in the definition for the embedding constant C_2 above. Moreover, in most of our examples we choose $\sigma \in [0,1]$, i.e., in our applications only the second case in the definition of C_4 is used.

Remark A.3. *We note that [74, Lemma 7.10] provides an analogue version of Lemma A.2 for dimensions $n \geq 3$.*

Then, Poincaré's inequality on the unbounded strip S is a direct consequence of Lemma A.2.

Lemma A.4 (Poincaré's inequality on the strip)**.**

(i) *On the strip S Poincaré's inequality holds true, i.e. we have*

$$\|u\|_{L^2(S)} \leq \frac{1}{\pi}\|\nabla u\|_{L^2(S,\mathbb{R}^2)} \quad \textit{for all } u \in H_0^1(S).$$

(ii) *For $u \in H_0^1(S,\mathbb{R}^2)$ we have*

$$\|u\|_{L^2(S,\mathbb{R}^2)} \leq \frac{1}{\pi}\|\nabla u\|_{L^2(S,\mathbb{R}^{2\times 2})}, \tag{A.4}$$

where the norms are similarly defined as in (2.1) *and* (2.4) *with Ω replaced by S.*

Proof. (i) Lemma A.2 applied with $\sigma = 0$ and $p = 2$ directly yields $C_2 = \frac{1}{\pi}$. Thus, from (A.3) the assertion follows.

(ii) The assertion is a direct consequence of part (i) and the definitions of the involved norms:

$$\begin{aligned} \|u\|^2_{L^2(S,\mathbb{R}^2)} &= \|u_1\|^2_{L^2(S)} + \|u_2\|^2_{L^2(S)} \\ &\leq \|\nabla u_1\|^2_{L^2(S,\mathbb{R}^2)} + \|\nabla u_2\|^2_{L^2(S,\mathbb{R}^2)} = \|\nabla u\|^2_{L^2(S,\mathbb{R}^{2\times 2})} \quad \text{for all } u \in H^1_0(S,\mathbb{R}^2). \end{aligned}$$

□

Remark A.5. *For any (sub) domain $G \subseteq S$ we can embed $H^1_0(G,\mathbb{R}^2)$ into $H^1_0(S,\mathbb{R}^2)$ by zero extension which directly implies that* (A.4) *is satisfied for all $u \in H^1_0(G,\mathbb{R}^2)$ as well. Especially, since our domain Ω is contained in the strip S, Poincaré's inequality* (A.4) *holds for $H^1_0(\Omega,\mathbb{R}^2)$ as well.*

Finally, we catch up with the proof that the operator Φ is actually defines an isometric isomorphism.

Lemma A.6. *$\Phi\colon H(\Omega) \to H(\Omega)'$, $\Phi u := -\Delta u + \sigma u$ defined in Section 2.3 is an isometric isomorphism, i.e, Φ is bijective and the equality $\|\Phi u\|_{H(\Omega)'} = \|u\|_{H^1_0(\Omega,\mathbb{R}^2)}$ holds true for all $u \in H(\Omega)$.*

Proof. First, by Riesz' Representation Lemma for bounded linear functionals for a given functional $l \in H(\Omega)'$ there exists an element $w_l \in H(\Omega)$ such that

$$l[\varphi] = \langle w_l, \varphi \rangle_{H^1_0(\Omega,\mathbb{R}^2)} = (\Phi w_l)[\varphi] \quad \text{for all } \varphi \in H(\Omega),$$

i.e., Φ is surjective.

Furthermore, using Cauchy-Schwarz' inequality we calculate

$$\|\Phi u\|_{H(\Omega)'} = \sup_{\substack{\varphi \in H(\Omega) \\ \varphi \neq 0}} \frac{|(\Phi u)[\varphi]|}{\|\varphi\|_{H^1_0(\Omega,\mathbb{R}^2)}} = \sup_{\substack{\varphi \in H(\Omega) \\ \varphi \neq 0}} \frac{\left|\langle u, \varphi \rangle_{H^1_0(\Omega,\mathbb{R}^2)}\right|}{\|\varphi\|_{H^1_0(\Omega,\mathbb{R}^2)}} \leq \|u\|_{H^1_0(\Omega,\mathbb{R}^2)} \quad \text{for all } u \in H(\Omega).$$

Moreover, testing Φu against $\varphi = u \neq 0$ we obtain $\|\Phi u\|_{H(\Omega)'} \geq \|u\|_{H^1_0(\Omega,\mathbb{R}^2)}$ and thus, we proved the equality $\|\Phi u\|_{H(\Omega)'} = \|u\|_{H^1_0(\Omega,\mathbb{R}^2)}$ for all $u \in H(\Omega)$ (note that the equality also holds for $u = 0$) which especially implies that Φ is one-to-one. □

A.2 Fourier Transform on the Strip

In this paragraph we present the proofs omitted in Section 2.4. First, we show a Lemma dealing with the invertibility of the Fourier transform introduced in Section 2.4.

Lemma A.7. *For the Fourier transform on $\mathcal{D}$ defined in* (2.19) *the identities*

$$\mathcal{F}_x \circ \mathcal{F}_x^{-1} = \mathcal{F}_x^{-1} \circ \mathcal{F}_x = \mathrm{id}_{\mathcal{D}}$$

hold true, where $\mathcal{F}^{-1}$ is defined in (2.20) *and $\mathcal{D}$ is given in* (2.18).

Proof. Let $u = \sum_{n=-N}^{N} u_n\varphi_n \in \mathcal{D}$. Then (2.19) and (2.20) together with well-known properties of the common Fourier transform $\mathcal{F}$ on $L^2(\mathbb{R},\mathbb{C})$ imply

$$\begin{aligned}\mathcal{F}_x\left[\mathcal{F}_x^{-1}[u]\right](x,y) &= \mathcal{F}_x\left[\sum_{n=-N}^{N}\mathcal{F}^{-1}[u_n]\varphi_n\right](x,y) = \sum_{n=-N}^{N}\mathcal{F}\left[\mathcal{F}^{-1}[u_n]\right](x)\varphi_n(y)\\ &= \sum_{n=-N}^{N} u_n(x)\varphi_n(y) = u(x,y) \quad \text{for all } (x,y)\in S.\end{aligned}$$

Hence, we obtain $\mathcal{F}_x \circ \mathcal{F}_x^{-1} = \mathrm{id}_{\mathcal{D}}$. Similar calculations show

$$\mathcal{F}_x^{-1}[\mathcal{F}_x[u]](x,y) = \sum_{n=-N}^{N}\mathcal{F}^{-1}[\mathcal{F}[u_n]](x)\varphi_n(y) = u(x,y) \quad \text{for all } (x,y)\in S,$$

yielding $\mathcal{F}_x^{-1}\circ\mathcal{F}_x = \mathrm{id}_{\mathcal{D}}$, which finishes the proof. □

In the further course, we give the proof of Lemma 2.4 omitted in Section 2.4 which provides some basic properties concerning derivatives of the new Fourier transform $\mathcal{F}_x$.

Proof of Lemma 2.4 . First, by (2.17) we get $\frac{\partial\varphi_n}{\partial y} = -2\mathrm{i}\pi n\varphi_n$ for all $n\in\mathbb{Z}$. Thus, iterative application of this formula, together with $c_n := -2\mathrm{i}\pi n$, yields $\frac{\partial^k\varphi_n}{\partial y^k} = c_n^k\varphi_n$ for all $n\in\mathbb{Z}$. Now, let $u = \sum_{n=-N}^{N} u_n\varphi_n \in \mathcal{D}$.

(i) For the first assertion we calculate

$$\frac{\partial^{j+k}u}{\partial x^j\partial y^k}(x,y) = \sum_{n=-N}^{N}\frac{\partial^j u_n}{\partial x^j}(x)\frac{\partial^k\varphi_n}{\partial y^k}(y) = \sum_{n=-N}^{N}\frac{\partial^j u_n}{\partial x^j}(x)c_n^k\varphi_n(y) \quad \text{for all } (x,y)\in S. \tag{A.5}$$

Since $\frac{\partial^j u_n}{\partial x^j}c_n^k \in \mathcal{S}(\mathbb{R},\mathbb{C})$, (A.5), (2.19) and the well-known properties of the Fourier transform on $\mathcal{S}(\mathbb{R},\mathbb{C})$ imply

$$\begin{aligned}\mathcal{F}_x\left[\frac{\partial^{j+k}u}{\partial x^j\partial y^k}\right](\xi,y) &= \sum_{n=-N}^{N}\mathcal{F}\left[\frac{\partial^j u_n}{\partial x^j}c_n^k\right](\xi)\varphi_n(y) = (\mathrm{i}\xi)^j\sum_{n=-N}^{N}\mathcal{F}[u_n](\xi)c_n^k\varphi_n(y)\\ &= (\mathrm{i}\xi)^j\sum_{n=-N}^{N}\mathcal{F}[u_n](\xi)\frac{\partial^k\varphi_n}{\partial y^k}(y) = (\mathrm{i}\xi)^j\left(\frac{\partial^k}{\partial y^k}\mathcal{F}_x[u]\right)(\xi,y)\end{aligned}$$

for all $(\xi,y)\in S$.

(ii) The definition of v directly yields

$$\frac{\partial^k v}{\partial y^k}(x,y) = (-\mathrm{i}x)^j\frac{\partial^k u}{\partial y^k}(x,y) = \sum_{n=-N}^{N}(-\mathrm{i}x)^j u_n(x)c_n^k\varphi_n(y) \quad \text{for all } (x,y)\in S.$$

The usual Fourier transform, applied to $v_n\colon\mathbb{R}\to\mathbb{C}$, $v_n(x) := (-\mathrm{i}x)^j u_n(x)c_n^k$ for all $n\in\mathbb{Z}$, gives

$$\mathcal{F}[v_n](\xi) = c_n^k\left(\frac{\partial^j}{\partial\xi^j}\mathcal{F}[u_n]\right)(\xi) \quad \text{for all } \xi\in\mathbb{R},\ n\in\mathbb{Z}.$$

Combining the arguments above implies

$$\begin{aligned}\mathcal{F}_x\left[\frac{\partial^k v}{\partial y^k}\right](\xi, y) &= \sum_{n=-N}^{N} \mathcal{F}[v_n](\xi)\varphi_n(y) = \sum_{n=-N}^{N} \left(\frac{\partial^j}{\partial \xi^j}\mathcal{F}[u_n]\right)(\xi) c_n^k \varphi_n(y) \\ &= \sum_{n=-N}^{N} \left(\frac{\partial^j}{\partial \xi^j}\mathcal{F}[u_n]\right)(\xi)\frac{\partial^k \varphi_n}{\partial y^k}(y) = \left(\frac{\partial^{j+k}}{\partial \xi^j \partial y^k}\mathcal{F}_x[u]\right)(\xi, y)\end{aligned}$$

for all $(\xi, y) \in S$.

□

After having completed the proofs for the Fourier transform $\mathcal{F}_x$, in the following, we prove related results for the distributional Fourier transform defined in (2.25). First, we start with the proof of a Lemma similar to Lemma A.7 but for the distributional Fourier transform.

Lemma A.8. *$\mathcal{F}_x \colon \mathcal{D}' \to \mathcal{D}'$ is invertible and its inverse is given by* (2.26)*.*

Proof. We consider the operator

$$G \colon \mathcal{D}' \to \mathcal{D}',\ (G[f])[\varphi] := f[\mathcal{F}_x^{-1}[\varphi]] \quad \text{for all } \varphi \in \mathcal{D}$$

defined by the right-hand side of (2.26). Then, applying Lemma A.7, for $f \in \mathcal{D}'$, we calculate

$$(\mathcal{F}_x[G[f]])[\varphi] = \left(\mathcal{F}_x^{-1}[f]\right)[\mathcal{F}_x[\varphi]] = f\left[\mathcal{F}_x^{-1}[\mathcal{F}_x[\varphi]]\right] = f[\varphi] \quad \text{for all } \varphi \in \mathcal{D}$$

and

$$(G[\mathcal{F}_x[f]])[\varphi] = (\mathcal{F}_x[f])\left[\mathcal{F}_x^{-1}[\varphi]\right] = f\left[\mathcal{F}_x\left[\mathcal{F}_x^{-1}[\varphi]\right]\right] = f[\varphi] \quad \text{for all } \varphi \in \mathcal{D},$$

i.e., G is the inverse of $\mathcal{F}_x$ and the assertion follows. □

Finally, we close this subsection with the proof of Lemma 2.7 introduced in Section 2.4 using the corresponding Lemma 2.4 for the Fourier transform on $\mathcal{D}$.

Proof of Lemma 2.7 . (i) Let $f \in \mathcal{D}'$ and $\varphi \in \mathcal{D}$. Then, the definition of the distributional Fourier transform (see (2.25)) and the definition of its derivative (see (2.28)) imply

$$\left(\mathcal{F}_x\left[\frac{\partial^{j+k} f}{\partial x^j \partial y^k}\right]\right)[\varphi] = \left(\frac{\partial^{j+k} f}{\partial x^j \partial y^k}\right)[\mathcal{F}_x[\varphi]] = (-1)^{j+k} f\left[\frac{\partial^{j+k}}{\partial x^j \partial y^k}\mathcal{F}_x[\varphi]\right].$$

Next, applying Lemma 2.4 (ii) with $v := (-\mathrm{i}x)^j \varphi$ we obtain

$$(-1)^{j+k} f\left[\frac{\partial^{j+k}}{\partial x^j \partial y^k}\mathcal{F}_x[\varphi]\right] = (-1)^{j+k} f\left[\mathcal{F}_x\left[\frac{\partial^k v}{\partial y^k}\right]\right]$$

which, together with (2.25), yields

$$(-1)^{j+k} f\left[\frac{\partial^{j+k}}{\partial x^j \partial y^k}\mathcal{F}_x[\varphi]\right] = (-1)^{j+k}(\mathcal{F}_x[f])\left[\frac{\partial^k v}{\partial y^k}\right].$$

Finally, (2.28) and (2.29), together with the definition of v and the linearity of the Fourier transform, show

$$(-1)^{j+k}(\mathcal{F}_x[f])\left[\frac{\partial^k v}{\partial y^k}\right] = (-1)^j\left(\frac{\partial^k}{\partial y^k}\mathcal{F}_x[f]\right)[v] = \left((\mathrm{i}x)^j\frac{\partial^k}{\partial y^k}\mathcal{F}_x[f]\right)[\varphi]$$

which proves the first assertion.

(ii) Again, let $f \in \mathcal{D}'$ and $\varphi \in \mathcal{D}$. Using the same arguments as in part (i), with Lemma 2.4 (i) (instead of (ii)), and the definition of g we get

$$\begin{aligned}\left(\mathcal{F}_x\left[\frac{\partial^k g}{\partial y^k}\right]\right)[\varphi] &= \left(\frac{\partial^k g}{\partial y^k}\right)[\mathcal{F}_x[\varphi]] = (-1)^k g\left[\frac{\partial^k}{\partial y^k}\mathcal{F}_x[\varphi]\right]\\ &= (-1)^{j+k} f\left[(\mathrm{i}\,\cdot)^j\frac{\partial^k}{\partial y^k}\mathcal{F}_x[\varphi]\right] = (-1)^{j+k} f\left[\mathcal{F}_x\left[\frac{\partial^{j+k}\varphi}{\partial x^j\partial y^k}\right]\right]\\ &= (-1)^{j+k}(\mathcal{F}_x[f])\left[\frac{\partial^{j+k}\varphi}{\partial x^j\partial y^k}\right] = \left(\frac{\partial^{j+k}}{\partial \xi^j\partial y^k}\mathcal{F}_x[f]\right)[\varphi].\end{aligned}$$

□

A.3 Properties of some Integral Terms

Here, we present some useful facts about the integrals contained in the definition of the operators B and B_w (for some $w \in W(\Omega)$). The proofs are rather technical but the results are used at several stages of this work. First, using matrix vector multiplication and the definition of the derivative introduced in Section 2.1 we calculate

$$[(u\cdot\nabla)v]\cdot\varphi = \left[u_1\frac{\partial v}{\partial x} + u_2\frac{\partial v}{\partial y}\right]\cdot\varphi = [(\nabla v)u]\cdot\varphi = \varphi^T(\nabla v)u \qquad \text{(A.6)}$$

for sufficiently smooth functions $u, v, \varphi\colon \Omega \to \mathbb{R}^2$.

Before proving several statements concerning integral terms we state the proofs of Proposition 3.1 (which was left out in Section 3.1) and Proposition 6.2 (postponed in Chapter 6):

Proof of Proposition 3.1 . (i) Let $v, w \in W(\Omega)$ and $u \in H(\Omega)$. Using the definition of B_{v+w} and the bilinearity of the inner product we conclude

$$\begin{aligned}\mathrm{B}_{v+w}\,u &= \mathit{Re}[(u\cdot\nabla)(v+w) + ((v+w)\cdot\nabla)u]\\ &= \mathit{Re}[(u\cdot\nabla)v + (v\cdot\nabla)u] + \mathit{Re}[(u\cdot\nabla)w + (w\cdot\nabla)u] = \mathrm{B}_v\,u + \mathrm{B}_w\,u.\end{aligned}$$

(ii) Let $v \in W(\Omega)$ and $u \in H(\Omega)$. The same arguments as in part (i) yield

$$\mathrm{B}_{-v}\,u = \mathit{Re}[(u\cdot\nabla)(-v) + ((-v)\cdot\nabla)u] = -\mathit{Re}[(u\cdot\nabla)v + (v\cdot\nabla)u] = -\,\mathrm{B}_v\,u.$$

(iii) Let $w \in H(\Omega) \cap W(\Omega)$ and $u, \varphi \in H(\Omega)$. Using the definitions of B and B_w together with the ideas presented in Section 2.2 we calculate

$$\begin{aligned}(\mathrm{B}(u,w) + \mathrm{B}(w,u))[\varphi] &= Re \int_\Omega [(u \cdot \nabla)w] \cdot \varphi \, \mathrm{d}(x,y) + Re \int_\Omega [(w \cdot \nabla)u] \cdot \varphi \, \mathrm{d}(x,y) \\ &= Re \int_\Omega [(u \cdot \nabla)w + (w \cdot \nabla)u] \cdot \varphi \, \mathrm{d}(x,y) \\ &= \int_\Omega (\mathrm{B}_w \, u) \cdot \varphi \, \mathrm{d}(x,y) = (\mathrm{B}_w \, u)[\varphi].\end{aligned}$$

□

Proof of Proposition 6.2 . (i) Let $v, w \in W(\Omega)$ and $u \in H(\Omega)$. The definition of $\hat{\mathrm{B}}_{v+w}$ together with the linearity of the matrix product implies

$$\begin{aligned}\hat{\mathrm{B}}_{v+w} \, u &= Re[(\nabla(v+w))^T u - (\nabla u)(v+w)] \\ &= Re[(\nabla v)^T u - (\nabla u)v] + Re[(\nabla w)^T u - (\nabla u)w] = \hat{\mathrm{B}}_v \, u + \hat{\mathrm{B}}_w \, u.\end{aligned}$$

(ii) Let $v \in W(\Omega)$ and $u \in H(\Omega)$. The same arguments as in part (i) yield

$$\hat{\mathrm{B}}_{-v} \, u = Re[(-\nabla v)^T u - (\nabla u)(-v)] = -Re[(\nabla v)^T u - (\nabla u)v] = -\hat{\mathrm{B}}_v \, u.$$

□

The following Lemma provides some results for estimating integral terms appearing at several stages of this thesis. We proof this Lemma in a slightly more general setting, i.e., for functions in $H_0^1(\Omega, \mathbb{R}^2)$, but since we have $H(\Omega) \subseteq H_0^1(\Omega, \mathbb{R}^2)$ the assertions remain valid also if we replace $H_0^1(\Omega, \mathbb{R}^2)$ by $H(\Omega)$ (which will be the standard application in most of the cases).

Lemma A.9. (i) *For $u, v, \varphi \in H_0^1(\Omega, \mathbb{R}^2)$ the following estimates are satisfied:*

$$\begin{aligned}\int_\Omega [(u \cdot \nabla)v] \cdot \varphi \, \mathrm{d}(x,y) = \int_\Omega \varphi^T (\nabla v) u \, \mathrm{d}(x,y) &\le \|u\|_{L^4(\Omega,\mathbb{R}^2)} \|\nabla v\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \|\varphi\|_{L^4(\Omega,\mathbb{R}^2)} \\ &\le C_4 \|u\|_{L^4(\Omega,\mathbb{R}^2)} \|\nabla v\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)} \\ &\le {C_4}^2 \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \|v\|_{H_0^1(\Omega,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}.\end{aligned}$$

(ii) *Let $w \in W^{1,\infty}(\Omega, \mathbb{R}^2)$, $u \in H^1(\Omega, \mathbb{R}^2)$ and define $Q := \mathrm{supp}(w) \cap \mathrm{supp}(u)$. Then for all $\varphi \in H_0^1(\Omega, \mathbb{R}^2)$ the following inequalities hold true:*

$$\begin{aligned}\int_\Omega [(u \cdot \nabla)w] \cdot \varphi \, \mathrm{d}(x,y) = \int_\Omega \varphi^T (\nabla w) u \, \mathrm{d}(x,y) &\le \|u\|_{L^2(Q,\mathbb{R}^2)} \|\nabla w\|_{L^\infty(Q,\mathbb{R}^{2\times 2})} \|\varphi\|_{L^2(Q,\mathbb{R}^2)} \\ &\le C_2 \|u\|_{L^2(Q,\mathbb{R}^2)} \|\nabla w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}.\end{aligned}$$

Furthermore, for $u \in H_0^1(\Omega, \mathbb{R}^2)$ we get

$$\int_\Omega [(u \cdot \nabla)w] \cdot \varphi \, \mathrm{d}(x,y) \le {C_2}^2 \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \|\nabla w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}.$$

(iii) *Let $w \in W^{1,\infty}(\Omega, \mathbb{R}^2)$, $A \in L^2(\Omega, \mathbb{R}^2)$ and set $Q := \operatorname{supp}(w) \cap \operatorname{supp}(A)$. Then for all $\varphi \in H_0^1(\Omega, \mathbb{R}^2)$ we have*

$$\int_\Omega (Aw) \cdot \varphi \, \mathrm{d}(x,y) \leq \|A\|_{L^2(Q,\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{L^2(Q,\mathbb{R}^2)}$$

implying

$$\begin{aligned} \int_\Omega [(w \cdot \nabla)u] \cdot \varphi \, \mathrm{d}(x,y) &= \int_\Omega \varphi^T (\nabla u) w \, \mathrm{d}(x,y) \\ &\leq \|\nabla u\|_{L^2(Q,\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{L^2(Q,\mathbb{R}^2)} \\ &\leq C_2 \|\nabla u\|_{L^2(Q,\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)} \end{aligned}$$

for any function $u \in H^1(\Omega, \mathbb{R}^2)$. Moreover, for $u \in H_0^1(\Omega, \mathbb{R}^2)$ we obtain

$$\begin{aligned} \int_\Omega [(w \cdot \nabla)u] \cdot \varphi \, \mathrm{d}(x,y) &\leq \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{L^2(Q,\mathbb{R}^2)} \\ &\leq C_2 \|u\|_{H_0^1(\Omega,\mathbb{R}^2)} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}. \end{aligned}$$

Proof. In each part the equality is a direct consequence of (A.6).

(i) By Hölder's inequality we get

$$\begin{aligned} &\int_\Omega \varphi^T (\nabla v) u \, \mathrm{d}(x,y) \\ &\leq \sum_{i=1}^2 \int_\Omega \left| u_1 \frac{\partial v_i}{\partial x} \varphi_i + u_2 \frac{\partial v_i}{\partial y} \varphi_i \right| \mathrm{d}(x,y) \\ &\leq \sum_{i=1}^2 \left[\|u_1\|_{L^4(\Omega)} \left\| \frac{\partial v_i}{\partial x} \right\|_{L^2(\Omega)} \|\varphi_i\|_{L^4(\Omega)} + \|u_2\|_{L^4(\Omega)} \left\| \frac{\partial v_i}{\partial y} \right\|_{L^2(\Omega)} \|\varphi_i\|_{L^4(\Omega)} \right] \end{aligned}$$

for all $u, v, \varphi \in H_0^1(\Omega, \mathbb{R}^2)$. Hence, applying Cauchy-Schwarz' inequality in $\mathbb{R}^2$ twice, we obtain

$$\begin{aligned} &\int_\Omega \varphi^T (\nabla v) u \, \mathrm{d}(x,y) \\ &\leq \|u\|_{L^4(\Omega,\mathbb{R}^2)} \sum_{i=1}^2 \|\varphi_i\|_{L^4(\Omega)} \left(\left\| \frac{\partial v_i}{\partial x} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_i}{\partial y} \right\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}} \\ &\leq \|u\|_{L^4(\Omega,\mathbb{R}^2)} \|\nabla v\|_{L^2(\Omega,\mathbb{R}^{2\times 2})} \|\varphi\|_{L^4(\Omega,\mathbb{R}^2)} \end{aligned}$$

for all $u, v, \varphi \in H_0^1(\Omega, \mathbb{R}^2)$ which proves the first part of the assertion. The second and the third inequality directly follow from the embedding (2.12).

(ii) Let $u \in H^1(\Omega,\mathbb{R}^2)$, $\varphi \in H_0^1(\Omega,\mathbb{R}^2)$, $w \in W(\Omega)$ and $Q = \operatorname{supp}(w)$. Using Cauchy-Schwarz' inequality we obtain

$$\begin{aligned}
&\int_\Omega \varphi^T(\nabla w)u \,\mathrm{d}(x,y)\\
&\quad= \int_Q \varphi^T(\nabla w)u \,\mathrm{d}(x,y)\\
&\quad\leq \sum_{i=1}^2 \int_Q \left|\varphi_i\left(\frac{\partial w_i}{\partial x}u_1 + \frac{\partial w_i}{\partial y}u_2\right)\right| \mathrm{d}(x,y)\\
&\quad\leq \sum_{i=1}^2 \left(\left\|\frac{\partial w_i}{\partial x}\right\|_{L^\infty(Q)} \int_Q |\varphi_i u_1|\,\mathrm{d}(x,y) + \left\|\frac{\partial w_i}{\partial y}\right\|_{L^\infty(Q)} \int_Q |\varphi_i u_2|\,\mathrm{d}(x,y)\right)\\
&\quad\leq \sum_{i=1}^2 \|\varphi_i\|_{L^2(Q)} \left(\left\|\frac{\partial w_i}{\partial x}\right\|_{L^\infty(Q)} \|u_1\|_{L^2(Q)} + \left\|\frac{\partial w_i}{\partial y}\right\|_{L^\infty(Q)} \|u_2\|_{L^2(Q)}\right).
\end{aligned}$$

Now, Cauchy-Schwarz' inequality in $\mathbb{R}^2$ implies

$$\begin{aligned}
&\left\|\frac{\partial w_i}{\partial x}\right\|_{L^\infty(Q)} \|u_1\|_{L^2(Q)} + \left\|\frac{\partial w_i}{\partial y}\right\|_{L^\infty(Q)} \|u_2\|_{L^2(Q)}\\
&\quad\leq \|u\|_{L^2(Q,\mathbb{R}^2)} \left(\left\|\frac{\partial w_i}{\partial x}\right\|_{L^\infty(Q)}^2 + \left\|\frac{\partial w_i}{\partial y}\right\|_{L^\infty(Q)}^2\right)^{\frac{1}{2}}
\end{aligned}$$

Inserting this result into the first calculation we get

$$\begin{aligned}
&\int_\Omega \varphi^T(\nabla w)u \,\mathrm{d}(x,y)\\
&\quad\leq \|u\|_{L^2(Q,\mathbb{R}^2)} \sum_{i=1}^2 \|\varphi_i\|_{L^2(Q)} \left(\left\|\frac{\partial w_i}{\partial x}\right\|_{L^\infty(Q)}^2 + \left\|\frac{\partial w_i}{\partial y}\right\|_{L^\infty(Q)}^2\right)^{\frac{1}{2}}\\
&\quad\leq \|u\|_{L^2(Q,\mathbb{R}^2)} \|\varphi\|_{L^2(Q,\mathbb{R}^2)} \left(\sum_{i=1}^2 \left(\left\|\frac{\partial w_i}{\partial x}\right\|_{L^\infty(Q)}^2 + \left\|\frac{\partial w_i}{\partial y}\right\|_{L^\infty(Q)}^2\right)\right)^{\frac{1}{2}}\\
&\quad= \|u\|_{L^2(Q,\mathbb{R}^2)} \|\nabla w\|_{L^\infty(Q,\mathbb{R}^{2\times 2})} \|\varphi\|_{L^2(Q,\mathbb{R}^2)},
\end{aligned}$$

where we again used Cauchy-Schwarz' inequality in $\mathbb{R}^2$ for the last estimate, which proves the first assertion.

Hence, the embedding result (2.12) yields

$$\|u\|_{L^2(Q,\mathbb{R}^2)} \|\nabla w\|_{L^\infty(Q,\mathbb{R}^{2\times 2})} \|\varphi\|_{L^2(Q,\mathbb{R}^2)} \leq C_2 \|u\|_{L^2(Q,\mathbb{R}^2)} \|\nabla w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}.$$

Finally, for $u \in H_0^1(\Omega,\mathbb{R}^2)$ (note that the calculations above remain true) we use the embedding (2.12) again to conclude the remaining assertion.

(iii) Let $A \in L^2(\Omega,\mathbb{R}^{2\times 2})$, $\varphi \in H(\Omega)$, $w \in W(\Omega)$ and $Q \coloneqq \operatorname{supp}(w)$. The same arguments as in part (ii) yield

$$\begin{aligned}
&\int_\Omega (Aw)\cdot\varphi \,\mathrm{d}(x,y)\\
&= \sum_{i,j=1}^{2}\int_Q \varphi_i A_{ij} w_j \,\mathrm{d}(x,y) \le \sum_{i,j=1}^{2} \|w_j\|_{L^\infty(Q)} \int_Q |\varphi_i A_{ij}| \,\mathrm{d}(x,y)\\
&\le \sum_{j=1}^{2} \|w_j\|_{L^\infty(Q)} \sum_{i=1}^{2} \|A_{ij}\|_{L^2(Q)} \|\varphi_i\|_{L^2(Q)}\\
&\le \sum_{j=1}^{2} \|w_j\|_{L^\infty(Q)} \left(\sum_{i=1}^{2} \|A_{ij}\|^2_{L^2(Q)}\right)^{\frac{1}{2}} \|\varphi\|_{L^2(Q,\mathbb{R}^2)}\\
&\le \|A\|_{L^2(Q,\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{L^2(Q,\mathbb{R}^2)}.
\end{aligned}$$

Since $u \in H^1(\Omega,\mathbb{R}^2)$ implies $\nabla u \in L^2(\Omega,\mathbb{R}^{2\times 2})$ the inequality above implies

$$\begin{aligned}
\int_\Omega [(w\cdot\nabla)u]\cdot\varphi \,\mathrm{d}(x,y) &= \int_\Omega \varphi^T(\nabla u)w \,\mathrm{d}(x,y)\\
&\le \|\nabla u\|_{L^2(Q,\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{L^2(Q,\mathbb{R}^2)}\\
&\le C_2 \|\nabla u\|_{L^2(Q,\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(Q,\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}
\end{aligned}$$

for $u \in H^1(\Omega,\mathbb{R}^2)$, where we used the embedding (2.12) again. Finally, once more the embedding (2.12) yields the last equalities for $u \in H_0^1(\Omega,\mathbb{R}^2)$.

□

Again, we note that by Lemma A.9 above the integrals of the form used in the definition of the operator B (cf. (3.1)) and $\hat{\mathrm{B}}$ (cf. (6.2)) are well-defined. The next Lemma provides some useful calculation rules when dealing with such integrals.

Lemma A.10. (i) *For all* $u, v, \tilde{v} \in H(\Omega)$

$$\begin{aligned}
\int_\Omega [(u\cdot\nabla)v]\cdot\tilde{v} \,\mathrm{d}(x,y) &= \int_\Omega \tilde{v}^T(\nabla v)u \,\mathrm{d}(x,y)\\
&= -\int_\Omega v^T(\nabla\tilde{v})u \,\mathrm{d}(x,y) = -\int_\Omega [(u\cdot\nabla)\tilde{v}]\cdot v \,\mathrm{d}(x,y)
\end{aligned}$$

holds true.

(ii) *For all* $u, v \in H(\Omega)$ *and* $w \in W(\Omega)$ *we have*

$$\begin{aligned}
\int_\Omega [(w\cdot\nabla)u]\cdot v \,\mathrm{d}(x,y) &\int_\Omega v^T(\nabla u)w \,\mathrm{d}(x,y)\\
&= -\int_\Omega u^T(\nabla v)w \,\mathrm{d}(x,y) = -\int_\Omega [(w\cdot\nabla)v]\cdot u \,\mathrm{d}(x,y)
\end{aligned}$$

and

$$\begin{aligned}
\int_\Omega [(u\cdot\nabla)w]\cdot v \,\mathrm{d}(x,y) &= \int_\Omega v^T(\nabla w)u \,\mathrm{d}(x,y)\\
&= -\int_\Omega w^T(\nabla v)u \,\mathrm{d}(x,y) = -\int_\Omega [(u\cdot\nabla)v]\cdot w \,\mathrm{d}(x,y).
\end{aligned}$$

(iii) *It holds*

$$\int_\Omega [(u \cdot \nabla)v] \cdot v \,\mathrm{d}(x,y) = \int_\Omega v^T(\nabla v)u \,\mathrm{d}(x,y) = 0 \quad \textit{for all } u, v \in H(\Omega)$$

and

$$\int_\Omega [(w \cdot \nabla)v] \cdot v \,\mathrm{d}(x,y) = \int_\Omega v^T(\nabla v)w \,\mathrm{d}(x,y) = 0 \quad \textit{for all } v \in H(\Omega), w \in W(\Omega).$$

Proof. (i) The first and last equality of the assertion directly follow from (A.6). Thus, the second equality remains to be proved. Using integration by parts we obtain for all $u, v, \tilde{v} \in H(\Omega)$:

$$\begin{aligned}
&\int_\Omega \tilde{v}^T(\nabla v)u \,\mathrm{d}(x,y)\\
&= \int_\Omega \left[u_1 \frac{\partial v}{\partial x} + u_2 \frac{\partial v}{\partial y}\right] \cdot \tilde{v} \,\mathrm{d}(x,y)\\
&= \sum_{i=1}^{2} \int_\Omega \left[u_1 \frac{\partial v_i}{\partial x}\tilde{v}_i + u_2 \frac{\partial v_i}{\partial y}\tilde{v}_i\right] \mathrm{d}(x,y)\\
&= -\sum_{i=1}^{2} \int_\Omega \left[\frac{\partial u_1}{\partial x} v_i\tilde{v}_i + \frac{\partial u_2}{\partial y} v_i\tilde{v}_i + u_1 v_i \frac{\partial \tilde{v}_i}{\partial x} + u_2 v_i \frac{\partial \tilde{v}_i}{\partial y}\right] \mathrm{d}(x,y)\\
&= -\sum_{i=1}^{2} \int_\Omega \operatorname{div}(u) v_i\tilde{v}_i \,\mathrm{d}(x,y) - \int_\Omega \left[u_1 \frac{\partial \tilde{v}}{\partial x} + u_2 \frac{\partial \tilde{v}}{\partial y}\right] \cdot v \,\mathrm{d}(x,y)\\
&= -\sum_{i=1}^{2} \int_\Omega \operatorname{div}(u) v_i\tilde{v}_i \,\mathrm{d}(x,y) - \int_\Omega v^T(\nabla \tilde{v})u \,\mathrm{d}(x,y).
\end{aligned}$$

Due to $u \in H(\Omega)$ we have $\operatorname{div}(u) = 0$ which implies the missing assertion.

(ii) The same arguments as in part (i) imply the assertion.

(iii) Let $u, v \in H(\Omega)$. Applying part (i) with $\tilde{v} = v$ we obtain

$$\int_\Omega [(u \cdot \nabla)v] \cdot v \,\mathrm{d}(x,y) = -\int_\Omega [(u \cdot \nabla)v] \cdot v \,\mathrm{d}(x,y)$$

and thus, $\int_\Omega [(u \cdot \nabla)v] \cdot v \,\mathrm{d}(x,y) = 0$. The remaining equality is proved mutatis mutandis using part (ii) instead of (i).

□

Finally, we close this Subsection with the proof of a Lemma providing estimates for the norms $\|\Phi^{-1} \mathrm{B}_w u\|_{H_0^1(\Omega,\mathbb{R}^2)}$ and $\|\Phi^{-1} \hat{\mathrm{B}}_w u\|_{H_0^1(\Omega,\mathbb{R}^2)}$ respectively.

Lemma A.11. *For fixed $w \in W(\Omega)$ the following inequalities hold true for all $u \in H(\Omega)$:*

(i) $\|\Phi^{-1} \mathrm{B}_w u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq 2Re\|w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^2)}\|u\|_{L^2(\mathrm{supp}(w),\mathbb{R}^2)}$.

(ii) $\|\Phi^{-1} \hat{\mathrm{B}}_w u\|_{H_0^1(\Omega,\mathbb{R}^2)} \leq Re\left(\|w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^2)} + C_2\|\nabla w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^{2\times 2})}\right)\|u\|_{L^2(\mathrm{supp}(w),\mathbb{R}^2)}$.

Proof. (i) Let $w \in W(\Omega)$, $u, \varphi \in H(\Omega)$. Using the definition (3.2) we obtain

$$(\mathrm{B}_w\, u)[\varphi] = Re \int_\Omega [(\nabla w)u + (\nabla u)w] \cdot \varphi \, \mathrm{d}(x,y).$$

Moreover, Lemma A.10 (ii) and (A.6) imply

$$\int_\Omega [(u \cdot \nabla)w] \cdot \varphi \, \mathrm{d}(x,y) = -\int_\Omega [(u \cdot \nabla)\varphi] \cdot w \, \mathrm{d}(x,y) = -\int_\Omega w^T (\nabla \varphi) u \, \mathrm{d}(x,y),$$

and

$$\int_\Omega [(w \cdot \nabla)u] \cdot \varphi \, \mathrm{d}(x,y) = -\int_\Omega [(w \cdot \nabla)\varphi] \cdot u \, \mathrm{d}(x,y) = -\int_\Omega u^T (\nabla \varphi) w \, \mathrm{d}(x,y).$$

Hence, we get

$$\begin{aligned}(\mathrm{B}_w\, u)[\varphi] &= -Re \int_\Omega [w^T (\nabla\varphi) u + u^T (\nabla\varphi) w] \, \mathrm{d}(x,y) \\ &= -Re \int_\Omega u^T [(\nabla\varphi)^T + (\nabla\varphi)] w \, \mathrm{d}(x,y).\end{aligned}$$

Now, Lemma A.9 (iii) yields

$$|(\mathrm{B}_w\, u)[\varphi]| \le Re \|u\|_{L^2(\mathrm{supp}(w),\mathbb{R}^2)} \|(\nabla\varphi)^T + (\nabla\varphi)\|_{L^2(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^2)}$$

Due to the definition of the L^2-norm (see (2.1)) and (2.6) we get

$$\begin{aligned}\|(\nabla\varphi)^T + (\nabla\varphi)\|_{L^2(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} &\le \|(\nabla\varphi)^T\|_{L^2(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} + \|\nabla\varphi\|_{L^2(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} \\ &= 2\|\nabla\varphi\|_{L^2(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} \le 2\|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}\end{aligned}$$

which, together with the definition of the dual norm and (2.14), implies the assertion.

(ii) Again, let $w \in W(\Omega)$, $u, \varphi \in H(\Omega)$. By (6.2) we obtain

$$(\hat{\mathrm{B}}_w\, u)[\varphi] = Re \int_\Omega \varphi^T [(\nabla w)^T u - (\nabla u) w] \, \mathrm{d}(x,y).$$

We note that due to the transposed derivative $(\nabla w)^T$ in the first term we cannot obtain an analogue result as in part (i) (i.e., with an estimate only depending on $\|w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^2)}$).

Now, applying Lemma A.10 (ii) and (A.6) we obtain

$$\begin{aligned}\int_\Omega \varphi^T (\nabla u) w \, \mathrm{d}(x,y) &= \int_\Omega [(w \cdot \nabla)u] \cdot \varphi \, \mathrm{d}(x,y) \\ &= -\int_\Omega [(w \cdot \nabla)\varphi] \cdot u \, \mathrm{d}(x,y) = -\int_\Omega u^T (\nabla\varphi) w \, \mathrm{d}(x,y)\end{aligned}$$

implying

$$(\hat{\mathrm{B}}_w\, u)[\varphi] = Re \int_\Omega [u^T (\nabla w) \varphi + u^T (\nabla\varphi) w] \, \mathrm{d}(x,y).$$

Hence, Lemma A.9 (ii) and (iii) together with (2.6) and (2.12) imply

$$\begin{aligned}&|(\mathrm{B}_w\, u)[\varphi]| \\ &\quad \le Re \|u\|_{L^2(\mathrm{supp}(w),\mathbb{R}^2)} \|\varphi\|_{L^2(\mathrm{supp}(w),\mathbb{R}^2)} \|\nabla w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} \\ &\qquad + Re \|u\|_{L^2(\mathrm{supp}(w),\mathbb{R}^2)} \|\nabla\varphi\|_{L^2(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} \|w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^2)} \\ &\quad \le Re \left(C_2 \|\nabla w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^{2\times 2})} + \|w\|_{L^\infty(\mathrm{supp}(w),\mathbb{R}^2)} \right) \|u\|_{L^2(\mathrm{supp}(w),\mathbb{R}^2)} \|\varphi\|_{H_0^1(\Omega,\mathbb{R}^2)}.\end{aligned}$$

Thus, again using the definition of the dual norm and (2.14) the assertion follows.

□

A.4 Argyris Reference Shape Functions

For the reference triangle $\hat{\mathcal{T}}$ we consider the functionals $\hat{\mathcal{L}}_1, \dots, \hat{\mathcal{L}}_{21} \colon \mathbb{P}^5(\hat{\mathcal{T}}) \to \mathbb{R}$ representing the degrees of freedom introduced in Section 4.2 for $\hat{\mathcal{T}}$ (cf. Figure 4.6). To calculate the desired Argyris reference shape functions $\hat{\zeta}_1, \dots, \hat{\zeta}_{21} \in \mathbb{P}^5(\hat{\mathcal{T}})$ corresponding to the degrees of freedom $\hat{\mathcal{L}}_1, \dots, \hat{\mathcal{L}}_{21}$ we insert the ansatz

$$\hat{\zeta}_j(\hat{x}, \hat{y}) = \sum_{k=0}^{5} \sum_{l=0}^{k} w_{k,l}^{(j)} \hat{x}^l \hat{y}^{k-l} \quad \text{for all } j = 1, \dots, 21$$

into the duality condition $\hat{\mathcal{L}}_i(\hat{\zeta}_j) = \delta_{i,j}$ for all $i, j = 1, \dots, 21$ which leads to a linear system of equations for the coefficients $w_{k,l}^{(j)}$. This system can be solved using a computer algebra software.

In the following, we list these Argyris shape functions defined on the reference cell $\hat{\mathcal{T}}$:

$$\begin{aligned}
\hat{\zeta}_1(\hat{x}, \hat{y}) &\coloneqq 1 - 10(\hat{x}^3 + \hat{y}^3) + 15(\hat{x}^2 - \hat{y}^2)(\hat{x}^2 - \hat{y}^2) - 6(\hat{x}^5 + \hat{y}^5) + 30\hat{x}^2\hat{y}^2(\hat{x} + \hat{y}), \\
\hat{\zeta}_2(\hat{x}, \hat{y}) &\coloneqq ((((-3\hat{x} + 8)\hat{x} + (\hat{y}^2 - 6))\hat{x} - 10\hat{y}^2(\hat{y} - 1))\hat{x} - (4\hat{y} + 1)(2\hat{y} - 1)(\hat{y} - 1)^2)\hat{x}, \\
\hat{\zeta}_3(\hat{x}, \hat{y}) &\coloneqq ((((-3\hat{y} + 8)\hat{y} + (\hat{x}^2 - 6))\hat{y} - 10\hat{x}^2(\hat{x} - 1))\hat{y} - (4\hat{x} + 1)(2\hat{x} - 1)(\hat{x} - 1)^2)\hat{y}, \\
\hat{\zeta}_4(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{2}\hat{x}^2(1 - \hat{x}^3 + 2\hat{y}^3 - 3(1 - \hat{x})(\hat{x} + \hat{y}^2)), \\
\hat{\zeta}_5(\hat{x}, \hat{y}) &\coloneqq \hat{x}\hat{y}(1 + (\hat{x} + z\hat{y})(-4 + (\hat{x} + \hat{y})(5 - 2(\hat{x} + \hat{y})))), \\
\hat{\zeta}_6(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{2}\hat{y}^2(1 - \hat{y}^3 + 2\hat{x}^3 - 3(1 - \hat{y})(\hat{y} + \hat{x}^2)), \\
\hat{\zeta}_7(\hat{x}, \hat{y}) &\coloneqq \hat{x}^2(10\hat{x} + 6\hat{x}^3 - 15(\hat{y}^2(\hat{y} - 1) + \hat{x}(\hat{x} + \hat{y}^2))), \\
\hat{\zeta}_8(\hat{x}, \hat{y}) &\coloneqq \hat{x}^2(\hat{x}(-4 + \hat{y}(7 - 3\hat{x})) - \tfrac{7}{2}\hat{y}^2(1 - \hat{x} - \hat{y})), \\
\hat{\zeta}_9(\hat{x}, \hat{y}) &\coloneqq (-5 + 2\hat{x}(7 - 4\hat{x}) + \tfrac{37}{2}\hat{y}(1 - \hat{x}) - \tfrac{27}{2}\hat{y}^2)\hat{x}^2\hat{y}, \\
\hat{\zeta}_{10}(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{4}\hat{x}^2(2\hat{x}(1 - \hat{x})^2 + \hat{y}^2(1 - \hat{x} - \hat{y})), \\
\hat{\zeta}_{11}(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{2}(2 + 2\hat{x}(-3 + 2\hat{x}) - 7\hat{y}(1 - \hat{x}) + 5\hat{y}^2)\hat{x}^2\hat{y} \\
\hat{\zeta}_{12}(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{4}\hat{x}^2\hat{y}^2(5 - 3\hat{x} - 5\hat{y}), \\
\hat{\zeta}_{13}(\hat{x}, \hat{y}) &\coloneqq \hat{y}^2(10\hat{y} + 6\hat{y}^3 - 15(\hat{x}^2(\hat{x} - 1) + \hat{y}(\hat{y} + \hat{x}^2))), \\
\hat{\zeta}_{14}(\hat{x}, \hat{y}) &\coloneqq (-5 + 2\hat{y}(7 - 4\hat{y}) + \tfrac{37}{2}\hat{x}(1 - \hat{y}) - \tfrac{27}{2}\hat{x}^2)\hat{y}^2\hat{x}, \\
\hat{\zeta}_{15}(\hat{x}, \hat{y}) &\coloneqq \hat{y}^2(\hat{y}(-4 + \hat{x}(7 - 3\hat{y})) - \tfrac{7}{2}\hat{x}^2(1 - \hat{x} - \hat{y})), \\
\hat{\zeta}_{16}(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{4}\hat{x}^2\hat{y}^2(5 - 3\hat{y} - 5\hat{x}), \\
\hat{\zeta}_{17}(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{2}(2 + 2\hat{y}(-3 + 2\hat{y}) - 7\hat{x}(1 - \hat{y}) + 5\hat{x}^2)\hat{y}^2\hat{x} \\
\hat{\zeta}_{18}(\hat{x}, \hat{y}) &\coloneqq \tfrac{1}{4}\hat{y}^2(2\hat{y}(1 - \hat{y})^2 + \hat{x}^2(1 - \hat{x} - \hat{y})), \\
\hat{\zeta}_{19}(\hat{x}, \hat{y}) &\coloneqq -16\hat{x}^2\hat{y}(1 - \hat{x} - \hat{y})^2, \\
\hat{\zeta}_{20}(\hat{x}, \hat{y}) &\coloneqq -8\sqrt{2}\hat{x}^2\hat{y}^2(1 - \hat{x} - \hat{y}), \\
\hat{\zeta}_{21}(\hat{x}, \hat{y}) &\coloneqq -16\hat{x}\hat{y}^2(1 - \hat{x} - \hat{y})^2
\end{aligned}$$

for all $(\hat{x}, \hat{y}) \in \hat{\mathcal{T}}$.

A.5 Enclosing Ranges via Bernstein Polynomials

In this Section we explain the representation of the local Argyris shape functions on a cell $\mathcal{T}$ by Bernstein polynomials, which, for instance, is needed to compute the ranges of our finite element solution(cf. Section 5.1).

Again, using the notations from Section 4.2 we get the representation $w\big|_{\mathcal{T}} = \sum_{i=1}^{21} w_i^{\mathcal{T}} \xi_i^{\mathcal{T}}$ of a finite element solution $w \in H^1(\Omega_0, \mathbb{R}^2)$ where $w_1^{\mathcal{T}}, \dots, w_{21}^{\mathcal{T}} \in \mathbb{R}$ denote the associated coefficients. Recall that the local finite element basis functions $\xi_1^{\mathcal{T}}, \dots, \xi_{21}^{\mathcal{T}}$ are defined using the gradients of the Argyris shape functions $\zeta_1^{\mathcal{T}}, \dots, \zeta_{21}^{\mathcal{T}}$. Thus, we obtain

$$w\big|_{\mathcal{T}} = \sum_{i=1}^{21} w_i^{\mathcal{T}} \xi_i^{\mathcal{T}} = \sum_{i=1}^{21} w_i^{\mathcal{T}} \begin{pmatrix} -\frac{\partial \zeta_i^{\mathcal{T}}}{\partial y} \\ \frac{\partial \zeta_i^{\mathcal{T}}}{\partial x} \end{pmatrix}. \tag{A.7}$$

To compute the desired Bernstein expansion for the components of our finite element function $w\big|_{\mathcal{T}}$ on $\mathcal{T}$, we exploit the fact that we can represent the derivatives $\frac{\partial \zeta_1^{\mathcal{T}}}{\partial x}, \dots, \frac{\partial \zeta_{21}^{\mathcal{T}}}{\partial x}$ and $\frac{\partial \zeta_1^{\mathcal{T}}}{\partial y}, \dots, \frac{\partial \zeta_{21}^{\mathcal{T}}}{\partial y}$, respectively, by the gradients of the reference functions on $\hat{\mathcal{T}}$ (cf. (9.4)). Therefore, we only need to expand the derivatives (of the Argyris reference shape functions) $\frac{\partial \hat{\zeta}_1^{\mathcal{T}}}{\partial x}, \dots, \frac{\partial \hat{\zeta}_{21}^{\mathcal{T}}}{\partial x}$ and $\frac{\partial \hat{\zeta}_1^{\mathcal{T}}}{\partial y}, \dots, \frac{\partial \hat{\zeta}_{21}^{\mathcal{T}}}{\partial y}$ (defined on the reference triangle $\hat{\mathcal{T}}$) in terms of Bernstein polynomials of degree 4 which are given by

$$p_{i,j}^{(4)}(\hat{x}, \hat{y}) := \binom{4}{i} \binom{4-i}{j} \hat{x}^i \hat{y}^j (1 - \hat{x} - \hat{y})^{4-i-j} \quad \text{for all } (\hat{x}, \hat{y}) \in \hat{\mathcal{T}}.$$

As already mentioned in Section 5.1 we use a reordered set $\left\{p_1^{(4)}, \dots, p_{15}^{(4)}\right\}$ of these polynomials defined as follows

$$\begin{aligned}
&p_1^{(4)} := p_{0,0}^{(4)}, \\
&p_2^{(4)} := p_{1,0}^{(4)}, \quad p_3^{(4)} := p_{0,1}^{(4)}, \\
&p_4^{(4)} := p_{2,0}^{(4)}, \quad p_5^{(4)} := p_{1,1}^{(4)}, \quad p_6^{(4)} := p_{0,2}^{(4)}, \\
&p_7^{(4)} := p_{3,0}^{(4)}, \quad p_8^{(4)} := p_{2,1}^{(4)}, \quad p_9^{(4)} := p_{1,2}^{(4)}, \quad p_{10}^{(4)} := p_{0,3}^{(4)}, \\
&p_{11}^{(4)} := p_{4,0}^{(4)}, \quad p_{12}^{(4)} := p_{3,1}^{(4)}, \quad p_{13}^{(4)} := p_{2,2}^{(4)}, \quad p_{14}^{(4)} := p_{1,3}^{(4)}, \quad p_{15}^{(4)} := p_{0,4}^{(4)}.
\end{aligned} \tag{A.8}$$

Next, we use the definition of the derivatives of the Argyris reference shape functions (see Appendix A.4 and [110, `ArgyrisShapes`]) to compute appropriate transformation matrices $T^x, T^y \in \mathbb{R}^{15 \times 21}$ such that

$$\frac{\partial \hat{\zeta}_k}{\partial \hat{x}} = \sum_{l=1}^{15} T_{l,k}^x p_l^{(4)} \quad \text{and} \quad \frac{\partial \hat{\zeta}_k}{\partial \hat{y}} = \sum_{l=1}^{15} T_{l,k}^y p_l^{(4)} \quad \text{for all } k = 1, \dots, 21,$$

where the transformation matrices are given by

$$T^x := \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-5 & -2 & 0 & -\frac{1}{4} & 0 & 0 & 5 & -2 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -\frac{11}{6} & \frac{1}{4} & -\frac{5}{12} & 0 & 0 & 0 & -\frac{5}{6} & 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{8}{3} & 0 & 0 \\
0 & -\frac{5}{6} & 0 & 0 & -\frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{5}{6} & 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & -\frac{8}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-5 & -2 & \frac{5}{6} & -\frac{1}{4} & \frac{1}{6} & 0 & 5 & -2 & \frac{11}{6} & \frac{1}{4} & -\frac{5}{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{8}{3} & 0 & 0 \\
-5 & \frac{5}{6} & -2 & 0 & \frac{1}{6} & -\frac{1}{4} & \frac{5}{2} & -\frac{7}{12} & \frac{17}{12} & \frac{1}{24} & -\frac{1}{4} & \frac{5}{24} & \frac{5}{2} & \frac{9}{4} & -\frac{7}{12} & \frac{5}{24} & -\frac{5}{12} & \frac{1}{24} & 0 & -\frac{4\sqrt{2}}{3} & \frac{8}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -\frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{5}{2} & -\frac{17}{12} & \frac{7}{12} & \frac{5}{24} & -\frac{1}{4} & \frac{1}{24} & -\frac{5}{2} & -\frac{17}{12} & \frac{7}{12} & -\frac{5}{24} & \frac{1}{4} & -\frac{1}{24} & 0 & \frac{4\sqrt{2}}{3} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{4} & -\frac{1}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

and

$$T^y = \begin{pmatrix}
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\frac{5}{6} & 0 & -\frac{1}{6} & 0 & 0 & 0 & -\frac{5}{6} & 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{8}{3} & 0 & 0 \\
0 & -\frac{11}{6} & 1 & 0 & -\frac{5}{12} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{5}{6} & 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & -\frac{8}{3} \\
-5 & 0 & -2 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 5 & 0 & -2 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-5 & -2 & \frac{5}{6} & -\frac{1}{4} & \frac{1}{6} & 0 & \frac{5}{2} & -\frac{7}{12} & \frac{9}{4} & \frac{1}{24} & -\frac{5}{12} & \frac{5}{24} & \frac{5}{2} & \frac{17}{12} & -\frac{7}{12} & \frac{5}{24} & -\frac{1}{4} & \frac{1}{24} & \frac{8}{3} & -\frac{4\sqrt{2}}{3} & 0 \\
-5 & \frac{5}{6} & -2 & 0 & \frac{1}{6} & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 5 & \frac{11}{6} & -2 & 0 & -\frac{5}{12} & \frac{1}{4} & 0 & 0 & \frac{8}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -\frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -\frac{5}{2} & \frac{7}{12} & -\frac{17}{12} & -\frac{1}{24} & \frac{1}{4} & -\frac{5}{24} & \frac{5}{2} & \frac{7}{12} & -\frac{17}{12} & \frac{1}{24} & -\frac{1}{4} & \frac{5}{24} & 0 & \frac{4\sqrt{2}}{3} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{4} & -\frac{1}{4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$

Using (A.7) and the representation formulas (9.4) again, we obtain

$$w_1(\Phi_{\mathcal{T}}(\hat{x},\hat{y})) = \sum_{l=1}^{15} \left(-\sum_{i,k=1}^{21} w_i^{\mathcal{T}} C_{k,i}^{\mathcal{T}} \left[T_{l,k}^x F_{1,2}^{\mathcal{T}} + T_{l,k}^y F_{2,2}^{\mathcal{T}} \right] \right) p_l^{(4)}(\hat{x},\hat{y}),$$

$$w_2(\Phi_{\mathcal{T}}(\hat{x},\hat{y})) = \sum_{l=1}^{15} \left(\sum_{i,k=1}^{21} w_i^{\mathcal{T}} C_{k,i}^{\mathcal{T}} \left[T_{l,k}^x F_{1,1}^{\mathcal{T}} + T_{l,k}^y F_{2,1}^{\mathcal{T}} \right] \right) p_l^{(4)}(\hat{x},\hat{y})$$

for all $(\hat{x},\hat{y}) \in \hat{\mathcal{T}}$. Thus, the coefficients needed in Section 5.1 to enclose the range on the cell $\mathcal{T}$ are now given by the terms on the right-hand side.

The same techniques can also be applied to the derivative

$$(\nabla w)\big|_{\mathcal{T}} = \sum_{i=1}^{21} w_i^{\mathcal{T}} \nabla \xi_i^{\mathcal{T}} = \sum_{i=1}^{21} w_i^{\mathcal{T}} \begin{pmatrix} -\frac{\partial^2 \zeta_i^{\mathcal{T}}}{\partial x \partial y} & -\frac{\partial^2 \zeta_i^{\mathcal{T}}}{\partial y^2} \\ \frac{\partial^2 \zeta_i^{\mathcal{T}}}{\partial x^2} & \frac{\partial^2 \zeta_i^{\mathcal{T}}}{\partial x \partial y} \end{pmatrix}.$$

Hence, we need to express the second derivatives of the Argyris shape functions on the reference triangle by Bernstein polynomials. Since the second derivatives are of polynomial degree at most 3 we need the set $\left\{p_1^{(3)},\dots,p_{10}^{(3)}\right\}$ (note that $\dim \mathbb{P}^3(\hat{\mathcal{T}}) = 10$) of Bernstein polynomials which are sorted similar as above (where the order 4 is replaced by 3 and the last line is omitted in (A.8)).

Hence, we compute transformation matrices $T^{xx}, T^{xy}, T^{yy} \in \mathbb{R}^{10\times 21}$ such that

$$\frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x}^2} = \sum_{l=1}^{10} T^{xx}_{l,k} p_l^{(3)}, \quad \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{x}\partial \hat{y}} = \sum_{l=1}^{10} T^{xy}_{l,k} p_l^{(3)} \quad \text{and} \quad \frac{\partial^2 \hat{\zeta}_k}{\partial \hat{y}^2} = \sum_{l=1}^{10} T^{yy}_{l,k} p_l^{(3)}$$

for all $k = 1,\dots,21$, where

$$T^{xx} = \begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-20 & -12 & 0 & -2 & 0 & 0 & 20 & -8 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{22}{3} & 1 & -\frac{8}{3} & 0 & 0 & 0 & -\frac{10}{3} & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{32}{3} & 0 & 0 \\
20 & 8 & 0 & 1 & 0 & 0 & -20 & 12 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-20 & -12 & \frac{32}{3} & -2 & \frac{7}{3} & 0 & 20 & -8 & \frac{32}{3} & 1 & -\frac{7}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{64}{3} & 0 & 0 \\
-20 & \frac{20}{3} & -8 & 0 & \frac{4}{3} & -1 & 10 & -\frac{7}{3} & \frac{17}{3} & \frac{1}{6} & -1 & \frac{5}{6} & 10 & \frac{37}{3} & -\frac{7}{3} & \frac{5}{6} & -\frac{7}{3} & \frac{1}{6} & 0 & -\frac{16\sqrt{2}}{3} & \frac{64}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
20 & 8 & -\frac{10}{3} & 1 & -\frac{2}{3} & 0 & -20 & 12 & -\frac{22}{3} & -2 & \frac{8}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{32}{3} & 0 & 0 \\
20 & -\frac{10}{3} & 8 & 0 & -\frac{2}{3} & 1 & 0 & -\frac{10}{3} & -\frac{10}{3} & \frac{2}{3} & 0 & -\frac{2}{3} & -20 & -\frac{44}{3} & \frac{14}{3} & -\frac{5}{3} & \frac{8}{3} & -\frac{1}{3} & 0 & \frac{32\sqrt{2}}{3} & -\frac{32}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{pmatrix},$$

$$T^{xy} = \begin{pmatrix}
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\frac{22}{3} & 0 & -\frac{5}{3} & 0 & 0 & 0 & -\frac{10}{3} & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{32}{3} & 0 & 0 \\
0 & -\frac{22}{3} & 0 & 0 & -\frac{5}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{10}{3} & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & -\frac{32}{3} \\
0 & 0 & \frac{10}{3} & 0 & \frac{2}{3} & 0 & 0 & 0 & \frac{22}{3} & 0 & -\frac{5}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{32}{3} & 0 & 0 \\
-20 & -\frac{2}{3} & -\frac{2}{3} & -1 & \frac{7}{3} & -1 & 10 & -\frac{7}{3} & 9 & \frac{1}{6} & -\frac{5}{3} & \frac{5}{6} & 10 & 9 & -\frac{7}{3} & \frac{5}{6} & -\frac{5}{3} & \frac{1}{6} & \frac{32}{3} & -\frac{16\sqrt{2}}{3} & \frac{32}{3} \\
0 & \frac{10}{3} & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{22}{3} & 0 & 0 & -\frac{5}{3} & 0 & 0 & 0 & \frac{32}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
20 & 8 & -\frac{10}{3} & 1 & -\frac{2}{3} & 0 & -10 & \frac{7}{3} & -5 & -\frac{1}{6} & \frac{2}{3} & \frac{1}{6} & -10 & -\frac{17}{3} & \frac{7}{3} & -\frac{5}{6} & 1 & -\frac{1}{6} & -\frac{32}{3} & \frac{16\sqrt{2}}{3} & 0 \\
20 & -\frac{10}{3} & 8 & 0 & -\frac{2}{3} & 1 & -10 & \frac{7}{3} & -\frac{17}{3} & -\frac{1}{6} & 1 & -\frac{5}{6} & -10 & -5 & \frac{7}{3} & \frac{1}{6} & \frac{2}{3} & -\frac{1}{6} & 0 & \frac{16\sqrt{2}}{3} & -\frac{32}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0
\end{pmatrix}$$

and

$$T^{yy} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\frac{22}{3} & 0 & 0 & -\frac{8}{3} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{10}{3} & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & -\frac{32}{3} \\
-20 & 0 & -12 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 20 & 0 & -8 & 0 & 0 & 1 & 0 & 0 & 0 \\
-20 & -8 & \frac{20}{3} & -1 & \frac{4}{3} & 0 & 10 & -\frac{7}{3} & \frac{37}{3} & \frac{1}{6} & -\frac{7}{3} & \frac{5}{6} & 10 & \frac{17}{3} & -\frac{7}{3} & \frac{5}{6} & -1 & \frac{1}{6} & \frac{64}{3} & -\frac{16\sqrt{2}}{3} & 0 \\
-20 & \frac{32}{3} & -12 & 0 & \frac{7}{3} & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 20 & \frac{32}{3} & -8 & 0 & -\frac{7}{3} & 1 & 0 & 0 & \frac{64}{3} \\
20 & 0 & 8 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -20 & 0 & 12 & 0 & 0 & -2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
20 & 8 & -\frac{10}{3} & 1 & -\frac{2}{3} & 0 & -20 & \frac{14}{3} & -\frac{44}{3} & -\frac{1}{3} & \frac{8}{3} & -\frac{5}{3} & 0 & -\frac{10}{3} & -\frac{10}{3} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{32}{3} & \frac{32\sqrt{2}}{3} & 0 \\
20 & -\frac{10}{3} & 8 & 0 & -\frac{2}{3} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -20 & -\frac{22}{3} & 12 & 0 & \frac{8}{3} & -2 & 0 & 0 & -\frac{32}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}.$$

Finally, using the representation formulas for the second derivatives given in (9.5) we obtain the coefficients (corresponding to the Bernstein basis) to enclose the range of the derivative ∇w, i.e., we obtain the following coefficients associated to the Bernstein basis

$$\begin{aligned}&(\nabla w)_{1,2}(\Phi_{\mathcal{T}}(\hat{x},\hat{y}))\\ &\quad=\sum_{l=1}^{10}\Bigg(-\sum_{i,k=1}^{21} w_i^{\mathcal{T}} C_{k,i}^{\mathcal{T}}\big[T_{l,k}^{xx}(F_{1,2}^{\mathcal{T}})^2+T_{l,k}^{xy}2F_{1,2}^{\mathcal{T}}F_{2,2}^{\mathcal{T}}+T_{l,k}^{yy}(F_{2,2}^{\mathcal{T}})^2\big]\Bigg)p_l^{(3)}(\hat{x},\hat{y}),\end{aligned}$$

$$\begin{aligned}&(\nabla w)_{2,1}(\Phi_{\mathcal{T}}(\hat{x},\hat{y}))\\ &\quad=\sum_{l=1}^{10}\Bigg(\sum_{i,k=1}^{21} w_i^{\mathcal{T}} C_{k,i}^{\mathcal{T}}\big[T_{l,k}^{xx}(F_{1,1}^{\mathcal{T}})^2+T_{l,k}^{xy}2F_{2,1}^{\mathcal{T}}F_{2,1}^{\mathcal{T}}+T_{l,k}^{yy}(F_{2,1}^{\mathcal{T}})^2\big]\Bigg)p_l^{(3)}(\hat{x},\hat{y}),\end{aligned}$$

$$\begin{aligned}&(\nabla w)_{2,2}(\Phi_{\mathcal{T}}(\hat{x},\hat{y}))\\ &\quad=\sum_{l=1}^{10}\Bigg(\sum_{i,k=1}^{21} w_i^{\mathcal{T}} C_{k,i}^{\mathcal{T}}\big[T_{l,k}^{xx}F_{1,1}^{\mathcal{T}}F_{1,2}^{\mathcal{T}}+T_{l,k}^{xy}(F_{1,1}^{\mathcal{T}}F_{2,2}^{\mathcal{T}}+F_{1,2}^{\mathcal{T}}F_{2,1}^{\mathcal{T}})+T_{l,k}^{yy}F_{2,1}^{\mathcal{T}}F_{2,2}^{\mathcal{T}}\big]\Bigg)p_l^{(3)}(\hat{x},\hat{y})\end{aligned}$$

for all $(\hat{x},\hat{y})\in\hat{\mathcal{T}}$.

A.6 Auxiliary Computations for the Eigenvalue Homotopy

In the further course we present some results needed for our eigenvalue homotopy. First, we prove a Lemma which provides a statement needed for the computation of the constant γ_1 and γ_2 respectively appearing in our extended coefficient homotopy (cf. Section 6.2.1.2).

Lemma A.12. *Let a, b, c be positive constants and $I := (0, \frac{b}{c})$. Then the function $g\colon I \to \mathbb{R}$, $g(t) := \frac{1}{t} + \frac{a}{b-ct}$ has exactly one minimum at $t_1 := \frac{b}{c+\sqrt{ac}} \in I$ with value $g(t_1) = \frac{(\sqrt{a}+\sqrt{c})^2}{b}$.*

Proof. First, from the definition of g we see

$$\lim_{t\to 0+} g(t) = \infty \quad \text{and} \quad \lim_{t\to \frac{b}{c}-} g(t) = \infty.$$

Moreover, since g is twice continuous differentiable on I we get the derivatives

$$g'(t) = -\frac{1}{t^2} + \frac{ac}{(b-ct)^2} \quad \text{and} \quad g''(t) = \frac{2}{t^3} + \frac{2ac^2}{(b-ct)^3} \quad \text{for all } t \in I.$$

Next, we solve

$$g'(t) = 0 \quad \Leftrightarrow \quad \frac{1}{t^2} = \frac{ac}{(b-ct)^2} \quad \Leftrightarrow \quad \sqrt{ac}t = |b-ct|$$

which in the case $a \neq c$ results in the two roots $t_1 = \frac{b}{c+\sqrt{ac}} < \frac{b}{c}$ and $t_2 = \frac{b}{c-\sqrt{ac}} > \frac{b}{c}$ where only the first one t_1 is contained in our interval I. In the case $a = c$ we directly obtain the single root $t_1 = \frac{b}{2c} < \frac{b}{c}$ in the interval I.

Since $t_1 < \frac{b}{c}$ we conclude $b - ct_1 > 0$ implying $g''(t_2) > 0$, i.e., $t_1 \in I$ is our desired minimum. Furthermore, a simple calculation shows

$$g(t_1) = \frac{c + \sqrt{ac}}{b} + \frac{a}{b - \frac{bc}{c+\sqrt{ac}}} = \frac{1}{b}\left(c + \sqrt{ac} + \frac{a(c+\sqrt{ac})}{\sqrt{ac}}\right) = \frac{(\sqrt{a}+\sqrt{c})^2}{b}.$$

□

Next, we shortly introduce the computation of the integrals appearing in the definition of the multiplication operator A (cf. (6.86) and (6.87)).

Lemma A.13. *For φ_n ($n \in \mathbb{N}$) introduced in (6.84) the following identity holds true*

$$\int_0^1 \varphi_n(y)^T \begin{pmatrix} 0 & 1-2y \\ 1-2y & 0 \end{pmatrix} \varphi_m(y)\,\mathrm{d}y = \begin{cases} \frac{16\lceil\frac{n}{2}\rceil\lceil\frac{m}{2}\rceil}{\pi^2\left(\lceil\frac{n}{2}\rceil^2 - \lceil\frac{m}{2}\rceil^2\right)^2}, & (n+m) \bmod 4 = 1, \\ 0, & (n+m) \bmod 4 \neq 1 \end{cases}$$

for all $n, m \in \mathbb{N}$.

Proof. As an abbreviation we denote the integral on the left-hand side of the assertion by $\mathcal{I}_{n,m}$, Moreover, recalling the definition of the basis functions φ_n we can distinguish the following cases:

- $\underline{n = 2k-1,\ m = 2l-1 \text{ or } n = 2k,\ m = 2l}$: In both cases the matrix vector product appearing in the integral reduces to zero which leads to $\mathcal{I}_{n,m} = 0$. Furthermore, we calculate

$$(2k - 1 + 2l - 1) \bmod 4 = 2(k + l - 1) \bmod 4 \in \{0, 2\}$$

and

$$(2k + 2l) \bmod 4 = 2(k + l) \bmod 4 \in \{0, 2\}$$

implying $(n + m) \bmod 4 \in \{0, 2\}$ which finally proves the assertion in these cases.

- $\underline{n = 2k-1,\ m = 2l \text{ or } n = 2k,\ m = 2l-1}$: Now, the matrix vector product reads as

$$\varphi_n(y)^T \begin{pmatrix} 0 & 1-2y \\ 1-2y & 0 \end{pmatrix} \varphi_m(y) = 2(1-2y)\sin(k\pi y)\sin(l\pi y)$$

for all $y \in [0, 1]$. Thus, we can calculate the integral and obtain

$$\begin{aligned} \mathcal{I}_{n,m} &= 2\int_0^1 (1-2y)\sin(k\pi y)\sin(l\pi y)\,\mathrm{d}y \\ &= \begin{cases} \frac{8kl(1-(-1)^{k+l})}{\pi^2(k^2-l^2)^2}, & k \neq l, \\ 0, & k = l, \end{cases} \\ &= \begin{cases} \frac{16kl}{\pi^2(k^2-l^2)^2}, & (k+l) \bmod 2 = 1, \\ 0, & (k+l) \bmod 2 = 0. \end{cases} \end{aligned}$$

In addition to that we get $n+m = 2(k+l)-1$ in both cases. Hence, if $(k+l) \bmod 2 = 1$ holds true we obtain $(n+m) \bmod 4 = 1$. Otherwise, for $(k+l) \bmod 2 = 0$ we calculate $(n + m) \bmod 4 = 3$ proving the assertion in these cases as well.

□

Finally, we calculate values of selected series needed in the computation of the functions θ_2 and θ_3 in Section 6.2.1.4.

Lemma A.14. *The following identities hold true for all $k \in \mathbb{N}$:*

(i) $\sum\limits_{\substack{l=1\\l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{k^2-l^2} = -\frac{1-(-1)^k}{2k^2}$,

(ii) $\sum\limits_{\substack{l=1\\l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{(k^2-l^2)^2} = \frac{\pi^2}{8k^2} - \frac{1-(-1)^k}{2k^4}$,

(iii) $\sum\limits_{\substack{l=1\\l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{(k^2-l^2)^3} = \frac{3\pi^2}{32k^4} - \frac{1-(-1)^k}{2k^6}$,

(iv) $\sum\limits_{\substack{l=1\\l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{(k^2-l^2)^4} = \frac{\pi^4}{384k^4} + \frac{5\pi^2}{64k^6} - \frac{1-(-1)^k}{2k^8}$,

(v) $\sum\limits_{\substack{l=1\\l\neq k}}^{\infty} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} = \frac{\pi^4}{384k^2} - \frac{\pi^2}{64k^4}$.

Proof. (i) We use the Fourier series expansion of $s \mapsto \sin(k|s|)$, where $k \in \mathbb{N}$ is an arbitrary natural number. Therefore, for all $l \in \mathbb{Z}$ with $l \neq \pm k$ we calculate

$$c_l = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \sin(k|s|)\mathrm{e}^{\mathrm{i}ls}\,\mathrm{d}s = \frac{2k}{\sqrt{2\pi}} \cdot \frac{1-(-1)^{k+l}}{k^2-l^2}.$$

Moreover, we get $c_{\pm k} = 0$. Since $s \mapsto \sin(k|s|)$ is continuous and (for some fixed $\delta > 0$) of bounded variation on $[s-\delta, s+\delta]$ for all $s \in \mathbb{R}$, [7, Theorem 15-18] implies that its fourier series converges to $\sin(k|\cdot|)$ pointwise on $\mathbb{R}$. Hence, we obtain

$$\begin{aligned}\sin(k|s|) &= \sum_{l=-\infty}^{\infty} c_l \frac{\mathrm{e}^{-\mathrm{i}ls}}{\sqrt{2\pi}} = \frac{c_0}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} \sum_{l=1}^{\infty} \left(c_l \mathrm{e}^{-\mathrm{i}ls} + c_{-l}\mathrm{e}^{\mathrm{i}ls}\right)\\ &= \frac{1-(-1)^k}{k\pi} + \frac{k}{\pi} \sum_{\substack{l=1\\l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{k^2-l^2}(\mathrm{e}^{-\mathrm{i}ls} + \mathrm{e}^{\mathrm{i}ls})\\ &= \frac{1-(-1)^k}{k\pi} + \frac{2k}{\pi} \sum_{\substack{l=1\\l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{k^2-l^2}\cos(ls)\end{aligned}$$

which (for $s = 0$) implies

$$\sum_{\substack{l=1\\l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{k^2-l^2} = \frac{\pi}{2k}\left(-\frac{1-(-1)^k}{k\pi}\right) = -\frac{1-(-1)^k}{2k^2}.$$

(ii) First, for all $k, l \in \mathbb{N}$ with $l \neq k$ we have the identity

$$\frac{1}{k-l} + \frac{1}{k+l} = \frac{2k}{(k-l)(k+l)} = \frac{2k}{k^2-l^2}. \tag{A.9}$$

Thus, we obtain

$$\begin{aligned}\sum_{\substack{l=1\\l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^2} &= \frac{1}{4k^2}\sum_{\substack{l=1\\l\neq k}}^{\infty}(1-(-1)^{k+l})\left(\frac{1}{k-l}+\frac{1}{k+l}\right)^2\\ &= \frac{1}{4k^2}\left(\sum_{\substack{l=1\\l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k-l)^2}+2\sum_{\substack{l=1\\l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{k^2-l^2}+\sum_{\substack{l=1\\l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k+l)^2}\right)\end{aligned} \tag{A.10}$$

Splitting the first sum into two parts and introducing index shifts ($j = k - l$ and $j = l - k$) yields

$$\begin{aligned}\sum_{\substack{l=1\\l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k-l)^2} &= \sum_{l=1}^{k-1}\frac{1-(-1)^{k+l}}{(k-l)^2}+\sum_{l=k+1}^{\infty}\frac{1-(-1)^{k+l}}{(k-l)^2}\\ &= \sum_{j=1}^{k-1}\frac{1-(-1)^{j}}{j^2}+\sum_{j=1}^{\infty}\frac{1-(-1)^{j}}{(-j)^2}.\end{aligned}$$

Since $1-(-1)^{k+k}=0$ we can omit $l\neq k$ in the third sum in (A.10) and an additional index shift ($j = k + l$) implies

$$\sum_{\substack{l=1\\l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k+l)^2} = \sum_{l=1}^{\infty}\frac{1-(-1)^{k+l}}{(k+l)^2} = \sum_{j=k+1}^{\infty}\frac{1-(-1)^{j}}{j^2}.$$

Thus, using the results form above together with part (i) we obtain

$$\begin{aligned}\sum_{\substack{l=1\\l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^2} &= \frac{1}{4k^2}\left(2\sum_{j=1}^{\infty}\frac{1-(-1)^{j}}{j^2}-\frac{1-(-1)^{k}}{k^2}-2\frac{1-(-1)^{k}}{2k^2}\right)\\ &= \frac{1}{2k^2}\left(\sum_{j=1}^{\infty}\frac{1-(-1)^{j}}{j^2}-\frac{1-(-1)^{k}}{k^2}\right).\end{aligned}$$

The well-known values $\sum_{j=1}^{\infty}\frac{1}{j^2}=\frac{\pi^2}{6}$ and $\sum_{j=1}^{\infty}\frac{(-1)^j}{j^2}=-\frac{\pi^2}{12}$ (cf. [36, Sections 0.233 and 0.234]) yield the identity $\sum_{j=1}^{\infty}\frac{1-(-1)^j}{j^2}=\frac{\pi^2}{4}$ which (inserted into the previous equation) proves the assertion.

(iii) Using the identity (for $l\neq k$)

$$\frac{1}{(k-l)^2(k+l)}+\frac{1}{(k-l)(k+l)^2}=\frac{2k}{(k-l)^2(k+l)^2}=\frac{2k}{(k^2-l^2)^2}, \tag{A.11}$$

applying similar arguments as before and inserting the result from part (ii) we obtain

$$\begin{aligned}\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^3} &= \frac{1}{8k^3}\sum_{\substack{l=1\\ l\neq k}}^{\infty}(1-(-1)^{k+l})\left(\frac{1}{k-l}+\frac{1}{k+l}\right)^3\\ &= \frac{1}{8k^3}\left(\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k-l)^3}+6k\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^2}+\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k+l)^3}\right)\\ &= \frac{1}{8k^3}\left(\sum_{j=1}^{k-1}\frac{1-(-1)^j}{j^3}+\sum_{j=1}^{\infty}\frac{1-(-1)^j}{(-j)^3}+6k\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^2}\right.\\ &\qquad\left.+\sum_{j=k+1}^{\infty}\frac{1-(-1)^j}{j^3}\right)\\ &= \frac{1}{8k^3}\left(-\frac{1-(-1)^k}{k^3}+6k\left(\frac{\pi^2}{8k^2}-\frac{1-(-1)^k}{2k^4}\right)\right),\end{aligned}$$

where in the first step we used (A.9) and in the second step (A.11). Finally, the assertion follows.

(iv) First, using (A.9) and (A.11) we calculate

$$\begin{aligned}\frac{1}{(k-l)^3(k+l)}+\frac{1}{(k-l)(k+l)^3} &= \frac{1}{2k}\left(\frac{1}{k-l}+\frac{1}{k+l}\right)\left(\frac{1}{(k-l)^2}+\frac{1}{(k+l)^2}\right)\\ &= \frac{1}{2k}\left(\frac{1}{(k-l)^3}+\frac{1}{(k+l)^3}\right)+\frac{1}{(k^2-l^2)^2}\end{aligned}$$

for $l \neq k$. Hence, we obtain

$$\begin{aligned}&\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^4}\\ &\quad= \frac{1}{16k^4}\sum_{\substack{l=1\\ l\neq k}}^{\infty}(1-(-1)^{k+l})\left(\frac{1}{k-l}+\frac{1}{k+l}\right)^4\\ &\quad= \frac{1}{16k^4}\left(\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k-l)^4}+\frac{2}{k}\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k-l)^3}+10\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^2}\right.\\ &\qquad\left.+\frac{2}{k}\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k+l)^3}+\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k+l)^4}\right).\end{aligned}$$

Applying the same index shifts as above and using part (ii) yields

$$\begin{aligned}\sum_{\substack{l=1\\ l\neq k}}^{\infty}\frac{1-(-1)^{k+l}}{(k^2-l^2)^4} &= \frac{1}{16k^4}\left(2\sum_{j=1}^{\infty}\frac{1-(-1)^j}{j^4}-\frac{1-(-1)^k}{k^4}-\frac{2}{k}\cdot\frac{1-(-1)^k}{k^3}\right.\\ &\qquad\left.+10\left(\frac{\pi^2}{8k^2}-\frac{1-(-1)^k}{2k^4}\right)\right).\end{aligned}$$

Using the well-known identities $\sum_{j=1}^{\infty} \frac{1}{j^4} = \frac{\pi^4}{90}$ and $\sum_{j=1}^{\infty} \frac{(-1)^j}{j^4} = -\frac{7\pi^4}{720}$ we conclude $\sum_{j=1}^{\infty} \frac{1-(-1)^j}{j^4} = \frac{\pi^4}{48}$. Inserting this result into the previous equation implies the assertion.

(v) For $l \neq k$ we calculate

$$\begin{aligned}\frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} &= \frac{(l^2-k^2)(1-(-1)^{k+l})}{(k^2-l^2)^4} + \frac{k^2(1-(-1)^{k+l})}{(k^2-l^2)^4} \\ &= -\frac{1-(-1)^{k+l}}{(k^2-l^2)^3} + k^2\frac{1-(-1)^{k+l}}{(k^2-l^2)^4}\end{aligned}$$

which implies

$$\sum_{\substack{l=1\\ l\neq k}}^{\infty} \frac{l^2(1-(-1)^{k+l})}{(k^2-l^2)^4} = -\sum_{\substack{l=1\\ l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{(k^2-l^2)^3} + k^2 \sum_{\substack{l=1\\ l\neq k}}^{\infty} \frac{1-(-1)^{k+l}}{(k^2-l^2)^4}.$$

Finally, inserting the results from part (iii) and (iv) proves the assertion.

□

List of Symbols

Notation	Description	Page
S	Infinite strip $\mathbb{R} \times (0,1)$	13
S_R	"Bounded strip" $(-R,R) \times (0,1)$	86
D	Compact obstacle $D \subseteq \overline{S}$	13
Ω	Domain $S \setminus D$	13
Ω_0	Computational domain	32
Ω_R	Bounded part of the domain $\Omega \cap S_R$	86
Re	Reynolds number	14
d_1, d_2, d_3	Constants describing the obstacle	13
U	Poiseuille flow	15
P	Pressure corresponding to the Poiseuille flow	15
V	Function to avoid the splitted boundary conditions	16
Γ	Difference of Poiseuille flow and V	16
g	Transformed forcing term	17
d_0	Constant in the definition on V	43
$H(\Omega)$	Divergence-free subspace of $H_0^1(\Omega, \mathbb{R}^2)$	17
$W(\Omega)$	Divergence-free subspace of $W^{1,\infty}(\Omega, \mathbb{R}^2)$	30
$\mathcal{L}(\Omega)$	Space for the pressure	125
$\mathcal{H}(\Omega)$	Subspace of compactly supported functions of $H_0^1(\Omega, \mathbb{R}^2)$	125
$H(\mathrm{div}, \Omega, \mathbb{R}^{2\times 2})$	Subspace of $L^2(\Omega, \mathbb{R}^{2\times 2})$ with row-wise divergence in $L^2(\Omega, \mathbb{R}^2)$	21
$\Vert \cdot \Vert_{\mathcal{B}}$	Operator norm	21
$\Vert \cdot \Vert_{L^p(\Omega,\mathbb{R}^2)}$	Norm on $L^p(\Omega, \mathbb{R}^2)$	19
$\Vert \cdot \Vert_{L^p(\Omega,\mathbb{R}^{2\times 2})}$	Norm on $L^p(\Omega, \mathbb{R}^{2\times 2})$	20
$\Vert \cdot \Vert_{H_0^1(\Omega,\mathbb{R}^2)}$	Norm on $H_0^1(\Omega, \mathbb{R}^2)$ and $H(\Omega)$	21
C_p, C_2, C_4	Sobolev's embedding constants	22
Φ	Isometric isomorphism	23
F	Zero-finding operator	31
$\tilde{\omega}$	Approximate solution to $\mathrm{F}\, u = 0$	32
ω	Transformed approximate solution	32

Notation	**Description**	**Page**
B	Bilinear form representing the non-linear part of F	30
$\mathrm{L}_{U+\omega}$	Linearization of F at $\tilde{\omega}$	32
$\mathrm{B}_{U+\omega}$	Second part of the linearization	30
$\hat{\mathrm{L}}_{U+\omega}$	Counterpart to $\mathrm{L}_{U+\omega}$ in the adjoint setting	66
$\hat{\mathrm{B}}_{U+\omega}$	Counterpart to $\hat{\mathrm{B}}_{U+\omega}$ in the adjoint setting	65
δ	Defect bound	33
K	Norm bound	33
K^*	"Adjoint" norm bound	34
$\mathcal{M}$	Finite element mesh	49
$\hat{\mathcal{T}}$	Reference triangle	49
$\mathcal{T}$	Triangular cell	49
$\Phi_{\mathcal{T}}$	Affine linear transformation from $\hat{\mathcal{T}}$ to $\mathcal{T}$	49
$\zeta_1^{\mathcal{T}}, \ldots, \zeta_{21}^{\mathcal{T}}$	Argyris shape functions on $\mathcal{T}$	50
$\xi_1^{\mathcal{T}}, \ldots, \xi_{21}^{\mathcal{T}}$	Divergence-free shape functions on $\mathcal{T}$	50

List of Figures

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[3] G. Alefeld and J. Herzberger. *Einführung in die Intervallrechnung*. Bibliographisches Institut, Mannheim-Vienna-Zurich, 1974. Reihe Informatik, Band 12.

[4] C. J. Amick. Steady solutions of the Navier-Stokes equations in unbounded channels and pipes. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 4(3):473–513, 1977.

[5] C. J. Amick. Properties of steady Navier-Stokes solutions for certain unbounded channels and pipes. *Nonlinear Anal.*, 2(6):689–720, 1978.

[6] C. J. Amick and L. E. Fraenkel. Steady solutions of the Navier-Stokes equations representing plane flow in channels of various types. *Acta Math.*, 144(1-2):83–151, 1980.

[7] T. M. Apostol. *Mathematical analysis : a modern approach to advanced calculus*. Addison-Wesley series in mathematics. Addison-Wesley, Reading, Mass., 1957.

[8] R. G. Bartle. Newton's method in Banach spaces. *Proc. Amer. Math. Soc.*, 6:827–831, 1955.

[9] H. Behnke and F. Goerisch. Inclusions for eigenvalues of selfadjoint problems. In *Topics in validated computations (Oldenburg, 1993)*, volume 5 of *Stud. Comput. Math.*, pages 277–322. North-Holland, Amsterdam, 1994.

[10] D. Boffi, F. Brezzi, and M. Fortin, editors. *Mixed finite element methods and applications*. Springer Series in Computational Mathematics. Springer, Berlin, Heidelberg, 2013.

[11] K. Boukir, Y. Maday, B. Métivet, and E. Razafindrakoto. A high-order characteristics/finite element method for the incompressible Navier-Stokes equations. *Internat. J. Numer. Methods Fluids*, 25(12):1421–1454, 1997.

[12] D. Braess. *Finite Elemente : Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Lehrbuch eBook Collection. Springer Spektrum, Berlin, Heidelberg, 5. aufl. 2013 edition, 2013.

[13] S. C. Brenner and L. R. Scott, editors. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer New York, New York, 2008.

[14] B. Breuer, J. Horák, P. McKenna, and M. Plum. A computer-assisted existence and multiplicity proof for travelling waves in a nonlinear supported beam. In *Journal of Differential Equations*, volume 224, pages 60–97. Elsevier, 2006.

[15] T. Buckmaster and V. Vicol. Nonuniqueness of weak solutions to the Navier-Stokes equation. *Ann. Math. (2)*, 189(1):101–144, 2019.

[16] M. Cannone. A generalization of a theorem by Kato on Navier-Stokes equations. *Rev. Mat. Iberoamericana*, 13(3):515–541, 1997.

[17] M. A. Case, V. J. Ervin, A. Linke, and L. G. Rebholz. A connection between Scott-Vogelius and grad-div stabilized Taylor-Hood FE approximations of the Navier-Stokes equations. *SIAM J. Numer. Anal.*, 49(4):1461–1481, 2011.

[18] A. J. Chorin. On the convergence of discrete approximations to the Navier-Stokes equations. *Math. Comp.*, 23:341–353, 1969.

[19] A. J. Chorin and J. E. Marsden. *A mathematical introduction to fluid mechanics*, volume 4 of *Texts in Applied Mathematics*. Springer-Verlag, New York, third edition, 1993.

[20] J. K. Cullum and R. A. Willoughby. *Lanczos algorithms for large symmetric eigenvalue computations. Vol. 1*, volume 41 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Theory, Reprint of the 1985 original [Birkhäuser Boston, Boston, MA; MR0808962 (87h:65064a)].

[21] D. Daney, G. Hanrot, V. Lefèvre, F. Rouillier, and P. Zimmermann. The mpfr library, 2001. `http://www.mpfr.org`.

[22] V. Domínguez and F.-J. Sayas. Algorithm 884: a simple Matlab implementation of the Argyris element. *ACM Trans. Math. Software*, 35(2):Art. 16, 11, 2009.

[23] D. A. Dunavant. High degree efficient symmetrical Gaussian quadrature rules for the triangle. *Internat. J. Numer. Methods Engrg.*, 21(6):1129–1148, 1985.

[24] F. Durst. *Grundlagen der Strömungsmechanik: eine Einführung in die Theorie der Strömungen von Fluiden*. Springer, Berlin, 2006.

[25] V. J. Ervin. Computational bases for RT_k and BDM_k on triangles. *Comput. Math. Appl.*, 64(8):2765–2774, 2012.

[26] L. C. Evans. *Partial Differential Equations*, volume 19. AMS, Providence, Rhode Island, graduate studies in mathematics edition, 1998.

[27] R. Farwig, G. P. Galdi, and H. Sohr. A new class of weak solutions of the Navier-Stokes equations with nonhomogeneous data. *J. Math. Fluid Mech.*, 8(3):423–444, 2006.

[28] R. Farwig, G. P. Galdi, and H. Sohr. Very weak solutions and large uniqueness classes of stationary Navier-Stokes equations in bounded domains of $\mathbb{R}^2$. *J. Differential Equations*, 227(2):564–580, 2006.

[29] L. E. Fraenkel. On a theory of laminar flow in channels of a certain class. *Proc. Cambridge Philos. Soc.*, 73:361–390, 1973.

[30] L. E. Fraenkel and P. M. Eagles. On a theory of laminar flow in channels of a certain class. II. *Math. Proc. Cambridge Philos. Soc.*, 77:199–224, 1975.

[31] G. P. Galdi. *An introduction to the mathematical theory of the Navier-Stokes equations.* Springer Monographs in Mathematics. Springer, New York, second edition, 2011. Steady-state problems.

[32] G. P. Galdi, P. Maremonti, and Y. Zhou. On the Navier-Stokes problem in exterior domains with non decaying initial data. *J. Math. Fluid Mech.*, 14(4):633–652, 2012.

[33] G. P. Galdi and P. J. Rabier. Sharp existence results for the stationary Navier-Stokes problem in three-dimensional exterior domains. *Arch. Ration. Mech. Anal.*, 154(4):343–368, 2000.

[34] G. P. Galdi and A. L. Silvestre. Strong solutions to the Navier-Stokes equations around a rotating obstacle. *Arch. Ration. Mech. Anal.*, 176(3):331–350, 2005.

[35] Y. Giga and T. Miyakawa. Navier-Stokes flow in $\mathbb{R}^3$ with measures as initial vorticity and Morrey spaces. *Comm. Partial Differential Equations*, 14(5):577–618, 1989.

[36] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products.* Elsevier/Academic Press, Amsterdam, seventh edition, 2007. Translated from the Russian.

[37] J. Guzmán and L. R. Scott. The Scott-Vogelius finite elements revisited. *Math. Comput.*, 88(316):515–529, 2019.

[38] R. Hammer, M. Hocks, U. Kulisch, and D. Ratz. *C++ toolbox for verified computing.* Springer, Berlin, 1995.

[39] J. G. Heywood. On uniqueness questions in the theory of viscous flow. *Acta Math.*, 136(1-2):61–102, 1976.

[40] M. Hillairet and P. Wittwer. Existence of stationary solutions of the Navier-Stokes equations in two dimensions in the presence of a wall. *J. Evol. Equ.*, 9(4):675–706, 2009.

[41] M. Hillairet and P. Wittwer. On the existence of solutions to the planar exterior Navier Stokes system. *J. Differential Equations*, 255(10):2996–3019, 2013.

[42] V. Hoang, M. Plum, and C. Wieners. A computer-assisted proof for photonic band gaps. *Z. Angew. Math. Phys.*, 60(6):1035–1052, 2009.

[43] W. Hofschuster and W. Krämer. C-xsc 2.0 – a c++ library for extended scientific computing. In R. Alt, A. Frommer, R. B. Kearfott, and W. Luther, editors, *Numerical Software with Result Verification*, pages 15–35, Berlin, Heidelberg, 2004. Springer.

[44] E. Hopf. Über die Anfangswertaufgabe für die hydrodynamischen Grundgleichungen. *Math. Nachr.*, 4:213–231, 1951.

[45] R. Hungerbühler and J. Garloff. Bounds for the range of a bivariate polynomial over a triangle. *Reliab. Comput.*, 4(1):3–13, 1998.

[46] R. Hungerbühler and J. Garloff. Computation of the Bernstein coefficients on subdivided triangles. *Reliab. Comput.*, 6(2):115–121, 2000.

[47] V. John. *Finite Element Methods for Incompressible Flow Problems.* Springer Series in Computational Mathematics. Springer, Cham, 2016.

[48] T. Kato. Strong L^p-solutions of the Navier-Stokes equation in $\mathbb{R}^m$, with applications to weak solutions. *Math. Z.*, 187(4):471–480, 1984.

[49] T. Kato. Strong solutions of the Navier-Stokes equation in Morrey spaces. *Bol. Soc. Brasil. Mat. (N.S.)*, 22(2):127–155, 1992.

[50] T. Kato. *Perturbation theory for linear operators.* Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[51] T. Kato and G. Ponce. Commutator estimates and the Euler and Navier-Stokes equations. *Comm. Pure Appl. Math.*, 41(7):891–907, 1988.

[52] H. Kim and H. Kozono. On the stationary Navier-Stokes equations in exterior domains. *J. Math. Anal. Appl.*, 395(2):486–495, 2012.

[53] R. Klatte, U. W. Kulisch, A. Wiethoff, C. Lawo, and M. Rauch. *C-XSC. A C++ class library for extended scientific computing. Transl. by G. F. Corliss, C. Lawo, R. Klatte, A. Wiethoff, C. Wolff.* Berlin: Springer-Verlag, 1993.

[54] H. Koch and D. Tataru. Well-posedness for the Navier-Stokes equations. *Adv. Math.*, 157(1):22–35, 2001.

[55] M. V. Korobkov, K. Pileckas, and R. Russo. Solution of Leray's problem for stationary Navier-Stokes equations in plane and axially symmetric spatial domains. *Ann. of Math. (2)*, 181(2):769–807, 2015.

[56] H. Kozono and T. Yanagisawa. Leray's problem on the stationary Navier-Stokes equations with inhomogeneous boundary data. *Math. Z.*, 262(1):27–39, 2009.

[57] J.-R. Lahmann and M. Plum. A computer-assisted instability proof for the Orr-Sommerfeld equation with Blasius profile. *ZAMM Z. Angew. Math. Mech.*, 84(3):188–204, 2004.

[58] N. J. Lehmann. Optimale Eigenwerteinschliessungen. *Numer. Math.*, 5:246–272, 1963.

[59] J. Leray. Sur le mouvement d'un liquide visqueux emplissant l'espace. *Acta Math.*, 63(1):193–248, 1934.

[60] X. Liu. Verified Finite Element Method (VFEM) libary. `https://ganjin.online/xfliu/vfem`.

[61] H. J. Maehly. Ein neues Variationsverfahren zur genäherten Berechnung der Eigenwerte hermitescher Operatoren. *Helvetica Phys. Acta*, 25:547–568, 1952.

[62] D. Maurer and C. Wieners. Parallel multigrid methods and coarse grid LDL^T solver for Maxwell's eigenvalue problem. In H.-G. Hegering, W. E. Nagel, and G. Wittum, editors, *Competence in High Performance Computing 2010 : Proceedings of an International Conference on Competence in High Performance Computing, June 2010,*

Schloss Schwetzingen, Germany, SpringerLinkSpringer eBook Collection. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[63] P. J. McKenna, F. Pacella, M. Plum, and D. Roth. A computer-assisted uniqueness proof for a semilinear elliptic boundary value problem. In *Inequalities and applications 2010*, volume 161 of *Internat. Ser. Numer. Math.*, pages 31–52. Birkhäuser/Springer, Basel, 2012.

[64] S. Monniaux. Behaviour of the Stokes operators under domain perturbation. *Sci. China Math.*, 62(6):1167–1174, 2019.

[65] R. E. Moore. *Interval analysis.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1966.

[66] P. B. Mucha. On Navier-Stokes equations with slip boundary conditions in an infinite pipe. *Acta Appl. Math.*, 76(1):1–15, 2003.

[67] K. Nagatou. A computer-assisted proof on the stability of the Kolmogorov flows of incompressible viscous fluid. *J. Comput. Appl. Math.*, 169(1):33–44, 2004.

[68] K. Nagatou, K. Hashimoto, and M. T. Nakao. Numerical verification of stationary solutions for Navier-Stokes problems. *J. Comput. Appl. Math.*, 199(2):445–451, 2007.

[69] M. T. Nakao. A numerical approach to the proof of existence of solutions for elliptic problems. *Japan J. Appl. Math.*, 5(2):313–332, 1988.

[70] M. T. Nakao. A numerical verification method for the existence of weak solutions for nonlinear boundary value problems. *J. Math. Anal. Appl.*, 164(2):489–507, 1992.

[71] M. T. Nakao. Numerical verification methods for solutions of ordinary and partial differential equations. *Numer. Funct. Anal. Optim.*, 22:321–356, 2001. International Workshops on Numerical Methods and Verification of Solutions, and on Numerical Function Analysis (Ehime/Shimane, 1999).

[72] M. T. Nakao, K. Hashimoto, and K. Kobayashi. Verified numerical computation of solutions for the stationary Navier-Stokes equation in nonconvex polygonal domains. *Hokkaido Math. J.*, 36(4):777–799, 2007.

[73] M. T. Nakao, K. Hashimoto, and Y. Watanabe. A numerical method to verify the invertibility of linear elliptic operators with applications to nonlinear problems. *Computing*, 75(1):1–14, 2005.

[74] M. T. Nakao, M. Plum, and Y. Watanabe. *Numerical verification methods and computer-assisted proofs for partial differential equations*, volume 53 of *Springer Series in Computational Mathematics.* Springer, Singapore, 2019.

[75] M. T. Nakao and Y. Watanabe. An efficient approach to the numerical verification for solutions of elliptic differential equations. *Numer. Algorithms*, 37(1-4):311–323, 2004.

[76] M. T. Nakao, N. Yamamoto, and K. Nagatou. Numerical verifications for eigenvalues of second-order elliptic operators. *Japan J. Indust. Appl. Math.*, 16(3):307–320, 1999.

[77] M. T. Nakao, N. Yamamoto, and Y. Watanabe. A posteriori and constructive a priori error bounds for finite element solutions of the Stokes equations. *J. Comput. Appl. Math.*, 91(1):137–158, 1998.

[78] S. A. Nazarov and B. A. Plamenevsky. *Elliptic problems in domains with piecewise smooth boundaries*, volume 13 of *De Gruyter Expositions in Mathematics*. Walter de Gruyter & Co., Berlin, 1994.

[79] H. Oertel jr. *Prandtl - Führer durch die Strömungslehre : Grundlagen und Phänomene*. SpringerLink. Springer Vieweg, Wiesbaden, 13., überarb. Aufl. 2012 edition, 2012.

[80] OpenBLAS Library. `https://www.openblas.net/`.

[81] C. Oseen. Über die Stokessche Formel, und über eine verwandte Aufgabe in der Hydrodynamik. *Ark. Mat. Astron. Fys.*, 6:1–20, 1910.

[82] F. Pacella, M. Plum, and D. Rütters. A computer-assisted existence proof for Emden's equation on an unbounded L-shaped domain. *Commun. Contemp. Math.*, 19(2):1750005, 21, 2017.

[83] F. Planchon. Global strong solutions in Sobolev or Lebesgue spaces to the incompressible Navier-Stokes equations in $\mathbb{R}^3$. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 13(3):319–336, 1996.

[84] R. Plato. Numerische Mathematik kompakt : Grundlagenwissen für Studium und Praxis, 2010.

[85] M. Plum. Existence and multiplicity proofs for semilinear elliptic boundary value problems by computer assistance. *Jahresber. Deutsch. Math.-Verein.*, 110(1):19–54, 2008.

[86] M. Plum and C. Wieners. New solutions of the Gelfand problem. *J. Math. Anal. Appl.*, 269(2):588–606, 2002.

[87] N. Revol and F. Rouillier. Motivations for an arbitrary precision interval arithmetic and the MPFI library. *Reliab. Comput.*, 11(4):275–290, 2005.

[88] T. Richter. Lecture Notes: Numerische Methoden der Strömungsmechanik, February 2015.

[89] D. Rütters. *Computer-assisted Multiplicity Proofs for Emden's Equation on Domains with Hole*. PhD thesis, Karlsruhe Institute of Technology (KIT), 2014. Karlsruhe, KIT, Diss., 2014.

[90] S. Rump. INTLAB - INTerval LABoratory. In T. Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, Dordrecht, 1999. `http://www.ti3.tuhh.de/rump/`.

[91] A. Russo. A note on the exterior two-dimensional steady-state Navier-Stokes problem. *J. Math. Fluid Mech.*, 11(3):407–414, 2009.

[92] B. Schweizer. *Partielle Differentialgleichungen*. Springer-Verlag, Berlin, 2013. Eine anwendungsorientierte Einführung. [An application-oriented introduction].

[93] V. Serov. *Fourier Series, Fourier Transform and Their Applications to Mathematical Physics*. Applied Mathematical Sciences. Springer, Cham, 2017.

[94] H. Sohr. *The Navier-Stokes equations.* Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Birkhäuser Verlag, Basel, 2001. An elementary functional analytic approach.

[95] H. Sohr and W. von Wahl. On the regularity of the pressure of weak solutions of Navier-Stokes equations. *Arch. Math. (Basel)*, 46(5):428–439, 1986.

[96] Steinbuch Centre for Computing (SCC). HoreKa. `https://www.scc.kit.edu/dienste/horeka.php`.

[97] G. Stokes. On the effect of the internal friction of fluids on the motion of penulums. *Trans. Cambridge Phil. Soc.*, 9:8–106, 1851.

[98] C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Internat. J. Comput. & Fluids*, 1(1):73–100, 1973.

[99] M. E. Taylor. Analysis on Morrey spaces and applications to Navier-Stokes and other evolution equations. *Comm. Partial Differential Equations*, 17(9-10):1407–1456, 1992.

[100] R. Temam. *Navier-Stokes equations*, volume 2 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York, revised edition, 1979. Theory and numerical analysis, With an appendix by F. Thomasset.

[101] L. Tobiska and R. Verfürth. Analysis of a streamline diffusion finite element method for the Stokes and Navier-Stokes equations. *SIAM J. Numer. Anal.*, 33(1):107–127, 1996.

[102] Y. Watanabe. A computer-assisted proof for the Kolmogorov flows of incompressible viscous fluid. *J. Comput. Appl. Math.*, 223(2):953–966, 2009.

[103] Y. Watanabe. An efficient numerical verification method for the Kolmogorov problem of incompressible viscous fluid. *J. Comput. Appl. Math.*, 302:157–170, 2016.

[104] Y. Watanabe, K. Nagatou, M. Plum, and M.T.Nakao. A computer-assisted stability proof for the Orr-Sommerfeld problem with Poiseuille flow. *A special issue of "Nonlinear Theory and Its Applications, IEICE" on "Recent Progress in Verified Numerical Computations"*, 2:123–127, 2011.

[105] Y. Watanabe, M. T. Nakao, and K. Nagatou. On the compactness of a nonlinear operator related to stream function-vorticity formulation for the Navier-Stokes equations. *JSIAM Lett.*, 9:77–80, 2017.

[106] Y. Watanabe, M. Plum, and M. T. Nakao. A computer-assisted instability proof for the Orr-Sommerfeld problem with Poiseuille flow. *ZAMM Z. Angew. Math. Mech.*, 89(1):5–18, 2009.

[107] Y. Watanabe, N. Yamamoto, and M. T. Nakao. A numerical verification method of solutions for the Navier-Stokes equations. In *Developments in reliable computing (Budapest, 1998)*, pages 347–357. Kluwer Acad. Publ., Dordrecht, 1999.

[108] H. F. Weinberger. *Variational methods for eigenvalue approximation.* Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1974. Based on a series of lectures presented at the NSF-CBMS Regional Conference on Approximation of

Eigenvalues of Differential Operators, Vanderbilt University, Nashville, Tenn., June 26–30, 1972, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 15.

[109] A. Weinstein and W. Stenger. *Methods of intermediate problems for eigenvalues.* Academic Press, New York-London, 1972. Theory and ramifications, Mathematics in Science and Engineering, Vol. 89.

[110] C. Wieners. Finite Element Software M++ (Meshes, Multigrid and More). `https://git.scc.kit.edu/mpp/mpp`, Releases: `https://git.scc.kit.edu/mpp/mpp/-/releases`.

[111] C. Wieners. Tutorial - Finite Element Software M++ (Meshes, Multigrid and More). `https://git.scc.kit.edu/mpp/tutorial`.

[112] C. Wieners. Numerical enclosures for solutions of the Navier-Stokes equation for small Reynolds numbers. In *Numerical methods and error bounds (Oldenburg, 1995)*, volume 89 of *Math. Res.*, pages 280–286. Akademie Verlag, Berlin, 1996.

[113] C. Wieners. A geometric data structure for parallel finite elements and the application to multigrid methods with block smoothing. *Comput. Vis. Sci.*, 13(4):161–175, 2010.

[114] C. Wieners et al. M++ group on gitLab. `https://git.scc.kit.edu/mpp`.

[115] P. Wittwer. On the structure of stationary solutions of the Navier-Stokes equations. *Comm. Math. Phys.*, 226(3):455–474, 2002.

[116] J. Wunderlich. Navier Stokes Project on gitLab. `https://git.scc.kit.edu/mpp/navierstokes`.

[117] J. Wunderlich and M. Plum. Computer-assisted existence proofs for one-dimensional Schrödinger-Poisson systems. *Acta Cybern.*, 24(3):373–391, 2020.

[118] S. Zhang. On the P1 Powell-Sabin divergence-free finite element for the Stokes equations. *J. Comput. Math.*, 26(3):456–470, 2008.

[119] S. Zimmermann and U. Mertins. Variational bounds to eigenvalues of self-adjoint eigenvalue problems with arbitrary spectrum. *Z. Anal. Anwendungen*, 14(2):327–345, 1995.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur unter Zuhilfenahme der ausgewiesenen Hilfsmittel angefertigt habe. Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach anderen gedruckten oder im Internet veröffentlichten Werken entnommen sind, habe ich durch genaue Quellenangaben kenntlich gemacht.

Karlsruhe, den 24.01.2022