Galerkin Least-Squares Stabilization Operators for the Navier-Stokes Equations

A Unified Approach

Mónika Polner

The research described in this dissertation was undertaken at the Numerical Analysis and Computational Mechanics Group, Department of Applied Mathematics, in the Faculty EWI, Universiteit Twente, P. O. Box 217, 7500 AE Enschede, The Netherlands.

The funding of this research was provided by the Maritime Research Institute Netherlands (MARIN). This funding is gratefully acknowledged.

This work was part of the research program of the J. M. Burgers School for Fluid Mechanics.

Front cover: Vorticity contours of a driven cavity flow on a 50×50 clustered mesh at a Reynolds number equal to 800. For further details we refer to Section 7.2.

Back cover: Streamlines for the solution of unsteady flow around a circular cylinder at Reynolds number equal to 200. For further details we refer to Section 7.3.

Printed by Wöhrmann Printing Service, Zutphen, The Netherlands

ISBN 90-365-2276-5

GALERKIN LEAST-SQUARES STABILIZATION OPERATORS FOR THE NAVIER-STOKES EQUATIONS

A UNIFIED APPROACH

DISSERTATION

to obtain
the doctor's degree at the University of Twente,
on the authority of the rector magnificus,
prof.dr. W.H.M. Zijm,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday 17 November 2005 at 15.00

by

Mónika Polner
born on 11 November 1972
in Nagykároly

This dissertation is approved by

the promoter
prof.dr.ir. J.J.W. van der Vegt

the assistant promoter
dr. R.M.J. van Damme

The most beautiful thing we can experience is the mysterious. It is the source of all true art and science.

Albert Einstein

This work is dedicated to Boudewijn.

Summary

In this dissertation we attempt to approach fluid mechanics problems from a unified point of view and to combine techniques developed originally for compressible or incompressible flows into a more general framework. In order to study the viability of a unified approach, it is necessary to choose a starting formulation, therefore the choice of variables in the governing equations is crucial. For example, conservative variables are not suitable for a unified formulation because they result in a singular limit for incompressible flows. When entropy or pressure primitive variables are used, then the incompressible limit of the Navier-Stokes equations is well-defined, hence they are a suitable choice for obtaining a unified formulation. These two sets of variables are investigated in detail.

Since each set of variables possesses unique properties, the accuracy, stability, robustness, and computational efficiency of a numerical method strongly depend on this choice.

The numerical discretization is a time-discontinuous Galerkin least-squares finite element method. An essential part in this algorithm and related methods is the stabilization operator. For compressible flows a stabilization operator in the finite element discretization is generally required to prevent numerical oscillations in regions with discontinuities or sharp gradients which are not accurately represented by the computational mesh. For incompressible flows, the concept of a stabilization operator is also crucial and eliminates the complications of designing elements which satisfy the inf-sup stability condition. Although very different concerns are present in solving compressible and incompressible flows which motivate the need of a stabilization operator in the variational formulation, this thesis shows that many ideas developed in one field can be used in the other field.

The most important ingredient to obtain a unified formulation is the design of a stabilization matrix which is valid for both type of flows. The choice of this matrix is crucial to ensure stability of the numerical discretization without compromising accuracy. Moreover, the stabilization matrix designed for incompressible flows might not be effective in the compressible flow regime and reversely, the compressible stabilization matrix might not be well-defined in the incompressible limit. This dissertation

presents a new technique to design stabilization matrices that can be used for both type of flows. The proposed class of stabilization matrices is obtained using dimensional analysis with respect to the flow variables. In the construction of the stabilization matrices we used the benefits of both entropy and pressure primitive variables. The obtained stabilization matrices are well-defined in the incompressible limit for both entropy and primitive variables and this is considered to be the main result of this thesis. The proposed class of dimensionally consistent stabilization matrices is further investigated to enhance stability of the Galerkin least-squares finite element discretization of the linearized incompressible Navier-Stokes equations and nonlinear stability in the compressible case. Therefore, we give necessary and sufficient condition on the positive definiteness of the designed stabilization matrix for entropy variables.

The time-discontinuous Galerkin least-squares finite element discretization results in a large system of nonlinear algebraic equations. For unsteady problems, a linear-in-time approximation of the space-time Galerkin least-squares variational equation is needed. In this thesis we propose a new method to solve the nonlinear algebraic system and compare the algorithm with the predictor multi-corrector method using the advection-diffusion equation as a model problem.

The newly designed stabilization matrix is demonstrated for the incompressible Navier-Stokes equations using some numerical examples. The main emphasis is on the influence of this stabilization matrix on the accuracy of the numerical discretization. The numerical examples show that when pressure primitive variables are used, the new class of stabilization matrices developed in this thesis perform well in stabilizing the numerical method without degrading accuracy .

Acknowledgments

This work could not have been accomplished without the support, encouragement and enthusiasm of many people surrounding me. It is not possible to thank everyone individually, I would, therefore, like to express my gratefulness to everybody who was there for me.

First, I would like to express my sincere gratitude to my promotor, Jaap van der Vegt, for giving me the opportunity to work in his group and for his valuable guidance, assistance, motivation and encouragement in my work.

I wish to acknowledge my supervisor Ruud van Damme, for his help and assistance in preparing numerical results in the last two years of my research. We had many interesting discussions and I never hesitated to come to him with any question.

The financial support of the Maritime Research Institute Netherlands (MARIN) is gratefully acknowledged. I particularly would like to thank Rene Huijsmans for his encouragement and enthusiastic discussions at Marin.

I am very grateful to Mike Botchev for his valuable suggestions in the linear algebra related parts of this thesis and Peter Gragert who introduced me in the world of computer science. I learned a lot from his knowledge and experience.

Many thanks to Helena who was my friend from my first day in The Netherlands, and also to my excellent English teacher Joanne, who taught me English just by being my friend.

Furthermore, I would like to thank my office mates: Agus, for our long discussions on Dutch culture during the three years we shared an office; Jaqueline, for her support and friendship; Davit and Pablo, for many good laughs and Bert, for the useful discussions on different numerical and physical problems. Many thanks go to Vita with whom I always shared both good and difficult times, happiness and complains. I would like to thank my present and former colleagues Timco, Lars, Chris, Kiran and Joris for their help and friendship and also Marielle for taking care of all the administrative work. Their friendship and support helped me a lot in completing this thesis.

Apart from work, a special place in my life is reserved for fencing. The possibility to train in The Netherlands had a big contribution to my research. Therefore, I would like to thank all my fencing friends and especially my trainer André for constantly challenging me.

A special word of thanks should be addressed to Wim, Egna, Laurien, Jorrit and Freia for their constant support, encouragement, happiness and for making me feel at home.

I wish to extend my deepest thanks to my parents, brother and sister, and their family, for their love, support, encouragement and understanding.

Finally, my sincere love and gratitude goes to Boudewijn, who brings out the best in me. I would like to thank you for being always there for me.

Mónika Polner
Enschede, October 2005

Contents

Chapter 1

Introduction

1.1 Motivation

Many applications in fluid dynamics require the solution of the Navier-Stokes equations in a time-dependent flow domain. Examples can be found in the analysis of fluid-structure interaction, moving spatial configurations and flows with (internal) free surfaces. In order to accurately represent the solution of these problems, the numerical method requires the use of moving and deforming meshes.

Several numerical techniques can deal with deforming meshes, e.g. finite volume methods, but in this thesis we focus on finite element methods since they provide an excellent framework for solving the Navier-Stokes equations and make it possible to efficiently deal with flow domains with a complicated geometry and to adapt the computational mesh to accurately capture boundary layers, vortical structures and other detailed flow phenomena.

Fluid flow problems that involve moving and deforming spatial configurations have been an area of great interest. In particular, arbitrary Lagrangian-Eulerian finite element techniques have been successfully used to deal with time-dependent fluid flow problems with changing spatial configurations, see [40] and the references therein. In [29], [36], [51], space-time techniques have been developed based on a *Galerkin least-squares* finite element method with fixed spatial domains and in [20], [37] for fluid and solid mechanics problems. The idea of the space-time Galerkin least-squares finite element method for the Navier-Stokes equations was extended in [40] to computations that involve changing configurations. The variational formulations of these methods employ the time-discontinuous Galerkin least-squares method, which is also the basis of our formulation. The time-discontinuous Galerkin least-squares method subdivides the space-time domain into so-called space-time slabs. The polynomial basis functions are discontinuous across space-time slab boundaries but continuous inside the space-time slab. This provides the flexibility to change the computational mesh from one

space-time slab to another in case of severe deformations and also provides a natural mechanism for incorporating adaptive remeshing in the formulation.

When applied to advection dominated problems, the Galerkin method, and likewise the time-discontinuous Galerkin method, lack stability. Unresolved internal and boundary layers locally result in severe oscillations in the solution, which pollute the entire solution. To overcome these difficulties, a *least-squares* operator is added to the basic Galerkin formulation. When properly defined, these operators guarantee stability without compromising accuracy. Stabilized methods will be discussed in great detail later on in this thesis.

The motivation of this research is the need to understand and predict the dynamic behavior of risers in waves and current, which is of great importance for the offshore industry. The risers are situated subsea, and consist of a structure containing pipes, valves, and connectors linked to the seabed and to a floating or fixed production platform. The pipes are used to extract oil and gas from the subsea, as illustrated in Figure 1.1. The riser is exposed to waves and current and its dynamic behavior and fatigue life is strongly influenced by the flow field surrounding its components, in particular periodic vortex shedding.

The Galerkin least-squares method, in combination with a suitable stabilization operator, is an excellent method to accurately compute the periodic shedding of vortices and, in combination with the space-time formulation, it is well suited for moving and deforming meshes. For example, in the case of a vibrating cylinder or when more cylinders move with respect to each other in a riser. The computation of this

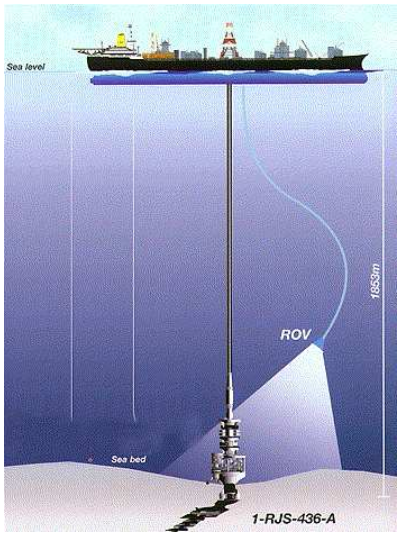

Figure 1.1: Riser connected to a floating production at water depth of 1853 m.

type of flows is, however, non trivial and in this thesis we focus on the mathematical formulation of Galerkin least-squares methods for the Navier-Stokes equations.

Stabilized methods have been around for some time. They were initially developed for incompressible flows in [8], and later extended to compressible flows [29], [30], [32], [33]. In combination with the developments in stabilized methods, several new space-time formulations were developed and tested for a large range of compressible and incompressible flow problems, [56], [15] and references therein. Although very different concerns are present in solving compressible and incompressible flows, there are several ideas developed in one field that can be used in the other field. Examples are the concept of an entropy variables formulation of the governing equations and the Galerkin least-squares method, that have been successfully applied to both types of flows, [22], [23]. The objective of this work is to approach fluid mechanics from a more unified point of view and to combine ideas for numerical methods developed originally for compressible or incompressible flows into a more general framework.

The compressible Navier-Stokes equations in conservative form represent the conservation of mass, momentum and energy. The system of conservation laws must be closed by adding two equations of state, or equivalently, by determining the fundamental equation of the system. Then, all relevant thermodynamic quantities can be obtained, which is sufficient to construct the flux vectors and coefficient matrices in the compressible Navier-Stokes equations. The conservative variables are, however, not suitable for a unified formulation since they result in a singular limit for incompressible flows. In [23], Hauke and Hughes demonstrated that, with the proper choice of variables, it is possible to obtain a formulation of the Navier-Stokes equations which is valid for both compressible and incompressible flows. This makes it possible to obtain a unified discretization valid for both type of flows.

In order to study the viability of a unified approach for compressible and incompressible flows, it is necessary to choose a starting formulation, therefore the choice of variables in the governing equation is crucial. Since each set of variables possesses unique properties, the accuracy, stability, robustness and computational efficiency of the numerical method depend on this choice. For a comparative study of different sets of variables we refer to the work of Hauke and Hughes [23]. It can be shown that when entropy or pressure primitive variables are used, then the incompressible limit of the Navier-Stokes equations is well-defined, therefore, they are a suitable choice for obtaining a unified formulation. This motivates the use of (physical) entropy variables as a starting point of our formulation. A detailed description of the entropy variables can be found in [9], [50], [30], and in several chapters of this thesis.

Another benefit of using entropy variables is that they symmetrize the Navier-Stokes equations. It is known that symmetric systems and notions of generalized entropy functions are closely linked and we discuss this link in later chapters of this thesis. The symmetric form of the compressible Navier-Stokes equations expresses the mathematical and physical stability provided by the second law of thermodynamics.

Entropy production is governed by the second law of thermodynamics and entropy variables provide a formulation which satisfies the entropy condition. Moreover, the discrete solution of the Galerkin formulation based on entropy variables automatically satisfies the Clausius-Duhem inequality, or second law of thermodynamics.

An incompressible fluid is characterized by constant density and therefore, the density is not a thermodynamic variable. The thermodynamic theory in this approximation has one less independent variable than for a simple compressible substance. Since pressure is no longer a thermodynamic but a mechanical variable, there is only one equation of state for incompressible flows. We are interested in a unified formulation, and special attention is therefore given to the study of the various thermodynamic limits. We address the important question of the existence of a general form of the fundamental equation, such that by taking the incompressible limit, the equation is well defined and describes an incompressible fluid.

An essential ingredient to obtain a unified formulation, valid for both compressible and incompressible flows, is provided by the stabilization operator. Compressible flow equations generally require a stabilization operator in the finite element discretization to prevent numerical oscillations in regions with discontinuities or sharp gradients which are not accurately represented on the computational mesh. The concept of a stabilization operator is also necessary for incompressible flows and eliminates the complications of designing elements which satisfy the inf-sup stability condition for a mixed formulation.

In this thesis we will therefore focus on the formulation of stabilization operators, and in particular the stabilization matrix. The choice of this matrix is crucial to ensure stability of the numerical discretization but it can also negatively influence the accuracy of the scheme. In the next section we give an overview of different aspects of the stabilization operator and the important stabilization parameters in it. If entropy or pressure primitive variables are used, the same formulation can be used to compute compressible and incompressible flows. The difficulty to overcome is, however, to design a stabilization matrix which is valid for both type of flows. The stabilization matrix designed for incompressible flows might not be effective in the compressible flow regime and reversely, the compressible stabilization matrix might not be well-defined in the incompressible limit. There have been earlier attempts to design stabilization matrices valid for both type of flows in [23] and in [42] using augmented conservation variables. In this thesis we propose a new class of stabilization matrices for both entropy and primitive variables and show that the incompressible limit is well-behaved and results in the stabilization matrix designed for incompressible flows. This unified formulation of stabilization matrices is considered to be the main contribution of this thesis.

1.2 Overview of stabilized methods

An excellent technique to compute fluid flow problems is provided by stabilized finite element methods. Although, stabilized methods have been mostly developed in the context of fluid mechanics, there have also been a number of successful applications of stabilized methods in structural mechanics. Stabilized finite elements are constructed by modifying the variational form of a particular problem, such that enhanced numerical stability is achieved without compromising consistency or accuracy. The origins of stabilized methods can be traced back to the early 80's when T. J. R. Hughes and coworkers realized the important origin of the lack of stability of the Galerkin method for advection-dominated diffusion problems. Stabilized methods were initially developed for advection-diffusion problems and for incompressible flows in [8], and later extended to compressible flows, [29], [30], [32] and [33]. In this section we discuss the key features of stabilized methods using two model problems, the advection-diffusion equation and the steady state Stokes equations.

1.2.1 Model problems

Model I. Consider the *advection-diffusion* equation

$$\mathcal{L}\phi = \phi_{,t} + a \cdot \nabla\phi - \nabla \cdot (\kappa \nabla\phi) - f = 0 \tag{1.2.1}$$

where $\phi = \phi(t, x)$ is the dependent variable, a scalar-valued function of the spatial coordinates $x \in \Omega \subset \mathbb{R}^d$ and time $t \in (0, T)$. Furthermore, the velocity vector is denoted by $a = a(x)$, $f = f(t, x)$ is the source function and the diffusivity tensor is $\kappa = \kappa(x)$. We assume for simplicity that κ is isotropic and positive-definite. Therefore, $\kappa = kI$, where I is the identity matrix, and the scalar $k = k(x)$ is positive. The domain Ω is assumed to be smooth and in addition, we consider a homogeneous Dirichlet boundary condition

$$\phi(t, x) = 0 \quad \text{on } \Gamma = \partial\Omega. \tag{1.2.2}$$

The *Galerkin variational formulation* corresponding to (1.2.1-1.2.2) is obtained by multiplying (1.2.1) by test functions and integrating the equation over the computational domain:

Find $\phi \in H_0^1(\Omega)$ such that

$$(\mathcal{L}\phi, w) = (\phi_{,t} + a \cdot \nabla\phi, w) + (\kappa\nabla\phi, \nabla w) - (f, w) = 0 \quad \forall w \in H_0^1(\Omega), \tag{1.2.3}$$

where $(\cdot, \cdot)$ indicates the inner product in $L^2(\Omega)$ and H_0^1 is the usual Sobolev space

$$H_0^1(\Omega) = \{v \mid v, Dv \in L^2(\Omega), v = 0 \text{ on } \Gamma\}.$$

Model II. The equations for Stokes flow with homogeneous Dirichlet boundary conditions are defined as:

$$\begin{aligned} -\nu\Delta u + \nabla p &= f \quad \text{in } \Omega \\ \nabla \cdot u &= 0 \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

where u is the velocity, p the pressure, ν the viscosity and f is the external force acting on the fluid. The mixed variational formulation of the Stokes equations can be given as follows:

Find $(u, p) \in H_0^1(\Omega) \times L_0^2(\Omega)$ such that

$$\begin{aligned} (\nabla u, \nabla w) - (\nabla \cdot w, p) &= (f, w) \qquad \forall w \in H_0^1(\Omega), \\ (\nabla \cdot u, q) &= 0 \qquad\qquad\quad \forall q \in L_0^2(\Omega), \end{aligned}$$

where $L_0^2(\Omega)$ is the subspace of $L^2(\Omega)$ consisting of all such functions in $L^2(\Omega)$ that have mean value zero. For this model problem we can take for simplicity, but without loss of generality, the viscosity to be equal to one.

1.2.2 Introduction to stabilized methods

Galerkin method. The standard Galerkin method, naturally associated with finite element methods, can be described as being an approximation of the variational formulation of a partial differential equation (PDE), or system of PDE's, on a space of functions that is spanned by piecewise polynomials. Therefore, the basic Galerkin method is constructed based on the variational formulation (1.2.3) by taking a subspace V_h of $H_0^1(\Omega)$ spanned by continuous piecewise polynomials, as

Find $\phi^h \in V_h$ such that

$$(\mathcal{L}\phi^h, w^h) = (\phi^h_{,t} + a \cdot \nabla\phi^h, w^h) + (\kappa\nabla\phi^h, \nabla w^h) - (f, w^h) = 0 \quad \forall w^h \in V_h.$$

The Galerkin method is a weighted residual formulation in which weighting and interpolation functions are from the same class of functions. The Galerkin method has been widely used in the mid 60's and early 70's and has been thought to be "THE" method to approximate PDE's. This idea was confirmed by the success of the method when applied to elliptic problems. The problems occur when the method is applied to more complicated problems, as stated by F. Brezzi et al. in [7]: *"However in the midst of this success the experts were aware that there were problems in applying this recipe to all problems under the sun."*

A typical class of problems where standard Galerkin methods fail are advection dominated problems, which are fundamental model problems in computational fluid dynamics since they expose the weakness of the classical numerical approaches, such

as central and upwind finite difference methods, as well as Galerkin finite element methods. It is well known that the Galerkin finite element method gives rise to central-difference type approximations of differential operators which are well suited only for elliptic problems. When the flow is dominated by advection, that is for high Peclet (or in the context of the Navier-Stokes equations, Reynolds) numbers, the Galerkin discretization gives rise to node-to-node oscillations of the solution or "wiggles." The only way to eliminate the oscillations seemed to be the refinement of the computational mesh in such a way that convection no longer dominates on the element level.

Upwind method. Wiggle-free solutions can be obtained by introducing *upwind* information to the convective term. Upwind convective terms can be constructed by adding *artificial diffusion* to the Galerkin method. For the steady one dimensional advection-diffusion model, the method can be formulated as:

Find $\phi^h \in V_h$ such that

$$(a\phi', w) + ((\kappa + \tilde{k})\phi', w') = (f, w) \quad \forall w^h \in V_h,$$

with $\phi' = \partial\phi/\partial x$ and the artificial diffusivity defined as

$$\tilde{k} = \frac{|a|h_e}{2}\bar{\xi}(\alpha_e), \tag{1.2.4}$$

where h_e is the element size and

$$\bar{\xi}(\alpha_e) = \coth(\alpha_e) - \frac{1}{\alpha_e},$$
$$\alpha_e = \frac{|a|h_e}{2k} \quad \text{(element Peclet number)}.$$

There are several drawbacks to this method, one is that upwind methods are generally less accurate. The loss of accuracy is manifested by overly diffuse solutions. Secondly, adding artificial diffusion is not related to the physics of the problem. In the framework of finite element methods, upwinded convective terms can be developed for example by modifying the weighting functions to achieve the upwind effect. This method has been successfully applied to the one dimensional advection-diffusion problem and was later extended to two dimensional cases. Unfortunately, when generalizing to more complicated problems, the method fails in the sense that it produces overly diffuse results or non accurate solutions. These effects can be observed in the presence of source terms, in time-dependent problems or when generalized to higher dimensions.

Summarizing, *upwind finite element methods* may be constructed by adding artificial diffusion to the Galerkin formulation, which results in exact nodal solutions for the one dimensional advection-diffusion problem, as discussed in [24], [25]. The failure of

this method for higher dimensions is due to the presence of *crosswind effects,* that is unnecessary diffusion in directions normal to the flow. Combining the success and failure of the upwind method, it is apparent that the upwinding effect is needed only in the direction of the flow.

Streamline upwind method. The *streamline upwind* method (SU) was introduced in [25] and designed to eliminate the crosswind diffusion problem. In this method, the artificial diffusivity, $\tilde{k}$, used in the one dimensional case (1.2.4), is replaced by the artificial diffusivity tensor $\tilde{k}_{ij}$, defined as

$$\tilde{k}_{ij} = \tilde{k}\hat{a}_i\hat{a}_j \tag{1.2.5}$$

where $\hat{a}_i = a_i/\|a\|$, $\|a\| = \sum_i a_i a_i$ and $\tilde{k}$ a scalar artificial diffusivity. Note that (1.2.5) represents a diffusivity acting only in the direction of the flow. The method can be formulated as:

Find $\phi^h \in V_h$, such that $\forall w^h \in V_h$, the following is valid

$$(\mathcal{L}\phi^h, w^h) + S^{SU}(\phi^h, w^h) = 0, \tag{1.2.6}$$

where

$$S^{SU}(\phi^h, w^h) = \sum_e \Big((a \cdot \nabla\phi^h), \tau_e(a \cdot \nabla w^h)\Big)_e \tag{1.2.7}$$

is the stabilization operator with τ_e a free parameter which determines the amount of upwind weighting. We will discuss in this thesis the development of this parameter for several problems in fluid mechanics and we will emphasize its importance in the stability of numerical methods. While integrals in the Galerkin method are defined over the entire computational domain, integrals in the stabilization term are restricted to elements and we indicate them with a subscript e. Note that stabilization terms are added only on the element interiors and not on the element boundaries. The SU method produces smooth solutions for high Reynolds number flows and the streamline upwinding solves also the crosswind problem. Modifying the test function for the convective term results, however, in a non-residual formulation. Consequently, the exact solution of the differential equation is no longer a solution of the variational problem as in the case of a Galerkin formulation. Non-residual formulations are known to produce inaccurate or wrong solutions when source terms are significant. Already in [8] some inconsistencies were realized since the source term and the time dependent part of the system was centrally weighted, resulting in overly diffuse solutions. Upwind weighting of all terms in the equation is therefore needed.

Streamline upwind Petrov-Galerkin method. A possibility to solve the inconsistencies of the SU method was proposed by Hughes and Brooks [25], [26], for a scalar advection-diffusion equation and consists in applying the streamline-upwind

test function to all terms in the equation. This method is called *streamline upwind Petrov-Galerkin* method (SUPG). Summarizing, the basic idea of the streamline upwind method is to add diffusion which acts only in the direction of the flow. This was then extended to a Petrov-Galerkin formulation, that means modifying the standard Galerkin weighting functions w^h (for all terms in the equation) by adding a streamline upwind perturbation which acts only in the flow direction:

$$\tilde{w}^h = w^h + \tau a \cdot \nabla w^h. \tag{1.2.8}$$

The SUPG method can be formulated as:

Find $\phi^h \in V_h$, such that $\forall w^h \in V_h$, the following is valid

$$(\mathcal{L}\phi^h, w^h) + S^{SUPG}(\phi^h, w^h) = 0, \tag{1.2.9}$$

where

$$\begin{aligned} S^{SUPG}(\phi^h, w^h) &= \sum_e (\mathcal{L}\phi^h, \tau_e a \cdot \nabla w^h)_e \\ &= \sum_e \Big(\underbrace{\phi^h_{,t} + a \cdot \nabla \phi^h - \nabla \cdot (\kappa \nabla \phi^h) - f}_{\text{advection-diffusion residual}}, \tau_e (a \cdot \nabla w^h) \Big)_e . \end{aligned} \tag{1.2.10}$$

At this point a separation has to be made between methods that apply an artificial diffusion and the SUPG method. It is important to note that SUPG is no longer associated with artificial diffusion and results in a consistent method.

In [8] SUPG has been applied to the linear scalar advection-diffusion equation and the incompressible Navier-Stokes equations. We will describe the development of SUPG for these two model problems later on in this section.

Galerkin least-squares method. More recently, a new class of stabilization methods was developed by observing that stabilization terms may be obtained by minimizing the square of the equation residual. This method, called *Galerkin least-squares* (GLS) method, is introduced in [29]. The basic idea is the following: start with the Galerkin finite element method and add least-squares terms of the residual. These terms enhance the stability of the Galerkin method without degrading accuracy. The stabilization term has the form

$$S^{GLS}(\phi^h, w^h) = \sum_e \Big(\mathcal{L}\phi^h, \tau_e \mathcal{L} w^h \Big)_e . \tag{1.2.11}$$

Both stabilization methods, viz. SUPG and GLS, are obtained by adding stabilization terms to the Galerkin formulation. The difference is in the structure of the stabilization terms and the GLS stabilization is a more general stabilization approach. This approach has been successfully applied to Stokes flows [27], compressible flows [29]

and [50], and incompressible flows [13], [52], and will be one of the main topics of this thesis.

Despite of its success, at this point we also have to address several shortcomings of the GLS and SUPG methods concerning the treatment of sharp boundary layers, source terms, time dependent flow problems and the generalization to multi-dimensions. For the remainder of this introduction we will discuss how these difficulties have been approached to successfully apply the GLS and SUPG methods to a large variety of fluid flow problems.

As discussed in [34], SUPG (and this applies also for GLS) is an excellent method for problems with smooth solutions, but local oscillations in the solution occur when discontinuities are present. To improve on this situation, in [34] a *discontinuity-capturing* term is added to the SUPG/GLS formulation. Such schemes are designed to introduce a dissipative effect in the neighborhood of discontinuities, without degrading the accuracy of the solution elsewhere in the flow field, see Figure 1.2. The discontinuity-capturing term has a form similar to the streamline term, but acts in the direction of the solution gradient rather than in the direction of the streamline. Since the discontinuity-capturing term is a function of the discrete solution gradient, the numerical method is nonlinear (even when the original equation is linear). Note that GLS or SUPG by itself is a linear method. In [34], the discontinuity-capturing term was defined for the scalar advection-diffusion equation and [33] deals with its generalization to systems.

The second topic of this introduction section on GLS/SUPG is its application to the *incompressible Navier-Stokes equations.* For incompressible flows, oscillations may arise not only from the convective nature of the flow, but also from the choice of finite element interpolation functions for the velocity and pressure. These numerical instabilities appear as oscillations in the pressure field. Treatment of the incompressibility constraint is one of the most difficult aspects of numerical algorithms for the incompressible Navier-Stokes equations. The classical Galerkin method applied to the incompressible Navier-Stokes equations gives rise to a so-called *mixed method*. The success of this method strongly depends upon the particular pair of velocity and pressure interpolations employed. For many combinations that would seem to be a

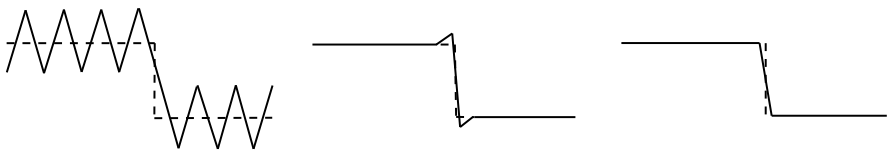

Figure 1.2: Oscillations present in the Galerkin method (left), controlled by stabilized methods (middle) and the overshoots/undershoots removed by the discontinuity capturing operator (right).

natural choice, severe oscillations in the pressure occur. The incompressibility constraint and its relation to the pressure field have been subject to extensive research. The mathematical framework for understanding the behavior of mixed methods for the Stokes problem was provided by Ladyzhenskaya, Babuška [1] and Brezzi [5]. Their result is one of the most important results in the theory of finite element methods and it was named Ladyzhenskaya-Babuška-Brezzi (LBB) or *inf-sup* stability condition. Due to the constraint nature of incompressible flows, the chosen combination of interpolation functions for velocity and pressure must satisfy the *inf-sup* stability condition. The GLS/SUPG method for the incompressible Navier-Stokes equations stabilizes the formulation to both the convective effects and the inf-sup condition. This makes it possible to use equal order interpolation functions for the velocity and pressure [28].

Due to their importance for the methods discussed in this thesis, we briefly summarize stabilized methods for the Stokes equations. The Stokes equations physically model slow motion of an incompressible fluid. In computational fluid dynamics, the Stokes equations provide an important model problem for analyzing finite element algorithms, such as stabilized methods. The reason that it is chosen as a model problem in many studies, and also by us, is that some of the difficulties encountered when solving the full incompressible Navier-Stokes equations are also present in the more simple Stokes equations.

The stabilized variational formulation for the Stokes equations was first formulated in [28]. In [27] the stabilized method for the Stokes equations was reformulated and almost at the same time in [11] an *absolutely stabilized* finite element formulation was given by Douglas and Wang, which can be viewed as a modification of the formulation given in [27]. The two methods can be formulated simultaneously as follows:

Find $(u^h, p^h) \in V_h \times P_h$ such that for all $(w^h, q^h) \in V_h \times P_h$

$$(\nabla u^h, \nabla w^h) - (\nabla \cdot w^h, p^h) - (\nabla \cdot u^h, q^h) - \alpha \sum_e h_e^2(-\Delta u^h + \nabla p^h, \boxed{\pm \Delta w^h} + \nabla q^h)$$
$$= (f, w^h) - \alpha \sum_e h_e^2(f, -\nu \Delta w^h + \nabla q^h), \qquad (1.2.12)$$

where α is a positive number and V_h, P_h are the finite element spaces for the velocity and pressure, respectively. Note that the two methods differ only in the sign of the term Δw^h in the stabilization term. The original methods, as proposed in [28], [27] and [11], have an additional jump term in the pressure, which is not important to our discussion. The "plus" method reduces to the method described in [27] and the "minus" formulation to the one given in [11]. The improvement from the original formulation given in [28] to the "plus" formulation in (1.2.12) is that the first one did not contain the Δw^h term and therefore, the latter one is symmetric and suitable for any combination of finite element spaces, either continuous or discontinuous for the pressure. The existence, uniqueness and convergence of the solution of the formulation

(1.2.12) still depends, however, on the stability constant α, which depends on the element shape. The main difference between the "plus" and the modified "minus" formulation is that the "minus" formulation is stable and convergent for all positive values of α. Furthermore, the results in [14] suggest that the method of Douglas and Wang in [11] is more robust, that means the accuracy of the finite element solution is less sensitive to the choice of α. On the other hand, the method of Douglas and Wang does not result in a symmetric system of equations.

Encouraged by the benefit that the method of Douglas and Wang has improved stability characteristics for Stokes flow compared with the GLS method when higher order interpolations are used, Franca et. al. [14], examined the consequence of its application to the incompressible Navier-Stokes equations. As a first step, the effect of this approach was analyzed on the advection-diffusion model and in [13] it was applied to the incompressible Navier-Stokes equations. In [14], Franca et. al. re-addressed the question of a careful design of the stability parameter. The motivation was to obtain good performance of the GLS method in the entire spectrum varying from advective to diffusive dominated flows. The design of the stabilization parameter is slightly improved by taking into account constants arising from inverse estimates. The usual element Peclet number was also modified to include the effect of the specific finite element polynomials employed.

It is important to mention that the GLS and SUPG stabilization methods have been successfully applied to not only incompressible, but also compressible flows. The SUPG method for hyperbolic systems and the compressible Euler equations was first introduced in [53] and later with additional examples in [35]. Other examples are the SUPG and GLS stabilization methods for compressible flows in entropy variables formulation in [31] and [50], respectively.

A central question which we will address in this thesis is the development of stabilization operators suitable for compressible flows which ensure that in the incompressible limit no pressure oscillations occur when equal order polynomials are used for the velocity and pressure.

Our work in compressible flows has emphasized the use of entropy variables. The method based on entropy variables possesses unique properties and provides a unified framework for both compressible and incompressible flows. There is a large variety of numerical methods developed for both type of flows and the main objective of this research is to collect ideas that have been successfully used in one field and can be applied in the other field. Examples of such ideas are the concept of stabilized methods and the use of entropy variables in the governing equations. The main topic of this thesis is to further investigate the choice of the stabilization parameters and give a suitable form that can be used for both compressible and incompressible flow computations.

1.2.3 Identification of the stabilization parameter τ.

An important component in the stabilization operator is the stabilization parameter (or matrix) τ, see for instance (1.2.10) and (1.2.11). The proper definition of the stabilization parameter τ is important for the stability and accuracy of the finite element discretization. An optimal parameter for the one dimensional advection-diffusion equation is available, which brings the next question, namely, how to generalize it to more dimensions, systems and more complicated flow problems. In this section we give a brief overview of stabilization parameters and emphasize what are the problems we want to overcome.

1 D case. For the one dimensional linear, steady advection-diffusion equation without source term, it was shown in [26] that when the *artificial diffusion parameter* k^e is chosen as

$$k^e = \frac{|a|h_e}{2}\bar{\xi}(\alpha_e), \tag{1.2.13}$$

where

$$\bar{\xi}(\alpha_e) = \coth(\alpha_e) - \frac{1}{\alpha_e}, \quad \alpha_e = \frac{|a|h_e}{2k} \quad \text{(element Peclet number)},$$

with h_e the element length, and when linear basis functions are employed, then the SUPG method gives rise to nodally exact solutions.

Multidimensional systems. A big step towards the generalization of the streamline operator in the SUPG method to *multidimensional* advective-diffusive systems was made in [32]. There have been earlier attempts to extend the definition of the stabilization term to one dimensional systems by using the same stabilization parameter for all components in the system. The failure of this approach was pointed out in [32]. Since in the original formulation only one τ parameter is used, there is no possibility to obtain simultaneously an optimal behavior with respect to all components in the system. If τ is too small for a particular component, spurious oscillations will result for that component. If τ is too large, it results in overly diffuse solutions. Consequently, a distinct τ is needed for each component. Successful generalizations to the multidimensional case must satisfy three important design requirements:

(1) It reduces consistently to the optimal one-dimensional system case.

(2) It is equivalent to SUPG/GLS for a scalar, multidimensional advection-diffusion equation.

(3) It reduces to SUPG/GLS on each uncoupled component of a multidimensional, diagonalizable advective-diffusive system.

In [32], a new definition of τ is presented, which represents a correct generalization of the streamline operator to coupled multidimensional systems. It is important to point out that a generalization of the concept of the absolute value of a matrix is required. The presentation of the SUPG method in [32] was restricted to steady cases. The generalization of τ to the unsteady space-time case in the framework of the GLS method is presented by Shakib in [50]. In this definition, τ is first formulated for pure advection, and then adjusted for the presence of diffusion.

A unified approach. In this thesis we address the question of designing a unified form of stabilization matrices, valid for a lage spectrum of flows. As we discussed earlier, the entropy variables provide a good starting point to obtain such a unified formulation. The main objective of this research is to design stabilization operators for entropy variables such that this operator is well defined in the incompressible limit. Although stabilized methods based on entropy variables possess unique properties, it is appealing to extend the formulation to other sets of variables, such as the pressure primitive variables. The benefits of obtaining alternative formulations include easier implementation and also easier analysis of possible stabilization operators. Since the pressure primitive variables are also suitable to study the incompressible limit, we use this set to obtain a dimensionally consistent stabilization operator for primitive variables and then extend it to entropy variables. For primitive variables, the stabilization matrix can be chosen to be of diagonal form, which has been successfully applied in previous research on incompressible flows, see [13]. When transforming this matrix to entropy variables, it results, however in a non-symmetric stabilization matrix, which destroys the symmetry of the weak formulation in entropy variables. It is possible to obtain a diagonal stabilization matrix for the entropy variables which is positive definite. This simple choice is, however, not good, as is shown by the numerical examples in [23]. In the same paper stabilization matrices for both entropy and primitive variables are proposed that can be employed for both compressible and incompressible flow computations, but there is still significant room for improvement and a need to further understand the properties of this type of stabilization operators, which will be addressed in this thesis.

1.3 Outline of this thesis

The contents of this thesis can be summarized as follows. The underlying physics of viscous flow is described in Chapter 2. This chapter recalls the basic thermodynamic relations that are used later in this thesis. These relations are constructed such that they satisfy Maxwell's fundamental thermodynamic relationships. In Chapter 2 a general equation of state is derived, using three measurable quantities. This equation of state can be used to describe both compressible and incompressible flows.

A comparative study of different sets of variables to give a unified formulation of the governing Navier-Stokes equations, valid for both compressible and incompressible flows, is given in Chapter 3. This chapter concludes that the incompressible limit of the Navier-Stokes equations is well defined only for entropy (or symmetrizing variables) and primitive variables (p, u, T).

In Chapter 4 we focus on the design and analysis of a dimensionally consistent class of stabilization operators suitable for space-time Galerkin least-squares finite element discretizations of the incompressible Navier-Stokes equations. In the analysis of the stability of the linearized incompressible Navier-Stokes equations, we give necessary and sufficient conditions for the positive definiteness of the proposed stabilization matrix for entropy variables.

If entropy or primitive variables are employed, the same formulation can be used to compute both compressible and incompressible flows. This gives the possibility to design stabilization operators that are well defined in the incompressible limit, which is very appealing. In Chapter 5, we define a class of stabilization operators for compressible flows without shocks and show that in the limit it is identical to the stabilization operator suitable for incompressible flows. The Galerkin least-squares method using the symmetrized form of the equations satisfies the Clausius-Duhem inequality, which is a non-linear stability condition. A crucial ingredient in the non-linear stability proof is the positive definiteness of the stabilization matrix. Therefore, the second topic of this chapter is to give necessary and sufficient conditions on the positive definiteness of the designed stabilization matrix.

The finite element discretization of the incompressible Navier-Stokes equations is presented in Chapter 6. The basis of our formulation is a time-discontinuous Galerkin least-squares method. The finite element discretization of the space-time variational equation leads to a system of nonlinear algebraic equations at each space-time slab. To reduce the nonlinear system to a sequence of linear systems, a predictor multi-corrector algorithm is used. The discretization presented in this chapter consists of a linear-in-time approximation.

In Chapter 7 we describe several test cases and applications to verify and demonstrate the Galerkin least-squares finite element method for the incompressible Navier-Stokes equations. Finally, we demonstrate the stabilization technique for the simulation of unsteady viscous flow about a circular cylinder. The main emphasis will lie on the accuracy of the simulation of the unsteady vortical structures in the wake, which is essential to obtain accurate predictions of unsteady lift and drag forces.

In Chapter 8, we draw conclusions and make recommendations for future research.

Chapter 2

Background and physics

This chapter discusses some basic aspects of thermodynamics which will be useful in this thesis. Thermodynamics is based on two general laws in nature, the first and the second law of thermodynamics. Based on these laws other observed properties of a given system can be derived. Since in this thesis we are dealing with both compressible and incompressible flows, we are interested in expressing or relating different thermodynamic properties.

In this chapter we only consider isolated systems, that is, systems that cannot exchange energy with the surroundings and are in a state of *thermodynamic equilibrium.* The quantities whose values determine the thermodynamic state of a system are called *state variables.* If a system is in thermodynamic equilibrium, all thermodynamic variables are defined and have a unique value. When we change the state variables, the state of the system will change and this change is called a *process.* There are two kinds of processes: *reversible* and *irreversible.* When studying the thermodynamic properties of a system, we mainly consider reversible processes, however, in reality a process is never truly reversible. The thermodynamic properties of a system can be of two types: *intensive* or *extensive.* An intensive variable is one whose value is independent of the mass of the system. Examples are pressure, temperature and density. On the other hand extensive variables are for example the volume, the internal energy or the entropy of a system. When an extensive variable is divided by the mass, it becomes an intensive variable and the ratio is called the *specific value* of that variable. In this chapter we will study both the intensive and extensive properties of some relevant variables and we will see that in many cases it is more convenient to work with the intensive form of the thermodynamic equations. For notational simplicity, we shall use capital letters for extensive variables (except the temperature) and small letters for the corresponding specific values.

2.1 Thermodynamics

There are two approaches to study thermodynamics, one is the classical way and the other is the axiomatic way. The two main concepts of the classical theory are the law expressing conservation of energy, or the first law of thermodynamics, and the second one relating thermodynamic properties to each other, the second law of thermodynamics. The key concept which enables us to relate thermodynamic properties is *entropy.* In the classical approach, entropy is defined using the first law combined with the concept of a reversible process. This is, however, too restrictive, since completely reversible processes do not exist in real life. Here comes the benefit of the axiomatical approach, which is emphasized in [44]. The fundamental assumption in the axiomatical approach is equivalent to the second law of thermodynamics. One of the key benefits of this approach is that it enables us to examine the thermodynamic state properties without explicitly using the first law. In this thesis we will follow the axiomatic approach.

2.2 Fundamental thermodynamic relations

When a system is in a state of thermodynamic equilibrium, all thermodynamic properties have definite and unique values. From experiments, it follows that fixing three extensive independent properties will determine the thermodynamic state of a system containing a single substance without chemical reactions. It is, however, important to notice that three arbitrary variables will not determine the thermodynamic equilibrium of the system, but these three variables must form an independent set. We will see later in this section that if we are interested in the intensive state of the system, only two independent intensive properties are needed. Following the approach in [44], we choose the following independent variables: the internal energy E, the volume V, and the number of moles N of the substance, all written in their extensive forms. Choosing values for E, V, and N fixes the thermodynamic state and determines the value of all other thermodynamic properties. Consider the following function for the entropy:

$$\text{(FE)} \qquad S = S(E, V, N). \tag{2.2.1}$$

The entropy S is the dependent variable in (2.2.1) and since it contains all the thermodynamic information about the substance it is called the *fundamental thermodynamic equation* for the system. This means that if this function is known, all thermodynamic properties can be found. Note that this is not true for any thermodynamic function. Assume now that we know the function $S = S(E, V, N)$. The total differential of the entropy is given by

$$\mathrm{d}S = \left(\frac{\partial S}{\partial E}\right)_{V,N} \mathrm{d}E + \left(\frac{\partial S}{\partial V}\right)_{E,N} \mathrm{d}V + \left(\frac{\partial S}{\partial N}\right)_{E,V} \mathrm{d}N. \tag{2.2.2}$$

The partial derivatives of this function are used to define the following thermodynamic variables:

$$\text{temperature} \qquad T = \left(\left(\frac{\partial S}{\partial E} \right)_{V,N} \right)^{-1} \tag{2.2.3}$$

$$\text{pressure} \qquad p = T \left(\frac{\partial S}{\partial V} \right)_{E,N} \tag{2.2.4}$$

$$\text{chemical potential} \qquad \tilde{\mu} = -T \left(\frac{\partial S}{\partial N} \right)_{E,V}. \tag{2.2.5}$$

A substance has three *equations of state*, $T = T(E, V, N)$, $p = p(E, V, N)$ and $\tilde{\mu} = \tilde{\mu}(E, V, N)$, which follow from (2.2.3), (2.2.4) and (2.2.5). By substituting the three equations of state into (2.2.2), we define the *fundamental differential equation of thermodynamics*

$$\text{(FDE)} \qquad \boxed{T \mathrm{d}S = \mathrm{d}E + p\,\mathrm{d}V - \tilde{\mu}\,\mathrm{d}N.} \tag{2.2.6}$$

Different substances can have different fundamental equations, but they will all satisfy the FDE given in (2.2.6). The most common way of relating thermodynamical properties is to give the equations of state of the system and then find all the properties of our interest. In this sense, it is important to notice that if we know the three equations of state, this is equivalent to knowing the fundamental equation. Consequently, all other thermodynamic properties can be determined. Due to its importance and since we will often use this later in this thesis, we illustrate this equivalence shortly. For details of the proof we refer to [44]. Using the property that the entropy S is a homogeneous function of degree one, we obtain the relation

$$\left(\frac{\partial S}{\partial E} \right)_{V,N} E + \left(\frac{\partial S}{\partial V} \right)_{E,N} V + \left(\frac{\partial S}{\partial N} \right)_{E,V} N = S. \tag{2.2.7}$$

Substituting the definition of the equations of state (2.2.3), (2.2.4) and (2.2.5) into (2.2.7), we obtain the so-called Euler equation

$$S = \frac{1}{T} E + \frac{p}{T} V - \frac{\tilde{\mu}}{T} N. \tag{2.2.8}$$

If we know the three equations of state for T, p and $\tilde{\mu}$, we can determine the fundamental equation by substituting them into (2.2.8). Hence, (2.2.8) shows the equivalence between the fundamental equation (2.2.1) and the three equations of state.

An important consequence of Euler's equation is that the three equations of state are not completely independent. When we differentiate Euler's equation (2.2.8) and subtract the FDE (2.2.6), we obtain

$$E\,\mathrm{d}\left(\frac{1}{T}\right) + V\mathrm{d}\left(\frac{p}{T}\right) - N\mathrm{d}\left(\frac{\tilde{\mu}}{T}\right) = 0. \tag{2.2.9}$$

Assuming that two equations of state are given, the third one can be obtained by integrating (2.2.9).

Remark 2.2.1 *Knowledge of two equations of state determines all thermodynamic information with the exception of a constant.*

Remark 2.2.2 *The most important issue concerning the axiomatic approach is that the major characteristics of entropy can be found only after we state the second law of thermodynamics. More precisely, the definition of thermodynamic properties by using the derivatives of the entropy can only be validated using the second law of thermodynamics. For further details on the validation we refer to [44].*

Next, we formulate thermodynamic properties and relations in their intensive forms by employing a unit-mass basis. Any intensive property x can be obtained from its extensive form X as

$$x = \frac{X}{\mathcal{M}N}$$

where $\mathcal{M}$ is the mass of one mole of substance. Since S is a homogeneous function of degree one, we obtain for the entropy per unit mass

$$s = \frac{S}{\mathcal{M}N} = S\left(e, v, \frac{1}{\mathcal{M}}\right) = s(e, v)$$

where $\mathcal{M}$ has been absorbed in the last equality. This shows that only two intensive properties define the thermodynamic state of the system. Consider the fundamental equation for the system, now in the intensive form

$$\text{(FE)} \qquad s = s(e, v). \tag{2.2.10}$$

The fundamental differential equation for the intensive function $s(e, v)$ is

$$\text{(FDE)} \qquad \mathrm{d}s = \frac{1}{T}\mathrm{d}e + \frac{p}{T}\,\mathrm{d}v. \tag{2.2.11}$$

Note here that (2.2.11) is also known as the fundamental equation of Gibbs and in classical thermodynamics it is derived for a reversible process from the first and second laws.

There are two intensive equations of state corresponding to (2.2.3) and (2.2.4):

$$\text{EOS1}: \quad \frac{1}{T} = \left(\frac{\partial s}{\partial e}\right)_v, \quad \text{that is} \quad T = T(e, v) \tag{2.2.12}$$

and

$$\text{EOS2}: \quad \frac{p}{T} = \left(\frac{\partial s}{\partial v}\right)_e, \quad \text{that is} \quad p = p(e, v). \tag{2.2.13}$$

The third equation of state cannot be derived directly from $s(e, v)$, but we can substitute the intensive forms into Euler's equation (2.2.8) and obtain

$$\frac{\tilde{\mu}}{\mathcal{M}} = e + pv - Ts = h - Ts, \tag{2.2.14}$$

where $h = e + pv$ denotes the specific enthalpy. Dividing the chemical potential by $\mathcal{M}$ changes it from a mole to a unit-mass basis. Note that for the remainder of this thesis we will use the unit-mass basis and for simplicity we denote it by $\tilde{\mu}$.

In order to define the incompressible limit of the Navier-Stokes equations two quantities need to be defined. These are the *volume expansivity* α_p and the *isothermal compressibility* β_T, which are defined as:

$$\alpha_p = \frac{1}{v}\left(\frac{\partial v}{\partial T}\right)_p, \ \beta_T = -\frac{1}{v}\left(\frac{\partial v}{\partial p}\right)_T. \tag{2.2.15}$$

Given the fundamental equation (2.2.10), we can determine the specific heat at constant pressure c_p and the specific heat at constant volume c_v, which are defined as:

$$c_p(T, p) = T\left(\frac{\partial s}{\partial T}\right)_p, \tag{2.2.16}$$

and

$$c_v(T, v) = T\left(\frac{\partial s}{\partial T}\right)_v. \tag{2.2.17}$$

In the next section we will show that α_p, β_T and c_p (or c_v) completely define the equilibrium thermodynamic state of a single substance. As a preliminary step towards this, in the remainder of this section we derive an important relation between the specific heats at constant pressure c_p and constant volume c_v using the definitions (2.2.15). Let us choose (v, T) as independent variables and introduce the Helmholtz free energy

$$f = f(v, T) = e - Ts. \tag{2.2.18}$$

Differentiation of (2.2.18) gives

$$\mathrm{d}f = \mathrm{d}e - s\mathrm{d}T - T\mathrm{d}s, \tag{2.2.19}$$

and introducing the FDE (2.2.11) results in

$$\mathrm{d}f = -p\mathrm{d}v - s\mathrm{d}T. \tag{2.2.20}$$

On the other hand, the total differential of f is given by

$$\mathrm{d}f = \left(\frac{\partial f}{\partial v}\right)_T \mathrm{d}v + \left(\frac{\partial f}{\partial T}\right)_v \mathrm{d}T. \tag{2.2.21}$$

Comparing the coefficients in (2.2.20) with the ones in (2.2.21), we obtain

$$s = -\left(\frac{\partial f}{\partial T}\right)_v, \tag{2.2.22}$$

$$p = -\left(\frac{\partial f}{\partial v}\right)_T \tag{2.2.23}$$

and the derivatives

$$\left(\frac{\partial s}{\partial v}\right)_T = -\frac{\partial^2 f}{\partial T \partial v}, \quad \left(\frac{\partial p}{\partial T}\right)_v = -\frac{\partial^2 f}{\partial T \partial v}. \tag{2.2.24}$$

Hence

$$\left(\frac{\partial s}{\partial v}\right)_T = \left(\frac{\partial p}{\partial T}\right)_v. \tag{2.2.25}$$

Let us choose now (p, T) to be the independent variables. Using Euler's equation (2.2.14) we can consider

$$\tilde{\mu} = \tilde{\mu}(p, T) = h - Ts.$$

The total differential of the enthalpy can then be written in the form

$$\begin{aligned} \mathrm{d}h &= \mathrm{d}\tilde{\mu} + s\mathrm{d}T + T\mathrm{d}s \\ &= \left(\frac{\partial \tilde{\mu}}{\partial p}\right)_T \mathrm{d}p + \left(\left(\frac{\partial \tilde{\mu}}{\partial T}\right)_p + s\right)\mathrm{d}T + T\mathrm{d}s. \end{aligned} \tag{2.2.26}$$

On the other hand,

$$\mathrm{d}h = \mathrm{d}e + p\mathrm{d}v + v\mathrm{d}p = T\mathrm{d}s + v\mathrm{d}p \tag{2.2.27}$$

where in the last equality we used the FDE. Comparing the coefficients in (2.2.26) with the ones in (2.2.27), we obtain

$$v = \left(\frac{\partial \tilde{\mu}}{\partial p}\right)_T,$$

$$s = -\left(\frac{\partial \tilde{\mu}}{\partial T}\right)_p.$$

Therefore, the volume expansivity can be expressed as

$$\alpha_p = \frac{1}{v}\left(\frac{\partial v}{\partial T}\right)_p = \frac{1}{v}\left(\frac{\partial^2 \tilde{\mu}}{\partial p \partial T}\right)_T = -\frac{1}{v}\left(\frac{\partial s}{\partial p}\right)_T. \tag{2.2.28}$$

The definitions (2.2.15), combined with (2.2.25) (2.2.28) gives the relation:

$$\left(\frac{\partial p}{\partial T}\right)_v = \left(\frac{\partial s}{\partial v}\right)_T = \frac{\left(\frac{\partial s}{\partial v}\right)_T \left(\frac{\partial v}{\partial p}\right)_T}{\left(\frac{\partial v}{\partial p}\right)_T} = \frac{\left(\frac{\partial s}{\partial p}\right)_T}{\left(\frac{\partial v}{\partial p}\right)_T} = \frac{\alpha_p}{\beta_T}. \tag{2.2.29}$$

Next we derive a relation between the specific heats and the volume expansivity α_p and isothermal compressibility β_T. For this, we first express the total derivative of the internal energy $e = e(v, T)$

$$\mathrm{d}e = \left(\frac{\partial e}{\partial T}\right)_v \mathrm{d}T + \left(\frac{\partial e}{\partial v}\right)_T \mathrm{d}v = c_v \mathrm{d}T + \left(\frac{\partial e}{\partial v}\right)_T \mathrm{d}v, \tag{2.2.30}$$

where the second equality follows from

$$\left(\frac{\partial s}{\partial T}\right)_v = \frac{1}{T}\left(\frac{\partial e}{\partial T}\right)_v = \frac{1}{T} c_v(T, v), \tag{2.2.31}$$

which is the FDE combined with definition (2.2.17). Using the differential of the Helmholtz free energy in combination with (2.2.23) and (2.2.29), we obtain

$$\left(\frac{\partial e}{\partial v}\right)_T = \left(\frac{\partial f}{\partial v}\right)_T + T\left(\frac{\partial s}{\partial v}\right)_T = -p + T\frac{\alpha_p}{\beta_T}. \tag{2.2.32}$$

Inserting (2.2.30) into the FDE, results in

$$\mathrm{d}s = \frac{1}{T} c_v \mathrm{d}T + \frac{1}{T}\left(\frac{\partial e}{\partial v}\right)_T \mathrm{d}v + \frac{p}{T}\mathrm{d}v,$$

which we can use in the definition of c_p to obtain the relation

$$c_p = c_v + \left(\left(\frac{\partial e}{\partial v}\right)_T + p\right)\left(\frac{\partial v}{\partial T}\right)_p. \tag{2.2.33}$$

Combining the above equation with (2.2.32) and using the definition of α_p, we obtain

$$c_p - c_v = \left(\left(\frac{\partial e}{\partial v}\right)_T + p\right) v\alpha_p = \frac{\alpha_p^2 v T}{\beta_T}. \tag{2.2.34}$$

2.2.1 Incompressible fluid

In this section special attention is given to an incompressible fluid. For a completely incompressible fluid both α_p and β_T, defined in (2.2.15), are zero. First we consider the speed of sound a, given by

$$a = \sqrt{\left(\frac{\partial p}{\partial \rho}\right)_s} = \sqrt{-v^2\left(\frac{\partial p}{\partial v}\right)_s}$$

where $\rho = 1/v$ is the density of the fluid. We express the speed of sound in terms of the isothermal compressibility β_T using the cyclic rule

$$\left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial z}\right)_x \left(\frac{\partial z}{\partial x}\right)_y = -1. \tag{2.2.35}$$

Therefore, from

$$\left(\frac{\partial s}{\partial T}\right)_v \left(\frac{\partial T}{\partial v}\right)_s \left(\frac{\partial v}{\partial s}\right)_T = -1$$

in combination with (2.2.31), it follows that

$$c_v = -T\left(\frac{\partial v}{\partial T}\right)_s \left(\frac{\partial s}{\partial v}\right)_T.$$

Similarly, from

$$\left(\frac{\partial s}{\partial T}\right)_p \left(\frac{\partial T}{\partial p}\right)_s \left(\frac{\partial p}{\partial s}\right)_T = -1,$$

and (2.2.16), we can express c_p as

$$c_p = -T\left(\frac{\partial p}{\partial T}\right)_s \left(\frac{\partial s}{\partial p}\right)_T.$$

The dimensionless variable γ can then be written as

$$\gamma = \frac{c_p}{c_v} = \frac{\left(\frac{\partial p}{\partial T}\right)_s \left(\frac{\partial s}{\partial p}\right)_T}{\left(\frac{\partial v}{\partial T}\right)_s \left(\frac{\partial s}{\partial v}\right)_T} = \left(\frac{\partial p}{\partial v}\right)_s \left(\frac{\partial v}{\partial p}\right)_T = \frac{a^2\beta_T}{v}$$

Therefore,

$$a^2 = \frac{\gamma}{\rho\beta_T}, \tag{2.2.36}$$

which leads to an infinite speed of sound in an incompressible fluid. For an incompressible fluid the density is constant and not a thermodynamic variable. The thermodynamic theory in this approximation has one less independent variable than for a simple compressible substance. The fundamental differential equation of an incompressible fluid is

$$\text{(FDEI)} \quad \mathrm{d}s = \frac{1}{T}\mathrm{d}e. \tag{2.2.37}$$

From the FDEI (2.2.37)

$$\left(\frac{\partial s}{\partial T}\right)_v = \frac{1}{T}\left(\frac{\partial e}{\partial T}\right)_v = \frac{1}{T}c_v(T, v),$$

where in the last equality we used (2.2.31). Since for an incompressible fluid the density is constant, the specific heat is $c_v = c_v(T)$. Therefore,

$$e = \int_{T_0}^{T} c_v(\tilde{T})\mathrm{d}\tilde{T},$$

and if we assume that c_v is constant, then $e = e_0 + c_v T$. Consequently, for incompressible fluids, the internal energy is a function of the temperature only. Using the thermal equation of state we can find the fundamental equation as

$$\frac{1}{T} = \left(\frac{\partial s}{\partial e}\right)_v = \left(\frac{\partial s}{\partial T}\right)_v \left(\frac{\partial T}{\partial e}\right)_v = \left(\frac{\partial s}{\partial T}\right)_v \frac{1}{c_v}, \;\Rightarrow\; s = s_0 + \int_{T_0}^{T} \frac{c_v(\tilde{T})}{\tilde{T}}\mathrm{d}\tilde{T}, \tag{2.2.38}$$

that is $s = s_0 + c_v \ln T$ for constant c_v. The pressure equation of state is obtained from (2.2.13), however, for an incompressible fluid the thermodynamic pressure is zero because s is independent of v. There is a pressure in an incompressible fluid, but it does not influence the thermodynamic state and is only a mechanical variable. This implies that other previously defined variables which involve the pressure, are no longer a purely thermodynamic variable. For instance, the enthalpy or the chemical potential,

$$h = e + pv, \quad \tilde{\mu} = \int_{T_0}^{T} c_v(\tilde{T})\mathrm{d}\tilde{T} + pv - s_0 T - T\int_{T_0}^{T} \frac{c_v(\tilde{T})}{\tilde{T}}\mathrm{d}\tilde{T},$$

and for constant c_v, $\tilde{\mu} = e + pv - T(s_0 + c_v \ln T)$. A conclusion that follows is that the incompressible fluid has only one specific heat:

$$c_p = \left(\frac{\partial h}{\partial T}\right)_p = \frac{\partial(e+pv)}{\partial T} = \frac{\mathrm{d}e}{\mathrm{d}T} = c_v. \tag{2.2.39}$$

Therefore, the ratio of specific heats, $\gamma = c_p/c_v$, is unity for an incompressible fluid.

2.3 General equation of state

In the general discussion of thermodynamic states of materials in Section 2.2, we showed that by knowing the FE, or equivalently the equations of state, we can express all thermodynamic quantities. For different materials, a different FE is valid. The question is, can we give a general form of the FE, such that by taking the incompressible limit, the FE is well defined and describes an incompressible fluid. The incompressible limit is defined using two measurable quantities, α_p and β_T. Our aim is to employ them in the FE.

In this section, we give a general form of the FE that can be used for both compressible and incompressible flows. We state that the thermodynamic state of a single species material is determined by three measurable quantities, α_p, β_T and c_p, defined in (2.2.15) and (2.2.16), respectively.

Consider p and T as independent variables. The total differential of the specific volume can be expressed as:

$$\mathrm{d}v = \left(\frac{\partial v}{\partial T}\right)_p \mathrm{d}T + \left(\frac{\partial v}{\partial p}\right)_T \mathrm{d}p = \alpha_p v\ \mathrm{d}T - \beta_T v\ \mathrm{d}p. \tag{2.3.1}$$

Integrating (2.3.1), we obtain

$$\ln v = \ln v_0 + \int_{T_0}^{T} \alpha_p(p_0, \tilde{T})\mathrm{d}\tilde{T} - \int_{p_0}^{p} \beta_T(\tilde{p}, T)\mathrm{d}\tilde{p}. \tag{2.3.2}$$

where $v_0 = v(p_0, T_0)$. Then,

$$v(p,T) = v_0 \exp\left(\int_{T_0}^{T} \alpha_p(p_0,\tilde{T})\mathrm{d}\tilde{T} - \int_{p_0}^{p} \beta_T(\tilde{p},T)\mathrm{d}\tilde{p}\right). \tag{2.3.3}$$

Hence, if α_p and β_T are known, for instance through measurements, then using (2.3.3) we can determine the specific volume as a function of p and T.

Next, we express the entropy in terms of α_p, β_T and c_p. The total differential of the entropy is

$$\mathrm{d}s = \left(\frac{\partial s}{\partial T}\right)_p \mathrm{d}T + \left(\frac{\partial s}{\partial p}\right)_T \mathrm{d}p = \frac{c_p}{T}\mathrm{d}T - \alpha_p v \ \mathrm{d}p, \tag{2.3.4}$$

where we used the relations (2.2.16) and (2.2.28). Integrate (2.3.4) and using the relation (2.3.3) for v, we obtain

$$\begin{aligned} s(p,T) = s_0 &+ \int_{T_0}^{T} \frac{c_p(p_0,\tilde{T})}{\tilde{T}}\mathrm{d}\tilde{T} \\ &- v_0 \int_{p_0}^{p} \alpha_p(\tilde{p},T) \exp\left(\int_{T_0}^{T} \alpha_p(p_0,\tilde{T})\mathrm{d}\tilde{T} - \int_{p_0}^{\tilde{p}} \beta_T(\hat{p},T)\mathrm{d}\hat{p}\right)\mathrm{d}\tilde{p}, \end{aligned} \tag{2.3.5}$$

with $s_0 = s(p_0, T_0)$. Therefore, s is defined as a function of the measurable quantities α_p, β_T and c_p.

By taking the limit $\alpha_p = \beta_T = 0$, we obtain for an incompressible substance the relation

$$s^{inc}(p,T) = s_0 + \int_{T_0}^{T} \frac{c_p(p_0,\tilde{T})}{\tilde{T}}\mathrm{d}\tilde{T},$$

which is only a function of T, that is

$$s^{inc}(T) = s_0 + \int_{T_0}^{T} \frac{c_p(p_0,\tilde{T})}{\tilde{T}}\mathrm{d}\tilde{T}. \tag{2.3.6}$$

This also implies that for incompressible flows the specific heats are equal,

$$c_v = T\left(\frac{\partial s}{\partial T}\right)_v = c_p. \tag{2.3.7}$$

Summarizing, given α_p, β_T and c_p, the entropy can be expressed as a function of pressure and temperature or only temperature for the incompressible case. Given the general expression for the entropy (2.3.5), all other thermodynamic variables can be determined in terms of α_p, β_T and c_p, with the independent variables p and T.

Next, we show that when there is a phase change, the limit is still valid [41]. Assume that α_p and β_T are constant in a range $[T_0, T]$ and $[p_0, p]$, with $\alpha_p(p,T) \approxeq \bar{\alpha}_p$ and $\beta_T(p,T) \approxeq \bar{\beta}_T$. Then,

$$\exp\left(\int_{T_0}^{T} \alpha_p(p_0,\tilde{T})\mathrm{d}\tilde{T} - \int_{p_0}^{\tilde{p}} \beta_T(\hat{p},T)\mathrm{d}\hat{p}\right)\mathrm{d}\tilde{p} \approxeq \exp\left(\bar{\alpha}_p(T-T_0) - \bar{\beta}_T(\tilde{p}-p_0)\right)$$

and

$$\begin{aligned}
&\int_{p_0}^{p} \alpha_p(\tilde{p},T)\exp\left(\int_{T_0}^{T}\alpha_p(p_0,\tilde{T})\mathrm{d}\tilde{T}-\int_{p_0}^{\tilde{p}}\beta_T(\hat{p},T)\mathrm{d}\hat{p}\right)\mathrm{d}\tilde{p}\\
&\approxeq \int_{p_0}^{p}\bar{\alpha}_p\exp\left(\bar{\alpha}_p(T-T_0)-\bar{\beta}_T(\tilde{p}-p_0)\right)\mathrm{d}\tilde{p}\\
&= \left(-\frac{\bar{\alpha}_p}{\bar{\beta}_T}\exp\left(\bar{\alpha}_p(T-T_0)-\bar{\beta}_T(\tilde{p}-p_0)\right)\right)\Big|_{\tilde{p}=p_0}^{p}\\
&= \frac{\bar{\alpha}_p}{\bar{\beta}_T}\exp(\bar{\alpha}_p(T-T_0))[1-\exp(-\bar{\beta}_T(p-p_0))].
\end{aligned}$$

Hence, in this temperature and pressure range, the entropy is equal to:

$$\begin{aligned}
s(p,T) = s_0 &+ \int_{T_0}^{T}\frac{c_p(p_0,\tilde{T})}{\tilde{T}}\mathrm{d}\tilde{T}\\
&- v_0\frac{\bar{\alpha}_p}{\bar{\beta}_T}\exp(\bar{\alpha}_p(T-T_0))[1-\exp(-\bar{\beta}_T(p-p_0))].
\end{aligned}$$

Note that if this range is the incompressibility regime, that is $\alpha_p \to 0$ and $\beta_T \to 0$, then, using

$$\lim_{\bar{\beta}_T\to 0}\frac{1-\exp(-\bar{\beta}_T(p-p_0))}{\bar{\beta}_T} = p-p_0,$$

we obtain

$$\lim_{\bar{\beta}_T\to 0} s(p,T) = s^{inc}(T).$$

The next important remark shows that the limit $\alpha_p \to 0$ and $\beta_T \to 0$ is reached in a certain way. Recall the thermodynamic relation between the specific heats (2.2.34) derived in Section 2.2, valid for any type of material. In the incompressible limit, $c_p = c_v$, therefore, for $\alpha_p = 0$ and $\beta_T = 0$, we have

$$\frac{\alpha_p^2 vT}{\beta_T} = 0. \tag{2.3.8}$$

Since v and T are bounded, the above relation is only possible when $O(\alpha_p^2) = O\left(\beta_T^{1+\epsilon}\right)$ for $\epsilon > 0$, as $\beta_T \to 0$. Equivalently,

$$O(\alpha_p) = O(\beta_T^{\delta}), \quad \text{with } \delta = \frac{1}{2}+\frac{\epsilon}{2},\ \epsilon > 0. \tag{2.3.9}$$

We conclude that the limiting behavior of $\alpha_p \to 0$ and $\beta_T \to 0$ must satisfy (2.3.9), hence the the limit of the two incompressibility parameters is not independent from each other.

2.4 Equations of state

Since the numerical algorithm discussed in this thesis has its roots also in compressible flows and to indicate its wide range of applicability, we discuss now several equations of state, both for gases and substances undergoing phase change. A detailed overview of equations of state can be found in [49].

The most frequently used thermal equation of state (EOS) is the *perfect gas* equation of state, in which the volume occupied by the molecules is negligible and in which the intermolecular forces are not taken into account. The perfect gas EOS is, however, not applicable in all conditions. For a gas at conditions of high pressure ($\approx 1000\,atm$), where intermolecular forces can no longer be neglected, it means that a *real gas* EOS should be used.

In Section 2.2, we saw that knowing the fundamental equation is equivalent to knowing the equations of state of a substance. In the next sections we give some examples of this equivalence for both perfect and real gases.

2.4.1 The perfect gas

At room temperature air is essentially a calorically perfect gas. It remains so until the temperature reaches approximately $600K$. Then, if temperature increases further, vibrational excitation becomes important, and air behaves as a thermally perfect gas. Above $200K$, chemical reactions occur and air becomes a chemically reacting mixture of perfect gases.

The *fundamental equation for a perfect gas* with the assumption of a constant specific heat c_v, is

$$s = s_0 + c_v \ln \frac{e}{e_0} - R \ln \frac{v_0}{v} \tag{2.4.1}$$

where v_0 and e_0 correspond to a reference state and R is the gas constant. Then, by definition the EOS2 (2.2.13) has the explicit form

$$\text{EOS2}: \quad pv = RT. \tag{2.4.2}$$

Since c_v is assumed to be constant, from (2.4.1) and the definition of EOS1 (2.2.12) we obtain

$$\frac{1}{T} = \left(\frac{\partial s}{\partial e}\right)_v = \frac{c_v}{e} \quad \Longrightarrow \quad e = c_v T.$$

Consequently, at constant temperature, the internal energy of a perfect gas is independent of the specific volume. It is independent of the pressure also, since

$$\left(\frac{\partial e}{\partial p}\right)_T = \left(\frac{\partial e}{\partial v}\right)_T \left(\frac{\partial v}{\partial p}\right)_T = 0.$$

The internal energy of a perfect gas is therefore a function of its temperature only when c_v is constant.

Using (2.2.15) and (2.4.2), we obtain $\alpha_p = 1/T$, $\beta_T = 1/p$ and combined with (2.2.34) this means that $R = c_p - c_v$. The above relations imply $pv = (\gamma - 1)e$, which is the ideal gas equation of state, where $\gamma = c_p/c_v$. Furthermore, the entropy in the fundamental equation (2.4.1) can be written as

$$s = c_v \ln\left(\frac{p}{\rho^\gamma}\right) + \underline{s_0 - c_v \ln e_0 + c_v \ln\left(\frac{\rho_0^{\gamma-1}}{\gamma - 1}\right)} \tag{2.4.3}$$

where we denote the underlined coefficient by s_0. The chemical potential $\tilde{\mu}$, is

$$\tilde{\mu}(p, T) = c_v T + RT - s_0 T + c_v T \ln\left(\frac{p}{\rho^\gamma}\right). \tag{2.4.4}$$

Note that, in the remainder of this thesis the reference values for e_0, ρ_0, s_0, etc. are not necessarily always the same. This prevents extensive relations for the coefficients.

2.4.2 Real gas

In practice, a gas behaves as a real gas under conditions of high pressure and moderate temperature, conditions which show the influence of intermolecular forces on the thermodynamics state of the system. The most familiar real gas thermal EOS are the Clausius and the van der Waals equations.

In this section we proceed reversely compared to the previous section. Given the EOS of a substance, we determine the fundamental equation.

Co-volume equation of state. First, we consider the generalization of the perfect gas EOS, the so-called Clausius or co-volume equation of state

$$p(v - b) = RT \tag{2.4.5}$$

where b is called the co-volume. Let us find the fundamental equation for this equation of state. Using (2.2.15), we obtain

$$\alpha_p = \frac{v - b}{Tv}, \quad \beta_T = \frac{(v - b)^2}{vRT}. \tag{2.4.6}$$

From (2.2.34) it follows that $c_p - c_v = R$. Differentiating (2.2.29) with respect to T at constant volume and (2.2.31) with respect to v at constant T, and equating the two results yields

$$\left(\frac{\partial c_v}{\partial v}\right)_T = T\left(\frac{\partial^2 p}{\partial T^2}\right)_v = T\left(\frac{\partial}{\partial T}\left(\frac{\alpha_p}{\beta_T}\right)\right)_v = T\left(\frac{\partial}{\partial T}\left(\frac{R}{v - b}\right)\right)_v = 0. \tag{2.4.7}$$

This means that c_v is a function of T only, i.e., $c_v = c_v(T)$. By combining (2.2.32) with (2.4.6), we obtain from (2.4.5) that $\left(\frac{\partial e}{\partial v}\right)_T = 0$. Hence, $e = e(T)$. From (2.2.31) it follows that

$$e(T) = \int_{T_0}^{T} c_v(\tilde{T})\mathrm{d}\tilde{T} + e_0 \tag{2.4.8}$$

where e_0 is the internal energy at T_0. Assume that c_v is constant. Then, $e = c_v T + e_0$. Using again (2.2.31), the specific entropy has the form

$$s = s_0 + c_v \ln T + g(v) \tag{2.4.9}$$

where s_0 is a reference entropy and $g(v)$ is a function depending only on the volume v. This function g can be uniquely determined, up to a constant, by solving the following ordinary differential equation obtained from (2.2.29):

$$\frac{\mathrm{d}}{\mathrm{d}v} g(v) = \frac{R}{v-b}. \tag{2.4.10}$$

Therefore, the entropy (or the fundamental equation for a real gas) can be calculated as

$$s = s_0 + c_v \ln T + R \ln(v-b). \tag{2.4.11}$$

Now we have all variables to calculate the chemical potential $\tilde{\mu}$ using Euler's equation (2.2.14) which yields:

$$\tilde{\mu}(p,T) = e_0 + c_v T + pb + RT - T\left(s_0 + c_v \ln T + R \ln \frac{RT}{p}\right). \tag{2.4.12}$$

Van der Waals equation of state. The second equation of state which we consider is the van der Waals equation

$$p = \frac{RT}{v-b} - \frac{a}{v^2} \tag{2.4.13}$$

where the term a/v^2 is a correction that accounts for the intermolecular forces of attraction. The compressibility coefficients have the following values

$$\alpha_p = \frac{(v-b)v^2 R}{v^3 RT - 2a(v-b)^2}, \quad \beta_T = \frac{(v-b)^2 v^2}{v^3 RT - 2a(v-b)^2}. \tag{2.4.14}$$

The difference between the specific heats follows from (2.2.34):

$$c_p - c_v = \frac{v^3 R^2 T}{v^3 RT - 2a(v-b)^2},$$

and from (2.2.32) we obtain

$$\left(\frac{\partial e}{\partial v}\right)_T = \frac{a}{v^2}. \tag{2.4.15}$$

Let us consider the specific internal energy expressed in terms of the independent variables T and v. Then,

$$\mathrm{d}e = \left(\frac{\partial e}{\partial T}\right)_v \mathrm{d}T + \left(\frac{\partial e}{\partial v}\right)_T \mathrm{d}v. \tag{2.4.16}$$

Inserting (2.4.15) and the definition of c_v in (2.4.16), we obtain

$$\mathrm{d}e = c_v \mathrm{d}T + \frac{a}{v^2}\mathrm{d}v$$

and after integration,

$$e = e_0 + \int_{T_0}^{T} c_v(\tilde{T})\mathrm{d}\tilde{T} + a\left(\frac{1}{v_0} - \frac{1}{v}\right)$$

where $e_0 = e(T_0, v_0)$ is the specific internal energy at a reference temperature T_0 and a reference specific volume v_0.

Next, we determine the specific entropy for constant c_v. Consider $s = s(v, T)$. The total differential of s is

$$\mathrm{d}s = \left(\frac{\partial s}{\partial v}\right)_T \mathrm{d}v + \left(\frac{\partial s}{\partial T}\right)_v \mathrm{d}T.$$

Inserting (2.4.16) in the FDE and using (2.2.31) and (2.4.15) we obtain

$$\mathrm{d}s = \frac{1}{T}c_v \mathrm{d}T + \frac{1}{T}\left(\frac{a}{v^2} + p\right)\mathrm{d}v.$$

Comparing the coefficients in the two relations above, we conclude that

$$\left(\frac{\partial s}{\partial T}\right)_v = \frac{1}{T}c_v \Rightarrow s = c_v \ln T + g_1(v)$$

and

$$\left(\frac{\partial s}{\partial v}\right)_T = \frac{1}{T}\left(\frac{a}{v^2} + p\right) = \frac{R}{v-b} \Rightarrow s = R\ln(v-b) + g_2(T).$$

Combining both relations we obtain

$$s = s_0 + c_v \ln T + R\ln(v-b) + g_1(v) + g_2(T) \tag{2.4.17}$$

where s_0 is the entropy at the reference temperature T_0 and specific volume v_0. Using (2.2.29), it follows that

$$\left(\frac{\partial s}{\partial v}\right)_T = \frac{\alpha_p}{\beta_T} = \frac{R}{v-b}.$$

On the other hand, from (2.4.17) we have

$$\left(\frac{\partial s}{\partial v}\right)_T = \frac{R}{v-b} + \frac{\partial g_1(v)}{\partial v}.$$

Therefore, $g_1(v) = constant$. Similarly, we obtain that $g_2(T) = constant$, and consequently,

$$s = s_0 + c_v \ln T + R \ln(v - b) \tag{2.4.18}$$

for some s_0. Finally, the chemical potential for the van der Waals equation can be expressed as

$$\tilde{\mu} = e_0 + c_v T - \frac{a}{v} + pv - T\Big(s_0 + c_v \ln T + R \ln(v - b)\Big). \tag{2.4.19}$$

Remark 2.4.1 *In Appendix A some measured values of the volume expansivity and isothermal compressibility are listed for several substances.*

2.5 Concluding remarks

In this chapter we summarized the basic thermodynamic concepts that are needed to study the incompressible limit. We proposed a general form of the fundamental equation that can be used for both compressible and incompressible flows. The thermodynamic state of a single species material is determined by three measurable quantities, α_p, β_T and c_p (or c_v).

Chapter 3

A unified formulation of the Navier-Stokes equations

Over the past decades a large variety of numerical schemes has been developed for the Navier-Stokes equations, but there has not been much synergy between the different approaches for compressible and incompressible flows. Many of the ideas developed in one field can, however, be useful in other fields. An example is the concept of symmetrized equations using entropy variables in compressible flow, investigated by Godunov [18], Mock [43], Harten [21], Hughes et al. [30], Dutt [12] and Johnson et al. [37]. One of the key benefits is that the use of the symmetrized compressible Navier-Stokes equations using the entropy variables, results in a global entropy stability which is automatically inherited by the numerical discretization, see for instance Shakib et al. [51]. This is not true when for instance conservative or primitive variables are used. For a comparative study of different sets of variables for solving compressible and incompressible flows we refer to the work of Hauke and Hughes [23]. For a detailed analysis of the entropy stability of the symmetrized Navier-Stokes equations see Barth [3] and Section 5.5 of this thesis. The use of entropy variables is also important for weakly compressible flows, in particular when one is interested in the proper limiting behavior of the compressible Navier-Stokes equations in the incompressible limit, discussed in details in Chapter 5 of this thesis.

In [23] Hauke and Hughes demonstrated that, with the proper choice of variables, it is possible to obtain a symmetrized formulation of the Navier-Stokes equations which is valid both for compressible and incompressible flows and does not result in a singular limit for incompressible flow. Therefore, this makes it possible to use the same set of variables for the whole spectrum of flows and to obtain a unified numerical discretization, valid in both the compressible and incompressible flow regime, and this is the main topic of this chapter.

3.1 Compressible flow governing equations

As starting point we discuss in this section the governing equations for compressible flows in three space dimensions. Fluids obey the general laws of mechanics, namely, conservation of mass, momentum and energy. In the Eulerian representation of the flow $\rho = \rho(t,x)$ is the density at $x = (x_1, x_2, x_3)$ at time t, and $u = u(t,x)$, $u = (u_1, u_2, u_3)$ is the velocity of a fluid particle at position x at time t. Then, *conservation of mass* is expressed by the continuity equation

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho u) = 0. \tag{3.1.1}$$

The *conservation of momentum* in a continuous medium can be expressed as

$$\rho a_i = \sum_{j=1}^{3} \frac{\partial \sigma_{ij}}{\partial x_j} + f_i, \quad i = 1,2,3, \tag{3.1.2}$$

where $a = a(t,x)$, $a = (a_1, a_2, a_3)$ is the acceleration vector of the particle x at time t, $f = (f_1, f_2, f_3)$ represents the volume forces applied to the fluid and $\sigma = \sigma(t,x)$ the Cauchy stress tensor at position x at time t. The acceleration vector of the particle is expressed as

$$a = \frac{\partial u}{\partial t} + u_j \frac{\partial u}{\partial x_j} \quad \text{or} \quad a_i = \frac{\partial u_i}{\partial t} + \sum_{j=1}^{3} u_j \frac{\partial u_i}{\partial x_j}, \quad i = 1,2,3. \tag{3.1.3}$$

Introduce here the operator $\mathrm{D}/\mathrm{D}t$, representing the time derivative following the particle, usually called material derivative, and if $\varphi(t,x)$ is a scalar valued function, then it is defined as

$$\frac{\mathrm{D}\varphi}{\mathrm{D}t} = \frac{\partial \varphi}{\partial t} + u_j \frac{\partial \varphi}{\partial x_j}.$$

Consequently, $a = \mathrm{D}u/\mathrm{D}t$. In the case of a Newtonian fluid the stress tensor is expressed in terms of the velocity and the pressure $p = p(t,x)$ as

$$\sigma_{ij} = \mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \left(\lambda \sum_{i=1}^{3} \frac{\partial u_i}{\partial x_i} - p \right) \delta_{ij} \tag{3.1.4}$$

where δ_{ij} is the Kronecker delta symbol and for the viscosity coefficients $\mu > 0$, $\lambda > 0$ the Stokes hypothesis is assumed, that is $\lambda = -2\mu/3$. The equations (3.1.1) and (3.1.2) are the general Navier-Stokes equations for Newtonian compressible fluids in Eulerian representation (see [39]). For a compressible fluid the equations for conservation of mass and momentum are supplemented by an independent equation expressing the *conservation of energy.*

The energy balance in a compressible fluid is determined by the internal energy, the conduction of heat, the convection of heat with the flow, the generation of heat

through friction and the expansion (or compression) work. The energy balance can be established on the basis of the First Law of Thermodynamics (see Section 2.1). Assuming Stokes' hypothesis, the *energy equation* can be written in the form

$$\rho c_v \frac{\mathrm{D}T}{\mathrm{D}t} + p \operatorname{div} u = \sum_{i=1}^{3} \frac{\partial}{\partial x_i} \left(\kappa \frac{\partial T}{\partial x_i} \right) + \mu \Phi \tag{3.1.5}$$

where κ is the coefficient of thermal conductivity. The dissipation function Φ in (3.1.5) is defined as

$$\begin{aligned} \Phi = 2 \left[\left(\frac{\partial u_1}{\partial x_1} \right)^2 + \left(\frac{\partial u_2}{\partial x_2} \right)^2 + \left(\frac{\partial u_3}{\partial x_3} \right)^2 \right] + \left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \right)^2 + \left(\frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1} \right)^2 \\ + \left(\frac{\partial u_2}{\partial x_3} + \frac{\partial u_3}{\partial x_2} \right)^2 - \frac{2}{3} \left(\sum_{i=1}^{3} \frac{\partial u_i}{\partial x_i} \right)^2 . \end{aligned} \tag{3.1.6}$$

For further details on the energy equation we refer to [47].

In the remainder of this thesis, we call the five equations representing conservation of mass, momentum and energy of a compressible fluid the three-dimensional compressible Navier-Stokes equations. The compressible Navier-Stokes equations can be written in conservative form as

$$U_{,t} + F^a_{i,i} = F^d_{i,i} \ , \ \text{for } i = 1, 2, 3, \tag{3.1.7}$$

where $U \in \mathbb{R}^5$ is the vector of conservative variables, and $F^a_i, F^d_i \in \mathbb{R}^5$ are, respectively, the advective and diffusive fluxes in the ith Cartesian coordinate direction, which are defined as:

$$U = \begin{pmatrix} \rho \\ \rho u_1 \\ \rho u_2 \\ \rho u_3 \\ \rho e^{tot} \end{pmatrix}, \quad F^a_i = \rho u_i \begin{pmatrix} 1 \\ u_1 \\ u_2 \\ u_3 \\ e^{tot} \end{pmatrix} + p \begin{pmatrix} 0 \\ \delta_{1i} \\ \delta_{2i} \\ \delta_{3i} \\ u_i \end{pmatrix}, \quad F^d_i = \begin{pmatrix} 0 \\ \tau_{1i} \\ \tau_{2i} \\ \tau_{3i} \\ \tau_{ij} u_j \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -q_i \end{pmatrix}. \tag{3.1.8}$$

In the above expressions, e^{tot} denotes the total energy, δ_{ij} the Kronecker delta symbol, $\tau = [\tau_{ij}]$ the viscous-stress tensor, and $q = (q_1, q_2, q_3)^T$ the heat-flux vector. An inferior comma represents partial differentiation (e.g. $U_{,t} = \partial U / \partial t$, the partial derivative with respect to time and $U_{,i} = \partial U / \partial x_i$) and the summation convention on repeated indices is used. It is useful to rewrite (3.1.7) in quasi-linear form:

$$U_{,t} + A_i(U) U_{,i} = (K_{ij}(U) U_{,j})_{,i}, \tag{3.1.9}$$

where $A_i(U) = F_{i,U}$ and $K_{ij}(U) U_{,j} = F^d_i$.

The compressible Navier-Stokes equations written in terms of the conservative variables are not suitable for an energy stability analysis, since the inviscid flux Jacobian

matrices are not symmetric and the viscosity matrices are neither symmetric nor positive semi-definite. In addition, the incompressible limit of these equations is singular, hence, the equations are not a good starting point for a unified approach suitable for both compressible and incompressible flows.

A more general approach is demonstrated by Hauke and Hughes [23] using entropy variables. In the next section we discuss this approach in detail.

3.2 Symmetrizing variables

Consider the independent state variables p and T and Euler's equation for the chemical potential

$$\tilde{\mu}(p,T) = e + p/\rho - Ts. \tag{3.2.1}$$

The system (3.1.9) can be symmetrized by introducing a new set of variables, the so-called entropy variables:

$$V = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \end{pmatrix} = \frac{1}{T} \begin{pmatrix} \tilde{\mu} - |u|^2/2 \\ u_1 \\ u_2 \\ u_3 \\ -1 \end{pmatrix}. \tag{3.2.2}$$

In terms of entropy variables (3.1.9) has the form:

$$\tilde{A}_0(V)V_{,t} + \tilde{A}_i(V)V_{,i} = (\tilde{K}_{ij}(V)V_{,j})_{,i}, \tag{3.2.3}$$

with $\tilde{A}_0 = U_{,V}$, $\tilde{A}_i = A_i\tilde{A}_0$ and $\tilde{K}_{ij} = K_{ij}\tilde{A}_0$. There are two important features which characterize this form of the equations. One is that the coefficient matrices have the properties:

- $\tilde{A}_0$ is symmetric positive definite,
- $\tilde{A}_i$ is symmetric for $i = 1, 2, 3$,
- $K = [\tilde{K}_{ij}]$ is symmetric (i.e., $\tilde{K}_{ij} = \tilde{K}_{ji}^T$) and positive semidefinite.

We summarize in Appendix B.1 the complete set of Jacobian matrices $\tilde{A}_i$ and $\tilde{K}_{ij}$ for the entropy variables in terms of the volume expansivity α_p, the isothermal compressibility β_T and specific heat at constant pressure c_p. These three variables can be measured experimentally and in Section 2.3 we demonstrated that the fundamental equation for a single substance can be expressed by these three quantities together with the independent thermodynamic variables, e.g. p and T. All other thermodynamic quantities, such as c_v, e, etc. can be derived from them, but are used in the Jacobian matrices to keep the notation concise.

Secondly, the Galerkin least-squares method based on this form of the compressible Navier-Stokes equations satisfies the Clausius-Duhem inequality or entropy condition, which results in the basic non-linear stability condition for the Navier-Stokes equations, discussed in details in Section 5.5 of this thesis, or see e.g. Shakib et al. [51] and Barth [3].

Remark 3.2.1 *The Jacobian matrices $A_i(U),\ i = 1, 2, 3$ can be determined explicitly only if the fluid constitutive relations (equations of state) are specified.*

Remark 3.2.2 *The expression for the relations between the entropy and conservative variables, i.e., the mappings $U \rightarrow V$ and $V \rightarrow U$, requires an explicit formulation of the equations of state. We give these mappings in the next sections for various equations of state.*

3.3 The incompressible limit

In this section we consider the incompressible limit of the symmetrized compressible Navier-Stokes equations in the formulation with the Jacobians expressed in terms of α_p, β_T and c_p, see Appendix B.1. We also establish the link with the standard formulation of the incompressible Navier-Stokes equations and the temperature equation.

The incompressible limit of the symmetrized Navier-Stokes equations is obtained when the volume expansivity α_p and the isothermal compressibility β_T approach zero. Note that the compressibility of the fluid is defined by two variables instead of only one, the Mach number, as is frequently assumed. Hauke and Hughes [23] demonstrate that using either the entropy variables V or the primitive variables $Y = (p, u, T)^T$ results in a well defined incompressible limit, but only the entropy variables provide a formulation which satisfies the entropy condition. Any formulation containing the density, e.g. (ρ, u, T), does not have a proper incompressible limit since the coefficients in the matrices A_i either become undefined or infinitely large. Note here that the variable Y also plays an important role in our analysis of the Galerkin least-squares stabilization operator in later chapters.

The symmetrized incompressible Navier-Stokes equations are obtained by setting α_p and β_T equal to zero in the flux Jacobian matrices in (3.2.3). The Jacobian matrices $\tilde{A}_i,\ i = 0, \ldots, 3$, given in Appendix B.1, when $\alpha_p = 0$ and $\beta_T = 0$, take the following form:

$$\tilde{A}_0 = \rho T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & u_1 \\ 0 & 0 & 1 & 0 & u_2 \\ 0 & 0 & 0 & 1 & u_3 \\ 0 & u_1 & u_2 & u_3 & r \end{pmatrix}$$

$$\tilde{A}_1 = \rho T \begin{pmatrix} 0 & 1 & 0 & 0 & u_1 \\ 1 & 3u_1 & u_2 & u_3 & 2u_1^2 + e_1 \\ 0 & u_2 & u_1 & 0 & 2u_1u_2 \\ 0 & u_3 & 0 & u_1 & 2u_1u_3 \\ u_1 & 2u_1^2 + e_1 & 2u_1u_2 & 2u_1u_3 & u_1(r + 2e_1) \end{pmatrix}$$

$$\tilde{A}_2 = \rho T \begin{pmatrix} 0 & 0 & 1 & 0 & u_2 \\ 0 & u_2 & u_1 & 0 & 2u_1u_2 \\ 1 & u_1 & 3u_2 & u_3 & 2u_2^2 + e_1 \\ 0 & 0 & u_3 & u_2 & 2u_2u_3 \\ u_2 & 2u_1u_2 & 2u_2^2 + e_1 & 2u_2u_3 & u_2(r + 2e_1) \end{pmatrix}$$

$$\tilde{A}_3 = \rho T \begin{pmatrix} 0 & 0 & 0 & 1 & u_3 \\ 0 & u_3 & 0 & u_1 & 2u_1u_3 \\ 0 & 0 & u_3 & u_2 & 2u_2u_3 \\ 1 & u_1 & u_2 & 3u_3 & 2u_3^2 + e_1 \\ u_3 & 2u_1u_3 & 2u_2u_3 & 2u_3^2 + e_1 & u_3(r + 2e_1) \end{pmatrix}$$

with $k = |u|^2/2$, $r = 2k + c_pT$ and $e_1 = h + k$. The viscous flux Jacobians $\tilde{K}_{ij}$, defined in Appendix B.1, are independent of α_p and β_T and therefore do not change in the incompressible limit. Note here that in the incompressible limit $\tilde{A}_0$ becomes positive-semidefinite. We have shown in Section 2.2.1 that when the incompressible limit is taken, there is only one equation of state, $e = c_vT$, and using the Euler equation, the fundamental equation $s = s_0 + c_v \ln T$ determines the chemical potential $\tilde{\mu}(p, T) = c_vT + pv - c_vT \ln T$. Therefore, all the entries in the flux Jacobian matrices are well defined.

Next, we will show that the symmetrized Navier-Stokes equations in the incompressible limit are identical to the incompressible Navier-Stokes equations. To illustrate the main step to obtain this identity, consider the set of primitive variables Y. First, we will demonstrate that the transformation matrices $V_{,Y}$ and $Y_{,V}$ are independent of α_p and β_T. Consider the independent variables p and T. The total differential of the first component of the entropy variables V is:

$$\mathrm{d}V_1 = \frac{1}{T}\left(\left(\frac{\partial\tilde{\mu}}{\partial p}\right)_T \mathrm{d}p + \left(\frac{\partial\tilde{\mu}}{\partial T}\right)_p \mathrm{d}T - u_j \mathrm{d}u_j\right) - \frac{V_1}{T}\mathrm{d}T = \frac{1}{T}\left(\frac{1}{\rho}dp - \frac{h-k}{T}dT - u_j du_j\right),$$

where in the last equality we used $v = 1/\rho$ and that the specific volume and entropy can be expressed in terms of the derivatives of the chemical potential as

$$v = \left(\frac{\partial\tilde{\mu}}{\partial p}\right)_T, \quad s = -\left(\frac{\partial\tilde{\mu}}{\partial T}\right)_p. \tag{3.3.1}$$

Furthermore, we obtain the following transformation for the total differential of all the components of the entropy variables:

$$dV = \frac{1}{T}\begin{pmatrix} \frac{1}{\rho}dp - \frac{h-k}{T}dT - u_j du_j \\ du_1 - V_2 dT \\ du_2 - V_3 dT \\ du_3 - V_4 dT \\ -V_5 dT \end{pmatrix}. \tag{3.3.2}$$

If we assume the Stokes hypothesis, viz. $\lambda = -2/3\mu$, where λ and μ are the viscosity coefficients, and introduce the transformation (3.3.2) into (3.2.3), we obtain after some lengthy algebra the following set of equations:

- The continuity equation:

$$u_{i,i} = 0, \qquad i = 1, 2, 3. \tag{3.3.3}$$

- The momentum equations:

$$u_{i,t} + u_j u_{i,j} = -\frac{1}{\rho}p_{,i} + (\nu s_{ij})_{,j}, \quad i, j = 1, 2, 3, \tag{3.3.4}$$

  with $\nu(T) = \mu(T)/\rho$ the kinematic viscosity, and $s_{ij} = u_{i,j} + u_{j,i}$ the shear stress tensor. Equations (3.3.3)-(3.3.4) are exactly the incompressible Navier-Stokes equations.

- The equation for the temperature field is:

$$T_{,t} + u_i T_{,i} = \frac{\mu}{\rho c_p}\sum_{i,j=1}^{3}\frac{1}{2}s_{ij}^2 + \frac{1}{\rho c_p}(\kappa T_{,i})_{,i}. \tag{3.3.5}$$

Remark 3.3.1 *This set of equations includes thermally coupled and isothermal incompressible flows. Thermally coupled incompressible flows are obtained by allowing the viscosity to depend on temperature. Isothermal incompressible flows can be obtained by specifying a constant viscosity.*

Observe that in the energy equation for compressible flows (3.1.5) the term due to the work of expansion (or compression), that is $p\,\mathrm{div}\,u$, vanishes for incompressible flows. Furthermore, inserting $\mathrm{div}\,u = 0$ in the dissipation function Φ in (3.1.6), we obtain for the energy equation exactly (3.3.5).

The symmetrized compressible Navier-Stokes equations in the incompressible limit are therefore identical to the incompressible Navier-Stokes equations and the temperature equation.

We compare now the "classical" definition of incompressibility with the incompressibility introduced by setting $\alpha_p = 0$ and $\beta_T = 0$. Incompressibility is usually expressed by the condition

$$u_{i,i} = 0. \tag{3.3.6}$$

Therefore, the equation expressing the conservation of mass (3.1.1) reduces to

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} = \frac{\partial \rho}{\partial t} + u \cdot \nabla \rho = 0, \tag{3.3.7}$$

that is, an incompressible fluid is characterized by constant density along the streamlines.

In terms of primitive variables Y, the mass conservation equation can be expressed as

$$\rho\beta_T(p_{,t} + u_i p_{,i}) - \rho\alpha_p(T_{,t} + u_i T_{,i}) + \rho u_{i,i} = 0.$$

In the incompressible limit, that is when $\alpha_p = 0$ and $\beta_T = 0$, we obtain the incompressibility constraint (3.3.6). Similarly, for entropy variables, when $\alpha_p = 0$ and $\beta_T = 0$, the mass conservation equation reduces to

$$\rho T \left(\frac{\partial V_{i+1}}{\partial x_i} + u_i \frac{\partial V_5}{\partial x_i} \right) = 0,$$

which is equivalent to (3.3.6). The assumption that for an incompressible fluid the density is constant is equivalent to $\alpha_p = 0$ and $\beta_T = 0$. This equivalence follows from the definitions of α_p and β_T.

3.3.1 Discussion on the Equation of State

The quasi-linear form of the Navier-Stokes equations, both in the conservative (3.1.9) and symmetrizing variables (3.2.3) depends on the equation of state. When the equations of state are given, all thermodynamic variables, including α_p, β_T, c_p and all entries of the flux Jacobian matrices, can be determined.

When the compressible Navier-Stokes equations are considered, different equations of state can be defined, which determine uniquely the quasi-linear form of the system both for the conservative and entropy variables. When the flow is incompressible, there is only one equation of state. Therefore, in the incompressible limit the compressible equations of state are no longer valid since they do not show in general the proper limiting behavior. This can be seen in the following examples.

When an ideal gas equation of state is assumed, viz. $pv = RT$, with R the gas constant, then the volume expansivity α_p and the isothermal compressibility β_T are equal to:

$$\alpha_p = \frac{1}{T}, \ \beta_T = \frac{1}{p}, \tag{3.3.8}$$

which results in an unbounded pressure and temperature in the incompressible limit. Consequently, the ideal gas law cannot be used in the incompressible limit.

Next, consider the co-volume EOS, viz. $p(v-b) = RT$. The incompressibility parameters are then

$$\alpha_p = \frac{v-b}{Tv}, \quad \beta_T = \frac{(v-b)^2}{vRT}. \tag{3.3.9}$$

The incompressible limit is obtained by taking $\alpha_p \to 0$ and $\beta_T \to 0$, or equivalently, $v \to b$. This limit leads again to an unbounded pressure when combined with the co-volume EOS. Furthermore, from (2.2.34) it follows that $c_p - c_v = R$, whereas in (2.2.39) we have shown that in the incompressible limit $c_p = c_v$.

As a last example, consider the van der Waals equation of state (2.4.13). From (2.4.14), it follows that the incompressible limit is equivalent with $v \to b$ or $v \to 0$. In the limit $v \to b$ we obtain $c_p - c_v = R$, which again violates the condition $c_p = c_v$ in the incompressible limit. The limit $v \to 0$ is also not applicable, since it results in the unphysical limit $\rho \to \infty$.

From this discussion, in combination with the results of Section 2.2.1, we conclude that none of the perfect or real gas assumptions are valid in the incompressible limit and in case of nearly incompressible flow it is important to use the proper measured equation of state for the specific substance, for an overview see [49], instead of idealized equations of state such as the ideal gas and van der Waals equations of state.

3.4 Symmetrization using entropy variables

In this section we give some examples of mappings from conservative to entropy variables and reversely, when various equations of state are used.

Ideal gas EOS. A detailed analysis of the entropy variables for the compressible Euler and Navier-Stokes equations, when combined with the ideal gas EOS is given in [51].

Co-volume EOS. Given the co-volume equation of state (2.4.5), the mappings $U \mapsto V$ and $V \mapsto U$ can be defined as follows. Use the relation for the internal energy $e = c_v T$ and

$$U_5 = \rho e^{tot} = \rho e + \rho k$$

where $k = |u|^2/2 = (U_2^2 + U_3^2 + U_4^2)/2U_1^2$. The temperature T then can be expressed in terms of the conservative variables as

$$T = \frac{1}{c_v}\left(\frac{U_5}{U_1} - k\right).$$

Furthermore,

$$p = \frac{RT}{v-b} = \frac{R(U_5 - kU_1)}{c_v(1-b\,U_1)}, \quad h = e + pv = \frac{(U_5 - kU_1)(c_p - c_v b\,U_1)}{c_v U_1(1-b\,U_1)}$$

The entropy variables can be expressed using the conservative variables and the expression for the chemical potential $\tilde{\mu}$ (2.4.12) as follows:

$$\begin{aligned}
V_1 &= \frac{c_p - bc_vU_1}{1-bU_1} - c_v \ln\left(\frac{U_5 - kU_1}{c_vU_1}\right) + R\ln\left(\frac{1}{U_1} - b\right) - \frac{c_vkU_1}{U_5 - kU_1},\\
V_2 &= \frac{u_1}{T} = \frac{c_vU_2}{U_5 - kU_1},\\
V_3 &= \frac{u_2}{T} = \frac{c_vU_3}{U_5 - kU_1},\\
V_4 &= \frac{u_3}{T} = \frac{c_vU_4}{U_5 - kU_1},\\
V_5 &= -\frac{1}{T} = -\frac{c_vU_1}{U_5 - kU_1}.
\end{aligned}$$

The inverse mapping $V \mapsto U$ can be defined as well. First, we need to express the pressure in terms of the entropy variables. The chemical potential $\tilde{\mu}$, can be written as

$$\tilde{\mu} = -\frac{V_1}{V_5} + k(V) \tag{3.4.1}$$

where $k(V) = k = (V_2^2 + V_3^2 + V_4^2)/2V_5^2$. Introducing the Euler equation for the chemical potential (2.2.14) into (3.4.1), we obtain

$$c_pT - c_pT\ln T - TR\ln R + pb + TR\ln p = V_1T + k(V) \tag{3.4.2}$$

where $T = -1/V_5$ and $c_p - c_v = R$. To simplify the notations, we rewrite (3.4.2) as

$$pb + RT\ln p + A = 0 \tag{3.4.3}$$

where $A = c_pT - c_pT\ln T - RT\ln R - V_1T - k(V)$. Since p in (3.4.3) is only defined implicitly, we use for an explicit expression of p a series expansion with respect to the variable b about the point $b = 0$, up to order 4 :

$$p = e^{-\frac{A}{RT}}\left(1 - \frac{e^{-\frac{A}{RT}}}{RT}b + \frac{3}{2}\left(\frac{e^{-\frac{A}{RT}}}{RT}\right)^2 b^2 - \frac{8}{3}\left(\frac{e^{-\frac{A}{RT}}}{RT}\right)^3 b^3 + o(b^4)\right).$$

The motivation for the series expansion about $b = 0$ is that for b sufficiently small, we obtain in the co-volume EOS an approximation of the pressure for an ideal gas. Finally, introducing $T = -1/V_5$ into the series expansion of the pressure, we obtain

$p = p(V)$. The conservative variables in terms of the entropy variables are approximated as:

$$
\begin{aligned}
U_1 &\approxeq \frac{p(V)V_5}{p(V)V_5 b - R}, \\
U_2 &= -\frac{U_1 V_2}{V_5} \approxeq -\frac{p(V)V_2}{p(V)V_5 b - R}, \\
U_3 &= -\frac{U_1 V_3}{V_5} \approxeq -\frac{p(V)V_3}{p(V)V_5 b - R}, \\
U_4 &= -\frac{U_1 V_4}{V_5} \approxeq -\frac{p(V)V_4}{p(V)V_5 b - R}, \\
U_5 &= U_1 \left(k - \frac{c_v}{V_5} \right) \approxeq \frac{p(V)(kV_5 - c_v)}{p(V)V_5 b - R}.
\end{aligned}
$$

The van der Waals EOS. Given the van der Waals equation of state, the mappings $U \mapsto V$ and $V \mapsto U$ can be defined. The explicit definition of $V(U)$ when the van der Waals equation of state is considered is identical to the expression for the co-volume EOS, except $V_1(U)$. Using

$$
p = \frac{RT}{v - b} = \frac{R(U_5 - kU_1)}{c_v(1 - b\,U_1)} - aU_1^2, \quad h = e + pv = \frac{(U_5 - kU_1)(c_p - c_v b\,U_1)}{c_v U_1 (1 - b\,U_1)} - aU_1
$$

we can express the chemical potential $\tilde{\mu}$ in (2.4.19) in terms of the conservative variables U. We have now all the terms necessary to express $V_1(U)$ as

$$
V_1 = \frac{c_p - b c_v U_1}{1 - b U_1} - c_v \ln \left(\frac{U_5 - kU_1}{c_v U_1} \right) + R \ln \left(\frac{1}{U_1} - b \right) - \frac{c_v U_1 (k + aU_1)}{U_5 - kU_1}.
$$

The explicit expression for the inverse mapping, that is $U(V)$, becomes complicated when the van der Waals EOS is used. Therefore, we do not consider it in this thesis. Note however, that these expressions can be obtained similarly as we obtained them when the co-volume EOS was used.

Remark 3.4.1 *The incompressible limit is well defined for entropy variables but not for conservation variables. Therefore, the mappings $U \mapsto V$ and $V \mapsto U$ are not defined in the incompressible case. Instead we use the mappings $Y \mapsto V$ and $V \mapsto Y$, where $Y = (p, u_1, u_2, u_3, T)^T$.*

Remark 3.4.2 *The expression for $U(V)$ becomes very cumbersome for more complicated, general equations of state, but they are only needed for postprocessing the data of the simulation.*

3.5 Dimensionless form of the equations for incompressible flow

In this section the governing incompressible Navier-Stokes equations are given in their dimensionless form. First, consider the incompressible Navier-Stokes equations and the temperature equation, using primitive variables, in dimensional form and assume that the viscosity is constant:

$$u_{i,i} = 0$$

$$\rho\left(u_{i,t} + u_j u_{i,j}\right) = -p_{,i} + \mu s_{ij,\,j}, \quad i = 1,2,3 \tag{3.5.1}$$

$$\rho c_p\left(T_{,t} + u_i T_{,i}\right) = \mu \sum_{i,j=1}^{3} \frac{1}{2} s_{ij}^2 + \kappa T_{,ii}. \tag{3.5.2}$$

The magnitudes of the dimensional quantities that are used to express the incompressible Navier-Stokes equations can be given using four fundamental magnitudes: mass M, length L, time τ, and temperature Θ, see Table 3.1. According to Buckingham's Pi theorem, the dimensionless form of the equations is obtained using four reference values which form an independent (or recurrent) set. Our choice is the set $\{\rho_r, |u|_r, L, \Delta T\}$, where reference values are denoted by $_r$ and $\Delta T = T_w - T_\infty$ is the temperature difference between the wall and some other location in the fluid. The flow field is defined by the following (nine) dimensional quantities: ρ_r, u_r, L, ΔT, μ_r, κ_r, c_{p_r}, and their dimensions are summarized in Table 3.1.

Introduce the dimensionless variables, denoted by a star:

$$x_i^\star = \frac{x_i}{L} \quad t^\star = \frac{t}{L/|u|_r} \quad p^\star = \frac{p}{\rho_r |u|_r^2} \quad u_i^\star = \frac{u_i}{|u|_r} \tag{3.5.3}$$

$$\rho^\star = \frac{\rho}{\rho_r} \quad \mu^\star = \frac{\mu}{\mu_r} \qquad \kappa^\star = \frac{\kappa}{\kappa_r} \qquad c_p^\star = \frac{c_p}{c_{p_r}}. \tag{3.5.4}$$

The temperature will be made dimensionless with reference to the temperature difference $\Delta T = T_w - T_\infty$ between the wall and some other location in the fluid, thus

$$T^\star = \frac{T}{\Delta T}.$$

It is important to note that the system of equations contains four fundamental magnitudes and nine dimensional variables define the flow field, therefore, according to the Pi theorem of Buckingham five dimensionless Pi groups can be formed.

Because of the choice of the recurrent set, the reference values for the viscosity μ_r, specific heat at constant pressure c_{p_r} and the heat conductivity κ_r are still left to be defined.

Quantity	*Symbol*	*Dimension*	*made dimensionless*
mass	m	M	$\rho_r L^3$
length	x_i	L	L
time	t	τ	$L/\|u\|_r$
temperature	T	Θ	T_r
density	ρ	ML^{-1}	ρ_r
velocities	u_i	$L\tau^{-1}$	$\|u\|_r$
pressure	p	$ML^{-1}\tau^{-2}$	$\rho_r\|u\|_r^2$
viscosity	μ	$ML^{-1}\tau^{-1}$	$\rho_r\|u\|_r L/\mathrm{Re}$
specific heat at constant volume	c_v	$L^2\tau^{-2}\Theta^{-1}$	$\|u\|_r^2/(T_r\,\mathrm{Ec})$
specific heat at constant pressure	c_p	$L^2\tau^{-2}\Theta^{-1}$	$\|u\|_r^2/(T_r\,\mathrm{Ec})$
thermal conductivity	κ	$ML\tau^{-3}\Theta^{-1}$	$\rho_r\|u\|_r^3 L/(T_r\,\mathrm{Re}\,\mathrm{Pr}\,\mathrm{Ec})$

Table 3.1: Physical quantities, their symbol, dimension and non-dimensionalization using the recurrent set $\{\rho_r, |u|_r, L, \Delta T\}$,

Introducing the dimensionless quantities into the system (3.5.1-3.5.2), we obtain the dimensionless incompressible Navier-Stokes equations:

$$\rho^\star\left(\frac{\partial u_i^\star}{\partial t^\star}+u_j^\star\frac{\partial u_i^\star}{\partial x_j^\star}\right)=-\frac{\partial p^\star}{\partial x_i^\star}+\boxed{\frac{\mu_r}{\rho_r|u|_r L}}\,\mu^\star\left(\frac{\partial^2 u_i^\star}{\partial x_1^{\star 2}}+\frac{\partial^2 u_i^\star}{\partial x_2^{\star 2}}+\frac{\partial^2 u_i^\star}{\partial x_3^{\star 2}}\right) \tag{3.5.5}$$

$$\rho^\star c_p^\star\left(\frac{\partial T^\star}{\partial t^\star}+u_i^\star\frac{\partial T^\star}{\partial x_i^\star}\right)=\boxed{\frac{\mu_r|u|_r}{\rho_r c_{p_r} L\Delta T}}\,\mu^\star\sum_{i,j=1}^{3}\frac{1}{2}s_{ij}^{\star 2}+\boxed{\frac{\kappa_r}{\rho_r|u|_r L c_{p_r}}}\,\kappa^\star\frac{\partial^2 T^\star}{\partial x_i^{\star 2}}. \tag{3.5.6}$$

Define the following dimensionless numbers:

- Reynolds number: $\mathrm{Re}=\dfrac{\rho_r|u|_r L}{\mu_r}$
- Prandtl number: $\mathrm{Pr}=\dfrac{\mu_r c_{p_r}}{\kappa_r}$
- Eckert number: $\mathrm{Ec}=\dfrac{|u|_r^2}{c_{p_r}\Delta T}$.

The flow parameters can now be expressed in terms of the five dimensionless Pi groups: Re, Pr, Ec, the flow angles α and β, and the recurrent set of reference values $\{\rho_r, |u|_r, L, \Delta T\}$:

$$u_r=|u|_r(\cos\alpha\sin\phi,\sin\beta,\sin\alpha)^T,\quad \text{with}\quad \phi=\arcsin(\sin\beta/\cos\alpha)$$

and the reference values for the viscosity μ_r, specific heat at constant pressure c_{p_r} and the heat conductivity κ_r can be defined as:

$$\mu_r = \frac{\rho_r |u|_r L}{\mathrm{Re}}, \quad c_{p_r} = \frac{|u|_r^2}{\mathrm{Ec}\Delta T}, \tag{3.5.7}$$

and

$$\kappa_r = \frac{\mu_r c_{p_r}}{\mathrm{Pr}} = \frac{\rho_r |u|_r L}{\mathrm{Re}} \frac{|u|_r^2}{\mathrm{Ec}\Delta T} \frac{1}{\mathrm{Pr}} = \frac{\rho_r |u|_r^3 L}{\Delta T \mathrm{Re}\ \mathrm{Ec}\ \mathrm{Pr}}. \tag{3.5.8}$$

Using these dimensionless Pi groups, it is straightforward to see that the marked terms in the system (3.5.5-3.5.6) are dimensionless.

Note that when the temperature equation is not included, only three dimensionless Pi groups, $\mathrm{Re}, \alpha, \beta$ and the recurrent set of reference values $\rho_r, |u|_r, L$ will define the flow parameters in the momentum equations. By including the temperature equation, new flow parameters need to be specified, therefore, there is a need for dimensionless Pi groups that relate these flow parameters. In this way we find the Prandtl and Eckert numbers.

We shall give now a small overview of the interpretation of the dimensionless numbers.

- The Reynolds number can be interpreted as the ratio between the inertia forces and viscous forces in a fluid. The Reynolds number is influenced by fluid properties (viscosity and density), flow conditions (velocity) and geometry (reference length scale).
- The Prandtl number is the ratio between the viscous and thermal effects and is a function of fluid properties only. An important interpretation of the Prandtl number is that it represents the ratio of the relative thickness of the velocity and thermal boundary layers. When $\mathrm{Pr} = 1$, both boundary layers are of equal thickness, $\mathrm{Pr} > 1$ shows that momentum transfer is more rapid than heat transfer.
- The Eckert number is the kinetic energy of the flow relative to the enthalpy difference in e.g. a boundary layer.

3.6 Dimensionless form of the symmetrized Navier-Stokes equations

In this section we discuss the non-dimensionalization of the Navier-Stokes equations in it's symmetrized form. Let us introduce the recurrent set $\{\rho_r, |u|_r, L, T_r\}$ as in Section 3.5. The dimensionless variables, marked with a star, are related to the dimensional variables as in (3.5.3-3.5.4)

First, we give the non-dimensionalization of the entries of the Jacobian matrices:

$$\alpha_p^\star = \frac{\alpha_p}{T_r}, \quad \beta_T^\star = \rho_r |u|_r^2 \beta_T, \quad k^\star = k/|u|_r^2, \quad e_1^\star = e_1/|u|_r^2,$$

and using (3.5.7-3.5.8), we can write the entry r in the flux Jacobian matrices $\tilde{A}_i$ as

$$r = 2k + c_p T = 2|u|_r^2 k^\star + c_{p_r} T_r c_p^\star T^\star = |u|_r^2 \left(2k^\star + \frac{c_p^\star}{\mathrm{Ec}} T^\star \right), \tag{3.6.1}$$

and the entries $d_i, i = 1, 2, 3$ in the viscous flux Jacobian matrices $\tilde{K}_{ii}$ as

$$d_i = \frac{1}{3}\mu u_i^2 + \mu |u|^2 + \kappa T = \rho_r |u|_r^3 L \left(\frac{1}{3} \frac{\mu^\star}{\mathrm{Re}} u_i^{\star 2} + \frac{\mu^\star}{\mathrm{Re}} |u^\star|^2 + \frac{\kappa^\star}{\mathrm{Re\ Pr\ Ec}} T^\star \right), \tag{3.6.2}$$

where we used the same notations for the entries in $\tilde{A}_i$ and $\tilde{K}_{ij}$ as in Appendix B.1. Observe that introducing (3.6.1) and (3.6.2) into the Jacobian matrices, we obtain that the dimensionless Jacobian matrices in the Navier-Stokes equations, marked with a star, have the same functional form as their dimensional counterparts, with the dimensionless entries

$$r^\star = 2k^\star + \frac{c_p^\star}{\mathrm{Ec}} T^\star, \quad d_i^\star = \frac{1}{3}\mu^\star u_i^{\star 2} + \mu^\star |u^\star|^2 + \frac{\kappa^\star}{\mathrm{Pr\ Ec}} T^\star$$

Define the following diagonal matrix

$$D_r = \mathrm{diag}(|u|_r, |u|_r^2, |u|_r^2, |u|_r^2, |u|_r^3). \tag{3.6.3}$$

Then, the advective flux Jacobian matrices for entropy variables can be made dimensionless using the relations:

$$\tilde{A}_0 = \frac{\rho_r T_r}{|u|_r^4} D_r \tilde{A}_0^\star D_r \,, \quad \tilde{A}_i = \frac{\rho_r T_r}{|u|_r^3} D_r \tilde{A}_i^\star D_r \,, \ i = 1, 2, 3. \tag{3.6.4}$$

Using the derived relations, we obtain

$$\frac{\partial}{\partial t} = \frac{|u|_r}{L} \frac{\partial}{\partial t^\star}, \quad \frac{\partial}{\partial x_i} = \frac{1}{L} \frac{\partial}{\partial x_i^\star}, \tag{3.6.5}$$

and the set of entropy variables can be made dimensionless in the following way

$$V = \frac{1}{T_r T^\star} \begin{pmatrix} |u|_r^2 (\tilde{\mu}^\star - \frac{1}{2}|u^\star|^2) \\ |u|_r u_1^\star \\ |u|_r u_2^\star \\ |u|_r u_3^\star \\ -1 \end{pmatrix} = \frac{|u|_r^3}{T_r} D_r^{-1} V^\star. \tag{3.6.6}$$

For the dimensionless form of the chemical potential we use

$$s = c_v \ln T - s_0 = \frac{|u|_r^2}{\mathrm{Ec}\ T_r} c_p^\star \ln(T^\star T_r) - s_0 = \frac{|u|_r^2}{T_r}\frac{c_p^\star}{\mathrm{Ec}} \ln T^\star = \frac{|u|_r^2}{T_r} s^\star,$$

with the dimensionless entropy given by

$$s^\star = \frac{c_p^\star}{\mathrm{Ec}} \ln T^\star,$$

and where we have chosen the reference entropy value as

$$s_0 = \frac{|u|_r^2}{T_r}\frac{c_p^\star}{\mathrm{Ec}} \ln T_r.$$

Therefore, the chemical potential can be made dimensionless using

$$\tilde{\mu} = h - Ts = |u|_r^2 \left(h^\star - T^\star s^\star\right) = |u|_r^2\, \tilde{\mu}^\star$$

and together with (3.6.4) and

$$\frac{\partial V}{\partial t} = \frac{|u|_r^4}{LT_r} D_r^{-1} \frac{\partial V^\star}{\partial t^\star}, \quad \frac{\partial V}{\partial x_i} = \frac{|u|_r^3}{LT_r} D_r^{-1} \frac{\partial V^\star}{\partial x_i^\star},$$

we obtain the following relation between the dimensional and dimensionless forms of the inviscid part of the Navier-Stokes equations

$$\tilde{A}_0 \frac{\partial V}{\partial t} + \tilde{A}_i \frac{\partial V}{\partial x_i} = \frac{\rho_r}{L}\, D_r \left(\tilde{A}_0^\star \frac{\partial V^\star}{\partial t^\star} + \tilde{A}_i^\star \frac{\partial V^\star}{\partial x_i^\star} \right). \tag{3.6.7}$$

For the viscosity matrixes, we obtain using (3.6.2) the following relations

$$\tilde{K}_{ij} = \frac{\rho_r T_r L}{|u|_r^3} \frac{1}{\mathrm{Re}} D_r \tilde{K}_{ij}^\star D_r, \quad i,j = 1,2,3, \tag{3.6.8}$$

therefore,

$$\frac{\partial}{\partial x_i}\left(\tilde{K}_{ij} \frac{\partial V}{\partial x_j} \right) = \frac{\rho_r}{L}\, D_r \frac{1}{\mathrm{Re}} \frac{\partial}{\partial x_i^\star}\left(\tilde{K}_{ij}^\star \frac{\partial V^\star}{\partial x_j^\star} \right). \tag{3.6.9}$$

The relation between the dimensionless and dimension full forms of the symmetrized Navier-Stokes equations can then be expressed as

$$\tilde{A}_0 V_{,t} + \tilde{A}_i V_{,i} - (\tilde{K}_{ij} V_{,j})_{,i} = \frac{\rho_r}{L} D_r \left(\tilde{A}_0^\star V_{,t^\star}^\star + \tilde{A}_i^\star \frac{\partial V^\star}{\partial x_i^\star} - \frac{1}{\mathrm{Re}} \frac{\partial}{\partial x_i^\star}\left(\tilde{K}_{ij}^\star \frac{\partial V^\star}{\partial x_j^\star} \right) \right). \tag{3.6.10}$$

We can conclude that the dimensional and dimensionless Navier-Stokes equations have the same functional form. Therefore, the stars can be dropped from the nondimensional equations. Observe that the factor 1/Re can be extracted from the viscous coefficient matrices and (3.6.10) is obtained.

Chapter 4

Stabilization operators for the incompressible Navier-Stokes equations

One of the main problems in the construction of finite element methods for the incompressible Navier-Stokes equations is to find stable and efficient discretizations which can deal with convection dominated flows and the incompressibility constraint. The two main approaches which address these issues are finite elements which satisfy the Ladyzhenskaya-Babuška-Brezzi (LBB) or *inf-sup* condition or the use of stabilized finite element formulations. The first approach results in successful finite element methods, but in general it is difficult to design elements which satisfy the LBB-condition. Detailed surveys of this approach can be found in [4, 6, 17, 19, 48]. The second approach uses *stabilized* finite element formulations and provides more flexibility in the construction of finite element discretizations. This technique requires, however, the design of a stabilization operator or the enrichment of the finite element spaces with special functions, such as bubble functions. Stabilized methods for convection dominated flows were introduced by Brooks and Hughes [8] and since then a vast amount of literature has been published on this subject.

In this chapter we focus on the design and analysis of a class of stabilization operators suitable for space-time Galerkin least squares finite element discretizations of the incompressible limit of a symmetrized formulation of the Navier-Stokes equations given in [23]. This analysis we also described in [46]. In the incompressible limit these equations consist of the incompressible Navier-Stokes equations and the heat equation. The symmetrized formulation presented in [23] and discussed in Chapter 3 is suitable for both the compressible and incompressible Navier-Stokes equations and is an important step towards unified numerical discretizations suitable for a wide range of flow conditions. The symmetrization of the Navier-Stokes equations also provides

a good starting point for finite element discretizations, see [3, 51], and is discussed in detail in [18, 21, 43]. In an extensive series of papers Hughes and co-workers have used this approach to develop stabilized finite element methods both for the compressible and incompressible Navier-Stokes equations, see e.g. [23, 30, 32, 40, 51].

The motivation for the present study is the need for a better understanding of the mathematical properties of stabilization operators suitable for the incompressible limit of the symmetrized formulation for the Navier-Stokes equations. In [23] a stabilization operator is proposed as a natural extension of previous research on incompressible flows using primitive variables [13]. The extensions suggested are, however, mainly based on numerical experiments. A detailed analysis of the properties of the stabilization operator and the resulting discretization is missing, in particular for the stability in the incompressible limit. We will give therefore a consistent mathematical derivation of a class of stabilization operators suitable for the incompressible limit of the Navier-Stokes equations in the symmetrized formulation given in [23]. First, we will use dimensional analysis to determine a class of dimensionally consistent stabilization operators and we will show that this class also yields the stabilization operator suggested in [23].

The second topic of this chapter is to analyze the resulting class of stabilization operators such that we can ensure that the Galerkin least-squares finite element discretization results in a stable discretization technique which provides a unique solution, at least for the locally linearized problem. This analysis is an extension of the work in [13, 14] to the space-time formulation of the linearized incompressible Navier-Stokes equations in the symmetrized formulation derived in [23]. This proof provides additional information on the admissible stabilization operators, ensures positive definiteness of the stabilization operator and coercivity of the Galerkin least squares discretization.

This chapter is organized as follows. We start with a discussion of the symmetrized formulation of the incompressible Navier-Stokes equations in Section 4.1. The space-time Galerkin least-squares finite element method for the symmetrized incompressible Navier-Stokes equations is discussed in Section 4.2. Next, we derive in Section 4.3 a dimensionally consistent stabilization operator for the incompressible limit of the symmetrized formulation of the Navier-Stokes equations. Finally, in Section 4.4 we state conditions on a class of stabilization operators which ensures the coercivity of the Galerkin least-squares finite element discretization for the linearized case. We conclude with a summary of the main results and some remarks in Section 4.5.

4.1 The governing equations

Consider the incompressible Navier-Stokes equations combined with the heat equation in a time-dependent flow domain $\Omega(t)$, which in the remainder will be denoted

incompressible Navier-Stokes equations for brevity. Since the flow domain boundary is moving and deforming in time, we do not make a separation between the space and time variables and consider directly the space $\mathbb{R}^{d+1}$, where d is the number of space dimensions. Assume that $d = 3$. Let $\mathcal{E} \subset \mathbb{R}^4$ be an open, bounded space-time domain. A point $x \in \mathbb{R}^4$ has coordinates (x_0, x_1, x_2, x_3), with $x_0 = t$ representing time. The flow domain $\Omega(t) \subset \mathcal{E}$ at time t is defined as: $\Omega(t) = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid (t, x_1, x_2, x_3) \in \mathcal{E}\}$. The space-time domain boundary $\partial\mathcal{E}$ consists of the hypersurfaces $\Omega(t_0) = \{x \in \partial\mathcal{E} \mid x_0 = t_0\}$, $\Omega(t_{N+1}) = \{x \in \partial\mathcal{E} \mid x_0 = t_{N+1}\}$, and $\mathcal{Q} = \{x \in \partial\mathcal{E} \mid t_0 < x_0 < t_{N+1}\}$.

Let $Y : \mathcal{E} \mapsto \mathbb{R}^5$ denote the vector of primitive variables $(p, u_1, u_2, u_3, T)^T$ and $\Phi : \mathbb{R}^5 \mapsto \mathbb{R}^{5\times 4}$ the flux tensor, with the flux vector in the ℓth coordinate direction F_ℓ, $(\ell = 0, \ldots, 3)$ given by the columns of Φ, i.e.,

$$\Phi = \begin{pmatrix} \rho & \rho u_1 & \rho u_2 & \rho u_3 \\ \rho u_1 & \rho u_1^2 + p & \rho u_1 u_2 & \rho u_1 u_3 \\ \rho u_2 & \rho u_1 u_2 & \rho u_2^2 + p & \rho u_2 u_3 \\ \rho u_3 & \rho u_1 u_3 & \rho u_2 u_3 & \rho u_3^2 + p \\ \rho E & \rho u_1 E + p u_1 & \rho u_2 E + p u_2 & \rho u_3 E + p u_3 \end{pmatrix}, \tag{4.1.1}$$

where ρ denotes the density, u_i the velocity component in the ith Cartesian coordinate direction, p the pressure and E the total energy. Using these notations, the incompressible Navier-Stokes equations can be written in conservative form as

$$F_\ell(Y(x))_{,\ell} - (K_{ij}(Y(x))Y_{,j})_{,i} = 0, \quad x \in \mathcal{E}, \tag{4.1.2}$$

where $K_{ij} \in \mathbb{R}^{5\times 5}$ for $i, j = 1, 2, 3$ denote the viscous flux Jacobian matrices and the summation convention is used on repeated indices.

Both for the analysis and construction of the finite element method it is beneficial to use the entropy variables, which are defined in (3.2.2). The key benefit of the entropy variables is that they symmetrize the quasi-linear form of (4.1.2), which can be expressed as

$$\tilde{A}_0(V)V_{,t} + \tilde{A}_i(V)V_{,i} = (\tilde{K}_{ij}(V)V_{,j})_{,i} \tag{4.1.3}$$

where $\tilde{A}_\ell$, $(\ell = 0, \ldots, 3)$ denote the flux Jacobian matrices and $\tilde{K}_{ij}$, $(i, j = 1, 2, 3)$ the diffusivity coefficient matrices, given for completeness in the Appendix B.1. Recall from Chapter 3 that these matrices have the following properties: $\tilde{A}_0$ is symmetric positive-semidefinite, $\tilde{A}_i$, $(i = 1, 2, 3)$ are symmetric, and $K = [\tilde{K}_{ij}]$ is symmetric (i.e., $\tilde{K}_{ij} = \tilde{K}_{ji}^T$ for all $i, j = 1, 2, 3$) and positive-semidefinite. A detailed analysis of the benefits of different sets of independent variables for both the compressible and incompressible Navier-Stokes equations is given in [23].

4.2 Galerkin least-squares finite element formulation

Consider a partitioning of the time interval $I = (t_0, t_{N+1})$ using the time levels $t_0 < t_1 < \ldots < t_{N+1}$. We denote by $I_n = (t_n, t_{n+1})$ the nth time interval and define a space-time slab as $\mathcal{E}_n = \mathcal{E} \cap I_n$. Each space-time slab $\mathcal{E}_n$ is bounded by the hypersurfaces $\Omega(t_n)$, $\Omega(t_{n+1})$ and $\mathcal{Q}_n = \partial\mathcal{E}_n \setminus (\Omega(t_n) \cup \Omega(t_{n+1}))$. In each space-time slab $\mathcal{E}_n$ we define a partition $\mathcal{T}_h^n$ into $(n_e)_n$ non-overlapping elements $\mathcal{E}_n^e$. The space-time elements $\mathcal{E}_n^e$ are obtained by splitting the spatial domain $\Omega(t_n)$ into a set of non-overlapping elements Ω_n^e and connecting them with a mapping Φ_t^n to the elements $\Omega_{n+1}^e \subset \Omega(t_{n+1})$ at time t_{n+1}.

We now introduce some notation. With $(\cdot,\cdot)_{\mathcal{D}}$ we denote the L^2 inner product in the open domain $\mathcal{D} \subset \mathbb{R}^{d+1}$. In case of vector arguments, the L^2-inner product is defined as

$$\begin{aligned}(\cdot,\cdot)_{\mathcal{D}} &: \mathcal{D}^m \times \mathcal{D}^m \longrightarrow \mathbb{R} \\ (V,W)_{\mathcal{D}} &= \int_{\mathcal{D}} W^T V \,\mathrm{d}\,\mathcal{D}, \quad \text{for all } V, W \in \mathcal{D}^m\end{aligned}$$

and $\|\cdot\|_{0,\mathcal{D}}$ is the corresponding norm in the space $L^2(\mathcal{D})$. For a symmetric positive definite matrix $A \in \mathbb{R}^{m\times m}$, define the following inner product and norm respectively,

$$\begin{aligned}(\cdot,\cdot)_{A,\mathcal{D}} &: \mathcal{D}^m \times \mathcal{D}^m \longrightarrow \mathbb{R} \\ (V,W)_{A,\mathcal{D}} &= \int_{\mathcal{D}} W^T A V \,\mathrm{d}\,\mathcal{D}, \quad \text{for all } V, W \in \mathcal{D}^m\end{aligned}$$

and $\|V\|_{A,\mathcal{D}}^2 = (V,V)_{A,\mathcal{D}}$.

The trial function space in each space-time slab $\mathcal{E}_n$ is denoted by V_h^n and the test function space by W_h^n. Their elements are assumed to be $\mathcal{C}^0$ continuous within each space-time slab, but discontinuous across the interfaces of the space-time slabs, namely at times $t_1, t_2, \ldots, t_{N-1}$. The finite element spaces are now defined as:

$$\begin{aligned}V_h^n = \big\{ V \in H^1(\mathcal{E}_n)^5 : V\,|_{\mathcal{E}_n^e} \circ G_n^e \in \Big(\hat{\mathcal{P}}_1(0,1) \otimes \hat{\mathcal{P}}_k(\hat{\Omega})\Big)^5,\ \forall\, \mathcal{E}_n^e \in \mathcal{T}_h^n, \\ \int_{\mathcal{E}_n} V_1 \mathrm{d}\mathcal{E} = 0,\ q_1(V) = \bar{q}_1 \text{ on } \mathcal{Q}_n \big\}\end{aligned}$$

$$\begin{aligned}W_h^n = \big\{ W \in H^1(\mathcal{E}_n)^5 : W\,|_{\mathcal{E}_n^e} \circ G_n^e \in \Big(\hat{\mathcal{P}}_1(0,1) \otimes \hat{\mathcal{P}}_k(\hat{\Omega})\Big)^5,\ \forall\, \mathcal{E}_n^e \in \mathcal{T}_h^n, \\ \int_{\mathcal{E}_n} W_1 \mathrm{d}\mathcal{E} = 0,\ q_2(W) = \bar{q}_2 \text{ on } \mathcal{Q}_n \big\},\end{aligned}$$

where G_n^e denotes the mapping from the space-time reference element $(0,1)\times\hat{\Omega}$, with $\hat{\Omega}$ the reference element in R^3 (e.g. a tetrahedron, cube or prism) to the element in physical space $\mathcal{E}_n^e$, and $\hat{\mathcal{P}}_k$ represent kth-order polynomials. Further, V_1, W_1 denote the first component of $V, W \in \mathbb{R}^5$, respectively, $q_1 : \mathcal{E}^5 \to \mathbb{R}^4$ are the (nonlinear) boundary conditions for the components V_2, V_3, V_4, and V_5 of V, with a similar expression for $q_2 : \mathcal{E}^5 \to \mathbb{R}^4$, and $\bar{q}_1, \bar{q}_2 \in \mathbb{R}^4$ are the prescribed boundary conditions. Note, not necessarily all components of V will have imposed boundary conditions, this depends on the type of boundary condition. If $q_1(V) = (V_2, V_3, V_4, V_5)^T = (0,0,0,0)^T$ then we denote V_h^n as V_{0h}^n, and analogously, we use W_{0h}^n when $q_2(W) = (W_2, W_3, W_4, W_5)^T = (0,0,0,0)^T$. When the finite element spaces are defined on the whole space time domain then the superscript n is omitted.

Let us recall the Galerkin least-squares variational formulation for the Navier-Stokes equations in terms of the entropy variables:

Within each space-time slab $\mathcal{E}_n$, find a $V \in V_h^n$, such that for all $W \in W_h^n$ the following relation is satisfied:

$$\int_{\mathcal{E}_n} \Big(-W_{,\ell} \cdot F_\ell(V) + W_{,i} \cdot (\tilde{K}_{ij} V_{,j})\Big)\, \mathrm{d}\mathcal{E} + B_{ls}(V,W) + B_{bc}(V,W)$$

$$+ \int_{\Omega(t_{n+1})} W(t_{n+1}^-) \cdot F_0(V(t_{n+1}^-))\mathrm{d}\Omega - \int_{\Omega(t_n)} W(t_n^+) \cdot F_0(V(t_n^-))\mathrm{d}\Omega = 0, \quad (4.2.1)$$

for $(\ell = 0, \ldots, 3)$, $(i, j = 1, 2, 3)$, where the second term is the least-squares stabilization operator, defined as

$$B_{ls}(V,W) = \sum_{e=1}^{(n_e)_n} \int_{\mathcal{E}_n^e} (\mathcal{L}_V V) \cdot \tilde{\tau} (\mathcal{L}_V W) \mathrm{d}\mathcal{E}, \quad (4.2.2)$$

with $\mathcal{L}_V$ the symmetrized Navier-Stokes operator

$$\mathcal{L}_V = \tilde{A}_\ell(V) \frac{\partial}{\partial x_\ell} - \frac{\partial}{\partial x_i}\left(\tilde{K}_{ij}(V) \frac{\partial}{\partial x_j}\right) \quad \text{for } \ell = 0, \ldots, 3,\ i, j = 1, 2, 3,$$

and $\tilde{\tau}$ the stabilization operator matrix. We use also the notation $\mathcal{L}_V^{inv}$ to denote the inviscid counterpart of $\mathcal{L}_V$ (hence $\tilde{K}_{ij} = 0$ for all $i, j = 1, 2, 3$). The boundary operator in (4.2.1) is obtained after integration by parts of the weak form of (4.1.2) and is defined as

$$B_{bc}(V,W) = \int_{\mathcal{Q}_n} n \cdot (W^T \Phi(V)) - W \cdot (\tilde{K}_{ij} V_{,j}) \bar{n}_i \, \mathrm{d}\mathcal{Q}.$$

Here n is the unit outward space-time normal vector at the boundary $\mathcal{Q}_n$ and $\bar{n}$ its the spatial component. Similarly, the last two integrals in (4.2.1) are obtained by combining the boundary integrals at $\Omega(t_n)$ and $\Omega(t_{n+1})$ with the so-called jump term

$$B_{jump}(V,W) = \int_{\Omega(t_n)} W(t_n^+) \cdot \Big(F_0(V(t_n^+)) - F_0(V(t_n^-))\Big)\mathrm{d}\Omega,$$

which ensures weak continuity between different space-time slabs. We do not give a detailed analysis of the boundary operator since it is beyond the scope of this chapter.

The stabilization operator is added to the weak formulation of the incompressible Navier-Stokes equations to ensure that the inf-sup or LBB condition is satisfied, see e.g [48], which is essential to obtain a unique solution. The stabilization operator technique allows the use of equal order polynomial basis functions for all quantities and provides more flexibility in the construction of finite element spaces. In the least-squares operator, the choice of the $\tilde{\tau}$ matrix is crucial, and is examined in detail in this chapter. This operator greatly influences the stability of the numerical scheme.

4.3 Explicit construction of stabilization operators

In this section a class of dimensionally consistent stabilization operators will be derived, which also includes, as a special case, the stabilization operator given in [23]. We recall that the standard definition of the stabilization matrix requires $\tilde{\tau}$ to be symmetric, positive definite, have dimensions of time, and scale linearly with the element size (see [32]). Due to the fully coupled structure of the system (4.1.3), it is, however, difficult to define a stabilization matrix directly in terms of entropy variables. Therefore, the choice of variables in which the system is expressed and used to define the stabilization operator is important.

Our starting point is a dimensional analysis of the stabilization matrix τ_Y related to the primitive variables $Y = (p, u_1, u_2, u_3, T)^T$. This stabilization operator is related to $\tilde{\tau}$ through the transformation

$$\tilde{\tau} = V_{,Y}\tau_Y, \tag{4.3.1}$$

where $V_{,Y}$ is given in the Appendix B.3. For primitive variables, the stabilization matrix τ_Y can be chosen to be of diagonal form, which has been successfully applied in previous research on incompressible flows, see [13]. Introducing this diagonal τ_Y into (4.3.1) results, however in a non-symmetric stabilization matrix for entropy variables, which does not satisfy our requirements on the symmetric formulation. It is possible to obtain a diagonal matrix for entropy variables, such as $\tilde{\tau} = \text{diag}\,(V_{,Y}\tau_Y)$, with the diagonal τ_Y which is therefore symmetric and also positive definite. This simple choice is however not good, as shown in the numerical examples in [23]. In order to find a stabilization operator in terms of the primitive variables, we first transform the Galerkin least-squares discretization defined in the entropy variables (4.2.1) to the primitive variables. All terms in the variational formulation (4.2.1) remain essentially unchanged with the least-squares contribution written in terms of the differential operator $\mathcal{L}_Y$.

In order to derive a dimensionally consistent stabilization operator we need to make the concept "a scales like b" mathematically more precise. For this we first introduce

some notation. Consider the set S of all flow variables (such as velocity, temperature, pressure, etc.), and its power set, denoted by $P(S)$.

Definition 4.3.1 *Given the set S, $P(S)$, a set $\mathcal{V} = \{\nu_1, \nu_2, \ldots, \nu_n\} \in P(S)$ and a set of functionals $F = \{f \mid f : P(S) \to \mathbb{R}\}$. Introduce the reference values for the elements of $\mathcal{V}$, $r(\mathcal{V}) = \{r(\nu_1), r(\nu_2), \ldots, r(\nu_n)\}$. Furthermore, define the following mapping*

$$\Lambda : (\nu_1, \nu_2, \ldots, \nu_n) \mapsto (\lambda_1\nu_1, \lambda_2\nu_2, \ldots, \lambda_n\nu_n) \text{ for any } \lambda_i > 0,$$

such that there are $m_i \in \mathbb{Z}$, $i = 1, \ldots, n$ with

$$f(\Lambda(\mathcal{V})) = \lambda_1^{m_1}\lambda_2^{m_2}\cdots\lambda_n^{m_n} f(\mathcal{V}), \quad \forall\, f \in F.$$

Then, an equivalence relation $\sim_{\mathcal{V}}$ over the set of functionals F is defined as:

$$f \sim_{\mathcal{V}} g \quad \Longleftrightarrow \quad \text{whenever } \begin{cases} f(\Lambda(\mathcal{V})) & = \lambda_1^{m_1}\lambda_2^{m_2}\cdots\lambda_n^{m_n} f(\mathcal{V}) \\ g(\Lambda(\mathcal{V})) & = \lambda_1^{k_1}\lambda_2^{k_2}\cdots\lambda_n^{k_n} g(\mathcal{V}), \end{cases}$$
$$\text{then } m_i = k_i,\ \forall i = 1, \ldots n.$$

We say that f is dimensionally equivalent (or has the same dimension) to g with respect to the set of flow variables $\mathcal{V}$.

Definition 4.3.2 *An equivalence class is a subset of F of the form $\{g : g \sim_{\mathcal{V}} f\}$, where f is an element in F. This equivalence class is represented as*

$$[f]_{\mathcal{V}} = (r(\nu_1))^{m_1}(r(\nu_2))^{m_2}\cdots(r(\nu_n))^{m_n}$$

and addition is defined between two elements of F if and only if they belong to the same equivalence class.

Definition 4.3.3 *The set $\mathcal{V}$ is called canonical if for $\nu_i \in \mathcal{V}$, $\nexists\, l_j \in \mathbb{Z}$ such that we can express $r(\nu_i) = (r(\nu_{i_1}))^{l_1}\cdots(r(\nu_{i_k}))^{l_k}$ with $\nu_{i_l} \in \mathcal{V} \setminus \{\nu_i\}$, for any $i = 1, \ldots, n$. If each element of F forms an equivalence class by itself, then F is called a set of independent variables with respect to the canonical set $\mathcal{V}$.*

Note here that each equivalence class in F is closed to linear combinations, i.e.,

$$f \sim_{\mathcal{V}} g \Longrightarrow c_1 f + c_2 g \sim_{\mathcal{V}} f,\ \forall c_1, c_2 \in \mathbb{R} \setminus \{0\}.$$

The inverse of an element and multiplication between two elements of F have the following properties, respectively

$$\left[\frac{1}{f}\right]_{\mathcal{V}} = ([f]_{\mathcal{V}})^{-1}, \quad [f_1 \cdot f_2]_{\mathcal{V}} = [f_1]_{\mathcal{V}} \cdot [f_2]_{\mathcal{V}}.$$

Definition 4.3.4 *Let $A = (a_{ij}) \in \mathbb{R}^{n\times m}$, $B = (b_{ij}) \in \mathbb{R}^{n\times m}$. Then, we define $A \sim_{\mathcal{V}} B$ if*

$$a_{ij} \sim_{\mathcal{V}} b_{ij} \quad \textit{for all} \quad i = 1, \ldots, n,\ j = 1, \ldots, m. \tag{4.3.2}$$

Our goal is now to derive a dimensionally consistent stabilization operator for the Navier-Stokes equations in primitive variables. We first transform the equations $\mathcal{L}_V V = 0$ to the primitive variables $Y = (p, u_1, u_2, u_3, T)^T$:

$$\mathcal{L}_Y Y = A_0(Y)Y_{,0} + A_i(Y)Y_{,i} - (K_{ij}(Y)Y_{,j})_{,i} = 0 \tag{4.3.3}$$

with the coefficient matrices $A_\ell(Y)$ and $K_{ij}(Y)$ given in the Appendix B.2. For the dimensional analysis we first need some preliminaries to specify the dimension of the various terms in the Navier-Stokes equations. Let us denote length, time, mass, and temperature by l, t, m, T, respectively, and introduce the reference values for length L, time τ, mass M and temperature Θ. Consider the canonical set $\{l, t, m, T\}$, then we obtain the following equivalence classes

$$[u]_{\{l,t,m,T\}} = \frac{L}{\tau}, \quad [\rho]_{\{l,t,m,T\}} = \frac{M}{L^3}, \quad [l]_{\{l,t,m,T\}} = L, \quad [T]_{\{l,t,m,T\}} = \Theta,$$

which imply that the set $\mathcal{V} = \{u, T, \rho, l\}$ is a set of independent variables with respect to the canonical set $\{l, t, m, T\}$. Moreover, $\mathcal{V}$ is a canonical set and we consider the set of reference values $r(\mathcal{V}) = \{U, \Theta, R, L\}$, with U and R the reference values for velocity and density, respectively. Using Definitions 4.3.1 and 4.3.4, we give a dimensional analysis of the derivatives of the primitive variables and the corresponding flux Jacobian matrices in (4.3.3) with respect to $\mathcal{V}$. For the velocity components we have $[u_i]_{\mathcal{V}} = U$, for all $i = 1, 2, 3$. Then,

$$[u_{,0}]_{\mathcal{V}} = \frac{U^2}{L},\ [u_{,i}]_{\mathcal{V}} = \frac{U}{L},\ [T_{,0}]_{\mathcal{V}} = \frac{U\Theta}{L},\ [T_{,i}]_{\mathcal{V}} = \frac{\Theta}{L},\ \text{for all } i = 1, 2, 3.$$

Since for an incompressible flow the pressure is not a thermodynamic but a mechanical variable, we obtain $[p]_{\mathcal{V}} = RU^2$. Then, $[p_{,i}]_{\mathcal{V}} = RU^2/L$ for all $i = 1, 2, 3$, and $[p_{,0}]_{\mathcal{V}} = RU^3/L$. Moreover, for the entries of the Jacobian matrices we have $[c_p]_{\mathcal{V}} = U^2/\Theta$, $[h-k]_{\mathcal{V}} = U^2$ with h the specific enthalpy and $k = \frac{1}{2}|u|^2$, $[\mu]_{\mathcal{V}} = RUL$ for the viscosity coefficient μ, and $[\kappa]_{\mathcal{V}} = RU^3L/\Theta$ for the thermal conductivity κ. Hence, using the definition of the various vectors and matrices given in the Appendix, the following dimensional equivalence is valid

$$[Y_{,0}]_{\mathcal{V}} = \frac{U}{L}\begin{pmatrix} RU^2 \\ U \\ U \\ U \\ \Theta \end{pmatrix},\ [Y_{,i}]_{\mathcal{V}} = \frac{U}{L}\begin{pmatrix} RU \\ 1 \\ 1 \\ 1 \\ \Theta/U \end{pmatrix},$$

$$[A_0(Y)]_{\mathcal{V}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & R & 0 & 0 & 0 \\ 0 & 0 & R & 0 & 0 \\ 0 & 0 & 0 & R & 0 \\ 0 & RU & RU & RU & RU^2/\Theta \end{pmatrix}, \tag{4.3.4}$$

$$[A_i(Y)]_{\mathcal{V}} = \begin{pmatrix} 0 & \delta_{1i}R & \delta_{2i}R & \delta_{3i}R & 0 \\ \delta_{1i} & RU & \delta_{2i}RU & \delta_{3i}RU & 0 \\ \delta_{2i} & \delta_{1i}RU & RU & \delta_{3i}RU & 0 \\ \delta_{3i} & \delta_{1i}RU & \delta_{2i}RU & RU & 0 \\ U & RU^2 & RU^2 & RU^2 & RU^3/\Theta \end{pmatrix}, \tag{4.3.5}$$

where δ_{ij} is the Kronecker delta symbol, and for the viscosity coefficient matrices we obtain for $i = j$:

$$[K_{ii}(Y)]_{\mathcal{V}} = L \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & RU & 0 & 0 & 0 \\ 0 & 0 & RU & 0 & 0 \\ 0 & 0 & 0 & RU & 0 \\ 0 & RU^2 & RU^2 & RU^2 & RU^3/\Theta \end{pmatrix}, \tag{4.3.6}$$

and for $i \neq j$ we have

$$[K_{ij}(Y)]_{\mathcal{V}} = L \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{12}RU & a_{13}RU & 0 \\ 0 & a_{21}RU & 0 & a_{23}RU & 0 \\ 0 & a_{31}RU & a_{32}RU & 0 & 0 \\ 0 & b_{11}RU^2 & b_{22}RU^2 & b_{33}RU^2 & 0 \end{pmatrix}, \tag{4.3.7}$$

with $i, j = 1, 2, 3$. The coefficients in the matrix $K_{ij}(Y)$ are defined as

$$a_{kl} = \begin{cases} 1 \text{ if } (k = i \wedge l = j) \vee (k = j \wedge l = i) \\ 0 \text{ otherwise} \end{cases} \qquad b_{kk} = \begin{cases} 1 \text{ if } k = i \vee k = j \\ 0 \text{ otherwise,} \end{cases}$$

for $k, l = 1, 2, 3$. By dimensional consistency, we can add the various contributions and obtain the following dimensional equivalence for the Navier-Stokes equations

$$[\mathcal{L}_Y Y]_{\mathcal{V}} = [A_0(Y)Y_{,0} + A_i(Y)Y_{,i} - (K_{ij}(Y)Y_{,j})_{,i}]_{\mathcal{V}} = \frac{RU}{L}(1, U, U, U, U^2)^T. \tag{4.3.8}$$

Our aim is now to construct a stabilized finite element method, which satisfies the following requirements:

(a) The method admits discrete solutions Y^h with the same dimension as the solution Y of (4.3.3).

(b) Elementwise the least-squares stabilization operator is dimensionally equivalent with the Galerkin operator.

Similar assumptions are made in [58], where a scaling analysis is performed to determine the appropriate low Mach number behavior of the stabilization matrix. These requirements imply

$$\left[(\mathcal{L}_Y^T W)\cdot\tau_Y(\mathcal{L}_Y Y)\right]_{\mathcal{V}} = \left[W\cdot(\mathcal{L}_Y Y)\right]_{\mathcal{V}}, \quad \forall W\in W_h^n,$$

which is equivalent with

$$\frac{1}{L}\left(\left(U[A_0^T(Y)]_{\mathcal{V}} + [A_i^T(Y)]_{\mathcal{V}} - \frac{1}{L}[K_{ij}^T(Y)]_{\mathcal{V}}\right)[W]_{\mathcal{V}}\right)^T [\tau_Y(\mathcal{L}_Y Y)]_{\mathcal{V}} = [W^T]_{\mathcal{V}}\,[(\mathcal{L}_Y Y)]_{\mathcal{V}}\,, \tag{4.3.9}$$

where the scales U and $1/L$ originate from the derivatives of the test function. Note that using $[\mu]_{\mathcal{V}} = RUL$ and $[\kappa]_{\mathcal{V}} = RU^3L/\Theta$, it follows that $A_i(Y)\sim_{\mathcal{V}}(1/L)K_{ij}(Y)$ for all $i,j=1,2,3$, hence the addition and subtraction are well defined in (4.3.9). Since the test functions are arbitrary, using (4.3.4)-(4.3.7), equation (4.3.9) is equivalent to

$$\frac{1}{L}\left(U[A_0(Y)]_{\mathcal{V}} + [A_i(Y)]_{\mathcal{V}} - \frac{1}{L}[K_{ij}(Y)]_{\mathcal{V}}\right)[\tau_Y]_{\mathcal{V}}\,[(\mathcal{L}_Y Y)]_{\mathcal{V}} = [(\mathcal{L}_Y Y)]_{\mathcal{V}}\,. \tag{4.3.10}$$

Therefore, requirements (a) and (b) provide an additional condition on the components of the stabilization matrix τ_Y, i.e., (4.3.10) is equivalent to:

$$\frac{1}{L}\begin{pmatrix} 0 & R & R & R & 0\\ 1 & RU & RU & RU & 0\\ 1 & RU & RU & RU & 0\\ 1 & RU & RU & RU & 0\\ U & RU^2 & RU^2 & RU^2 & RU^3/\Theta \end{pmatrix}[\tau_Y]_{\mathcal{V}}\,[(\mathcal{L}_Y Y)]_{\mathcal{V}} = [(\mathcal{L}_Y Y)]_{\mathcal{V}}\,. \tag{4.3.11}$$

The scaling relation (4.3.11) shows that a suitable stabilization matrix τ_Y is dimensionally equivalent to the inverse of a non-singular matrix represented by the expanded matrix in (4.3.11), i.e.,

$$[\tau_Y]_{\mathcal{V}} = L\begin{pmatrix} U & 1 & 1 & 1 & 0\\ 1/R & 1/(RU) & 1/(RU) & 1/(RU) & 0\\ 1/R & 1/(RU) & 1/(RU) & 1/(RU) & 0\\ 1/R & 1/(RU) & 1/(RU) & 1/(RU) & 0\\ \Theta/(RU) & \Theta/(RU^2) & \Theta/(RU^2) & \Theta/(RU^2) & \Theta/(RU^3) \end{pmatrix},$$

which defines the structure of a dimensionally consistent stabilization matrix τ_Y. Summarizing, we write the general form of the stabilization matrix in primitive variables,

indicating the dimension of the entries of the matrix, as

$$\tau_Y = L \begin{pmatrix} c_{11}U & c_{12} & c_{13} & c_{14} & 0 \\ \frac{c_{21}}{R} & \frac{c_{22}}{RU} & \frac{c_{23}}{RU} & \frac{c_{24}}{RU} & 0 \\ \frac{c_{31}}{R} & \frac{c_{32}}{RU} & \frac{c_{33}}{RU} & \frac{c_{34}}{RU} & 0 \\ \frac{c_{41}}{R} & \frac{c_{42}}{RU} & \frac{c_{43}}{RU} & \frac{c_{44}}{RU} & 0 \\ \frac{c_{51}\Theta}{RU} & \frac{c_{52}\Theta}{RU^2} & \frac{c_{53}\Theta}{RU^2} & \frac{c_{54}\Theta}{RU^2} & \frac{c_{55}\Theta}{RU^3} \end{pmatrix} \tag{4.3.12}$$

where c_{ij} are functions of dimensionless flow variables and R, U, Θ, L are the reference density, velocity, temperature and length, respectively. Note here that the zeroes in the last column of (4.3.12) result from the inverse of a non-singular matrix represented by the expanded matrix in (4.3.11). In this form the matrix still has 21 unknowns which need to be specified.

In the remaining part of this section we will specify the coefficients c_{ij} in the stabilization matrix τ_Y (4.3.12) using the symmetry property of $\tilde{\tau}$.

Theorem 4.3.1 *Given the stabilization matrix τ_Y for primitive variables in the dimensionally consistent form (4.3.12), then a unique class of stabilization matrices τ_Y and $\tilde{\tau}$, related by transformation (4.3.1), can be defined as:*

$$\tau_Y = \begin{pmatrix} \tau_c & (\omega+1)\rho u_1 \tau_m & (\omega+1)\rho u_2 \tau_m & (\omega+1)\rho u_3 \tau_m & 0 \\ \omega u_1 \tau_m & \tau_m & 0 & 0 & 0 \\ \omega u_2 \tau_m & 0 & \tau_m & 0 & 0 \\ \omega u_3 \tau_m & 0 & 0 & \tau_m & 0 \\ -(h-k)\tau_e & -u_1\tau_e & -u_2\tau_e & -u_3\tau_e & \tau_e \end{pmatrix} \tag{4.3.13}$$

and

$$\tilde{\tau} = \left(\begin{array}{c|c|c|c|c} \tilde{\tau}_{11} & \frac{\omega u_1}{T}\tau_m + \frac{(h-k)}{T}\frac{u_1}{T}\tau_e & \frac{\omega u_2}{T}\tau_m + \frac{(h-k)}{T}\frac{u_2}{T}\tau_e & \frac{\omega u_3}{T}\tau_m + \frac{(h-k)}{T}\frac{u_3}{T}\tau_e & -\frac{(h-k)}{T^2}\tau_e \\ \dots & \frac{1}{T}\tau_m + \frac{u_1^2}{T^2}\tau_e & \frac{u_1 u_2}{T^2}\tau_e & \frac{u_1 u_3}{T^2}\tau_e & -\frac{u_1}{T^2}\tau_e \\ \dots & \dots & \frac{1}{T}\tau_m + \frac{u_2^2}{T^2}\tau_e & \frac{u_2 u_3}{T^2}\tau_e & -\frac{u_2}{T^2}\tau_e \\ \dots & symm & \dots & \frac{1}{T}\tau_m + \frac{u_3^2}{T^2}\tau_e & -\frac{u_3}{T^2}\tau_e \\ \dots & \dots & \dots & \dots & \frac{1}{T^2}\tau_e \end{array}\right) \tag{4.3.14}$$

where h is the specific enthalpy, $k = |u|^2/2$, $\omega \in \mathbb{R}$ is a parameter, $\tau_c, \tau_m, \tau_e \in \mathbb{R}^+$ and $\tilde{\tau}_{11} = \frac{1}{\rho T}\tau_c - \frac{\omega|u|^2}{T}\tau_m + \left(\frac{h-k}{T}\right)^2 \tau_e$.

Proof:
Using the symmetry of $\tilde{\tau}$ and transformation (4.3.1), we obtain the following relations for the coefficients c_{ij} in (4.3.12):

$$c_{23} = c_{32},\ c_{24} = c_{42},\ c_{34} = c_{43}, \tag{4.3.15}$$

$$\begin{aligned} &c_{12} = \frac{\rho(c_{22}u_1 + c_{23}u_2 + c_{24}u_3 + c_{21}U)}{RU},\ c_{13} = \frac{\rho(c_{23}u_1 + c_{33}u_2 + c_{34}u_3 + c_{31}U)}{RU} \\ &c_{14} = \frac{\rho(c_{24}u_1 + c_{34}u_2 + c_{44}u_3 + c_{41}U)}{RU},\ c_{51} = -\frac{(h-k)c_{55}}{U^2}, \\ &c_{52} = -\frac{u_1 c_{55}}{U},\ c_{53} = -\frac{u_2 c_{55}}{U},\ c_{54} = -\frac{u_3 c_{55}}{U}. \end{aligned} \tag{4.3.16}$$

Since $[k]_{\mathcal{V}} = [h]_{\mathcal{V}} = U^2$, all operations in the above relations are valid and only for convenience we leave the elements of $V_{,Y}$ in their dimension full form. Consider now the middle 3×3 block in (4.3.12), which corresponds to the three momentum equations. Relation (4.3.15) implies that this block is symmetric. Moreover, this block must be rotational invariant, which together with it's symmetry, implies that it is a constant times the identity matrix. The coefficients must therefore satisfy the relation $c_{22} = c_{33} = c_{44} = c$ and $c_{23} = c_{32} = c_{24} = c_{42} = c_{34} = c_{43} = 0$. For simplicity we introduce the following notation for the diagonal entries in τ_Y,

$$\tau_c := c_{11}UL, \quad \tau_m := \frac{cL}{RU}, \quad \tau_e := \frac{c_{55}L\Theta}{RU^3}. \tag{4.3.17}$$

Then, the relations in (4.3.16) can be written as

$$\begin{aligned} c_{51} &= -\frac{RU(h-k)}{L\Theta}\tau_e, \quad c_{52} = -\frac{Ru_1U^2}{L\Theta}\tau_e, \\ c_{53} &= -\frac{Ru_2U^2}{L\Theta}\tau_e, \qquad c_{54} = -\frac{Ru_3U^2}{L\Theta}\tau_e \end{aligned}$$

and obtain

$$c_{12} = \frac{\rho u_1}{L}\tau_m + \frac{\rho}{R}c_{21},\ c_{13} = \frac{\rho u_2}{L}\tau_m + \frac{\rho}{R}c_{31},\ c_{14} = \frac{\rho u_3}{L}\tau_m + \frac{\rho}{R}c_{41}. \tag{4.3.18}$$

Since $\tau_m \neq 0$, it follows from (4.3.18) that there are at least three additional non-vanishing entries in the matrix τ_Y. The vector a, composed of $a = (c_{12}, c_{13}, c_{14})^T$, is multiplied in the least-squares operator with the momentum equations, which are rotational invariant, and this implies that a must also be rotational invariant. We can therefore write $a = \alpha u$, with the scalar $[\alpha]_{\mathcal{V}} = 1/U$, and using (4.3.18) obtain

$$\frac{\rho}{R}b = \left(\alpha - \frac{\rho}{L}\tau_m\right)u,$$

with $b = (c_{21}, c_{31}, c_{41})^T$. Since $[\alpha]_{\mathcal{V}} = [\frac{\rho}{L}\tau_m]_{\mathcal{V}}$, we can choose $\alpha = (\omega+1)\frac{\rho\tau_m}{L}$ with $\omega \in \mathbb{R}$. Therefore,

$$a = (\omega+1)\frac{\rho\tau_m}{L}u, \quad \frac{b}{R} = \omega\frac{\tau_m}{L}u.$$

Inserting all relations on the constants $c_{ij},\ i,j=1,\ldots,5$ into (4.3.12), we obtain the general form of the stabilization matrix (4.3.13).

The stabilization operator $\tilde{\tau}$ can be obtained directly using (4.3.1) with τ_Y given by (4.3.13). □

In order to ensure the positive definiteness of $\tilde{\tau}$, we need to set conditions on the parameters ω, τ_c, τ_m and τ_e.

Theorem 4.3.2 *The matrix $\tilde{\tau}$ in (4.3.14) is positive definite if and only if the following conditions on the stabilization parameters τ_c, τ_m, τ_e and ω are satisfied*

$$\begin{cases} \tau_m > 0 \\ \tau_e > 0 \\ \tau_c > \rho|u|^2\omega(\omega+1)\tau_m \end{cases} \tag{4.3.19}$$

Proof:
Assume that $\tilde{\tau}$ is positive definite. Then, all eigenvalues of $\tilde{\tau}$ are real and positive. Since τ_m is an eigenvalue of $\tilde{\tau}$, it follows that $\tau_m > 0$. From the positive definiteness of $\tilde{\tau}$ it follows that all it's principal submatrices are also positive definite, therefore $\tau_e > 0$. Moreover, it follows that all minor principals of $\tilde{\tau}$ are positive definite. Since $\tau_e > 0$ and $\tau_m > 0$, this implies the following five inequalities:

$$\tau_c > \rho\left(\omega|u|^2\tau_m - \frac{(h-k)^2}{T}\tau_e\right), \tag{4.3.20}$$

$$\tau_c > \frac{\rho\tau_m\left(\omega u_1^2(\omega T\tau_m + 2h\tau_e) + \omega T\tau_m|u|^2 - (h-k)^2\tau_e\right)}{T\tau_m + u_1^2\tau_e}, \tag{4.3.21}$$

$$\tau_c > \frac{\rho\tau_m\left(\omega(u_1^2+u_2^2)(\omega T\tau_m + 2h\tau_e) + \omega T\tau_m|u|^2 - (h-k)^2\tau_e\right)}{T\tau_m + (u_1^2+u_2^2)\tau_e}, \tag{4.3.22}$$

$$\tau_c > \frac{\rho\tau_m\left(\omega|u|^2(\omega T\tau_m + 2h\tau_e) + \omega T\tau_m|u|^2 - (h-k)^2\tau_e\right)}{T\tau_m + |u|^2\tau_e}, \tag{4.3.23}$$

$$\tau_c > \rho\omega|u|^2(\omega+1)\tau_m. \tag{4.3.24}$$

Consider now the right-hand side of (4.3.20-4.3.24) written in the following functional form

$$F(X,\tau_e) = \frac{\rho\tau_m\left(\omega X(\omega T\tau_m + 2h\tau_e) + \omega T\tau_m|u|^2 - (h-k)^2\tau_e\right)}{T\tau_m + X\tau_e}$$

with $0 \le X \le |u|^2$ and $\tau_e \ge 0$. We can consider now the following cases:

Case 1. Assume that $\tau_e = 0$. Then

$$F(X,0) = \rho\omega\tau_m(\omega X + |u|^2)$$

Since $\frac{\partial F(X,0)}{\partial X} = \rho\omega^2\tau_m > 0$, it follows that $F(X,0)$ is a monotone increasing function of X. Hence, $F(X,0) \leq F(|u|^2,0) = \rho\omega|u|^2(\omega+1)\tau_m$ for all $0 \leq X \leq |u|^2$. Consequently, when $\tau_e = 0$ combining inequalities (4.3.20-4.3.24) leads to (4.3.24).

Case 2. Assume that $\tau_e > 0$. Then, for any fixed $0 \leq X^* \leq |u|^2$ it follows that $F(X^*, \tau_e)$ is a monotone decreasing function of τ_e, since

$$\frac{\partial F(X^*,\tau_e)}{\partial \tau_e} = -\frac{\rho\tau_m^2 T(h-k-\omega X)^2}{(T\tau_m + X\tau_e)^2} < 0.$$

Therefore, $F(X^*,0) > F(X^*,\tau_e)$ for all $\tau_e > 0$ and for all $0 \leq X^* \leq |u|^2$.

Summarizing, from the two cases discussed above, it follows that

$$F(X^*,\tau_e) < F(X^*,0) \leq F(|u|^2,0), \quad \forall\ X^* \in [0,|u|^2],\ \text{and}\ \forall \tau_e > 0,$$

which leads to (4.3.19).

The proof of the reverse statement of this lemma is straightforward, since (4.3.19) implies that the inequalities (4.3.20-4.3.24) are valid, i.e., all minor principals of $\tilde{\tau}$ are positive definite, which is a sufficient condition for positive definiteness of $\tilde{\tau}$. □

Remark 4.3.1 *We can give further information on the lower right* 4×4 *principal submatrix of* $\tilde{\tau}$*, denoted by* $\hat{\tau}$*. Note that* $\hat{\tau}$ *does not depend on* τ_c *and* ω*. The positive definiteness of* $\tilde{\tau}$ *implies that* $\hat{\tau}$ *is also positive definite and it's eigenvalues are*

$$\lambda_1 = \lambda_2 = \frac{\tau_m}{T} > 0, \tag{4.3.25}$$

and

$$\lambda_{3,4} = \frac{\tau_m}{2T} + \frac{\tau_e(|u|^2+1)}{2T^2} \pm \sqrt{\left(\frac{\tau_m}{2T} + \frac{\tau_e(|u|^2+1)}{2T^2}\right)^2 - \frac{\tau_m}{T}\frac{\tau_e}{T^2}}\ > 0. \tag{4.3.26}$$

Remark 4.3.2 *Setting* $\omega = 0$ *in Theorem 4.3.1, we obtain the stabilization matrix proposed in [23], which is based on numerical experiments.*

Remark 4.3.3 *Using the dimensional analysis described in this section, we can state the symmetrized Navier-Stokes equations in dimensionless quantities:*

$$\tilde{A}_\ell V_{,\ell} - \frac{1}{Re}(\tilde{K}_{ij}V_{,j})_{,i} = 0, \quad \textit{for}\ \ell = 0,\ldots,3,\ i,j = 1,2,3, \tag{4.3.27}$$

where Re denotes the Reynolds number.

This follows directly from the following relations. Define the reference values for the viscosity μ_r, specific heat at constant pressure c_{p_r} and heat conductivity κ_r using the independent set of reference values $\{U, \Theta, R, L\}$ and the dimensionless Pi groups $\mathrm{Re}, \mathrm{Pr}, \mathrm{Ec}$, as:

$$\mu_r = \frac{RUL}{\mathrm{Re}}, \quad c_{p_r} = \frac{U^2}{\Theta \mathrm{Ec}}, \quad \kappa_r = \frac{\mu_r c_{p_r}}{\mathrm{Pr}} = \frac{RU^3 L}{\Theta \mathrm{Re}\,\mathrm{Pr}\,\mathrm{Ec}},$$

where Pr denotes the Prandtl number and Ec the Eckert number. Introducing the reference values into the symmetrized system and using (4.3.8), it follows that the dimensional and dimensionless forms of the Navier-Stokes equations are related in the following way

$$\tilde{A}_0 V_{,t} + \tilde{A}_i V_{,i} - (\tilde{K}_{ij} V_{,j})_{,i} = \frac{RU}{L} D_r \left(\tilde{A}_0^\star V_{,t^\star}^\star + \tilde{A}_i^\star \frac{\partial V^\star}{\partial x_i^\star} - \frac{1}{\mathrm{Re}} \frac{\partial}{\partial x_i^\star} \left(\tilde{K}_{ij}^\star \frac{\partial V^\star}{\partial x_j^\star} \right) \right),$$

where the dimensionless variables are marked with a star and $D_r = \mathrm{diag}\,(1, U, U, U, U^2)$. Note that the dimensionless Jacobian matrices have the same functional form as their dimensionfull counterparts, only the entries r and d_i, $i = 1, 2, 3$ in the Jacobian matrices $\tilde{A}_\ell$ and $\tilde{K}_{ii}$, respectively, appear in their dimensionless forms in $\tilde{A}_\ell^\star$ and $\tilde{K}_{ii}^\star$ with different coefficients

$$r^\star = 2k^\star + \frac{c_p^\star}{\mathrm{Ec}} T, \quad d_i^\star = \frac{1}{3}\mu^\star u_i^{\star 2} + \mu^\star |u^\star|^2 + \frac{\kappa^\star}{\mathrm{Pr}\ \mathrm{Ec}} T^\star,$$

where we used the same notations for the matrix entries as in the Appendix. Therefore, the stars can be dropped from the non-dimensional equations. Observe that the factor $1/\mathrm{Re}$ can be extracted from the viscous coefficient matrices and (4.3.27) is obtained.

Remark 4.3.4 *It is straightforward to prove that the dimensionless symmetrized incompressible Navier-Stokes equations (4.3.27) are equivalent with the continuity equation*

$$u_{i,i} = 0, \quad i = 1, 2, 3,$$

the momentum equations

$$u_{i,t} + u_j u_{i,j} = -\frac{1}{\rho} p_{,i} + \frac{1}{Re} (\nu s_{ij})_{,j}, \quad i, j = 1, 2, 3,$$

with $\nu(T) = \mu(T)/\rho$ the kinematic viscosity, and $s_{ij} = u_{i,j} + u_{j,i}$ the shear stress tensor, and the equation for the temperature field

$$\frac{\rho c_p}{Ec} (T_{,t} + u_i T_{,i}) = \frac{\mu}{Re} \sum_{i,j=1}^{3} \frac{1}{2} s_{ij}^2 + \frac{1}{Re\, Pr\, Ec} (\kappa T_{,i})_{,i} \quad i = 1, 2, 3.$$

In the remainder of this chapter we will use now the same symbols for the dimensionless quantities.

4.4 Analysis of a class of stabilization operators

The class of stabilization matrices for the entropy variables derived in Section 4.3 is dimensionally consistent and positive definite. These are essential requirements for the stabilization operator, but not sufficient to ensure that the numerical scheme is stable and has a unique solution. In this section we investigate necessary conditions for this by analyzing the coercivity of the linearized Galerkin least-squares finite element discretization for the incompressible Navier-Stokes equations. This will result in sufficient conditions to ensure coercivity for a class of stabilization operators which belong to the framework given by Theorems 4.3.1 and 4.3.2.

If we specify the reference variables U, L, R and Θ in (4.3.17), and introduce the function ξ, then we can specify the coefficients τ_c, τ_m and τ_e in (4.3.17) in terms of the flow variables and element size.

Definition 4.4.1 *The stabilization parameters $\tau_c,\ \tau_m$ and τ_e on the elements $\mathcal{E}_n^e \in \mathcal{T}_h^n$ are defined as*

$$\tau_c(x) = \frac{h_e|u(x)|}{2},\ \tau_m(x) = \frac{h_e}{2\rho|u(x)|}\xi(Re_e(x)),\ \tau_e(x) = \frac{\tau_m(x)}{c_v}$$

for all $x \in \mathcal{E}_n^e$, with

$$Re_e(x) = \frac{m_k\rho|u(x)|h_e}{\mu(x)}, \quad m_k = \min\{1, C_k\}, \tag{4.4.1}$$

$$\xi(Re_e(x)) = \begin{cases} Re_e(x), & 0 \le Re_e(x) < 1 \\ 1, & Re_e(x) \ge 1, \end{cases} \tag{4.4.2}$$

where μ is the fluid viscosity, h_e denotes the element diameter and C_k is a positive constant independent of the physical properties and element diameter h_e.

The motivation for the function ξ in the definition of τ_m and τ_e is the need in the coercivity proof for an upper bound on these parameters which depends on h_e^2. Let us first show that the stability parameter τ_m is bounded in each element by a constant. By definition, for all $x \in \mathcal{E}_n^e$,

$$\tau_m(x) = \frac{h_e}{2\rho|u(x)|}, \quad \mathrm{Re}_e(x) \ge 1, \tag{4.4.3}$$

$$\tau_m(x) = \frac{m_k h_e^2}{2\mu(x)}, \qquad 0 \le \mathrm{Re}_e(x) < 1. \tag{4.4.4}$$

Therefore, for $\mathrm{Re}_e(x) \ge 1$,

$$\tau_m(x) = \frac{h_e}{2\rho|u(x)|}\frac{1}{\mathrm{Re}_e(x)}\frac{m_k\rho|u(x)|h_e}{\mu(x)} \le \frac{m_k h_e^2}{2\mu(x)}, \quad \forall x \in \mathcal{E}_n^e, \tag{4.4.5}$$

and combined with (4.4.4), we conclude that the bound (4.4.5) is valid for all values of $\mathrm{Re}_e(x)$. A similar estimate is valid for τ_e,

$$\tau_e(x) \leq \frac{m_k h_e^2}{2c_v \mu(x)}, \quad \forall x \in \mathcal{E}_n^e \,. \tag{4.4.6}$$

The next lemma provides sufficient conditions such that the stabilization matrix $\tilde{\tau}$ satisfies the requirements of Theorem 4.3.2, which ensures that the stabilization matrix is positive definite.

Lemma 4.4.1 *Using Definition 4.4.1, the stabilization matrix $\tilde{\tau}$ in (4.3.14) is positive definite for all $\omega \in \left(\frac{-1-\sqrt{5}}{2}, \frac{-1+\sqrt{5}}{2}\right)$.*

Proof:
Using Theorem 4.3.2 and Definition 4.4.1, we obtain that (4.3.19) is equivalent with

$$\omega^2 + \omega - \frac{1}{\xi(\mathrm{Re}_e(x))} < 0 \quad \text{for all } x \in \mathcal{E}_n^e, \tag{4.4.7}$$

and combined with (4.4.2) this completes the proof. □

For the analysis of the coercivity of the Galerkin discretization it is convenient to separate the continuity equation, which is the first equation in the system (4.1.3), from the other equations. For this purpose we introduce the variable $\hat{V}$, which is the image of V under the projection $\pi : \mathcal{E}^5 \to \mathcal{E}^4$, such that $\pi(V) = \hat{V} = (V_2, V_3, V_4, V_5)^T$, with a similar definition for $\hat{W}$. Further, we denote with $\hat{A}_\ell$ and $\hat{K}_{ij}$ the lower right 4×4 part of $\tilde{A}_\ell$, $\tilde{K}_{ij} \in \mathbb{R}^{5\times 5}$, respectively. In the remainder of this section we assume that these Jacobian matrices are constant.

Introduce the following linear operators:

$$\hat{\mathcal{L}} : \mathcal{E}^4 \to \mathbb{R}^4, \quad \hat{\mathcal{L}}\hat{V} = \hat{A}_\ell(\bar{V})\frac{\partial \hat{V}}{\partial x_\ell} - \frac{1}{\mathrm{Re}}\frac{\partial}{\partial x_i}\left(\hat{K}_{ij}(\bar{V})\frac{\partial \hat{V}}{\partial x_j}\right), \tag{4.4.8}$$

where $\hat{\mathcal{L}}^{inv}$ denotes the inviscid part of $\hat{\mathcal{L}}$, (i.e., $\hat{K}_{ij} = 0$),

$$\hat{\mathcal{D}} : \mathcal{E}^4 \to \mathbb{R}, \quad \hat{\mathcal{D}}\hat{V} = \bar{\rho}\bar{T}\left(\frac{\partial \hat{V}_i}{\partial x_i} + \bar{u}_i \frac{\partial \hat{V}_4}{\partial x_i}\right), \tag{4.4.9}$$

$$\hat{\mathcal{F}} : \mathcal{E} \to \mathbb{R}^4, \quad \hat{\mathcal{F}}V_1 = -\bar{\rho}\bar{T}\left(\frac{\partial V_1}{\partial x_1}, \frac{\partial V_1}{\partial x_2}, \frac{\partial V_1}{\partial x_3}, \bar{u}_i \frac{\partial V_1}{\partial x_i}\right)^T, \tag{4.4.10}$$

for $\ell = 0, \ldots, 3,\ i, j = 1, 2, 3$ and $\bar{V}, \bar{\rho}, \bar{T}, \bar{u}$ denote prescribed (vector) fields.

We consider now the linearized incompressible Navier-Stokes equations, defined in entropy variables on a bounded domain Ω, together with a source term $\hat{\mathcal{S}} : \mathcal{E} \to \mathbb{R}^4$:

$$\hat{\mathcal{D}}\hat{V} = 0, \qquad \text{in } \Omega, \tag{4.4.11}$$

$$\hat{\mathcal{L}}\hat{V} - \hat{\mathcal{F}}V_1 = \hat{\mathcal{S}}, \qquad \text{in } \Omega, \tag{4.4.12}$$

and boundary conditions

$$\hat{V} = g_w, \qquad \text{at } \partial\Omega, \tag{4.4.13}$$

with $\sum_{i=1}^{3} \int_{\partial\Omega} g_{wi} \bar{n}_i dS = 0$ and g_{wi}, $(i = 1, \cdots, 4)$ given entropy functions at the boundary $\partial\Omega$. We also introduce the variable $\tilde{V} = \hat{V} - g_w$ into (4.4.11)-(4.4.13), hence $\tilde{V}$ has homogeneous boundary conditions at $\partial\Omega$, and assume that there exists a bounded linear trace lifting operator ℓ for each component of g_w. This will make the treatment of the inhomogeneous boundary conditions much easier.

The splitting of V into V_1 and $\hat{V}$ is also applied to the stabilization matrix $\tilde{\tau}$, given by (4.3.14):

$$\tilde{\tau} = \begin{pmatrix} \delta & \sigma^T \\ \sigma & \hat{\tau} \end{pmatrix}, \tag{4.4.14}$$

where $\delta = \tilde{\tau}_{11}$, $\hat{\tau}$ is the lower right 4×4 submatrix of $\tilde{\tau}$ and

$$\sigma^T = \left(\tfrac{\omega u_1}{T}\tau_m + \tfrac{(h-k)u_1}{T^2}\tau_e, \quad \tfrac{\omega u_2}{T}\tau_m + \tfrac{(h-k)u_2}{T^2}\tau_e, \quad \tfrac{\omega u_3}{T}\tau_m + \tfrac{(h-k)u_3}{T^2}\tau_e, \quad -\tfrac{(h-k)}{T^2}\tau_e \right).$$

With the coefficients given in Definition 4.4.1 and ω satisfying Lemma 4.4.1, we have $\delta > 0$ and $\hat{\tau}$ is a symmetric positive definite matrix. Hence, using (4.3.25-4.3.26), it follows that we can write the smallest and largest eigenvalues of $\hat{\tau}$ as:

$$\lambda_{\min} = c_{\min}\tau_m > 0, \quad \text{and} \quad \lambda_{max} = c_{\max}\tau_m > 0,$$

where $c_{\min}$ and $c_{\max}$ are positive and functions of $|u|, T$ and c_v.

The weak formulation for (4.4.11)-(4.4.13) is obtained by multiplication of (4.4.11) with the test function W_1 and (4.4.12) with $\hat{W}$, integration by parts over the space time slab $\mathcal{E}_n$ and applying the homogeneous boundary conditions for $\tilde{V}$:

Within each space-time slab $\mathcal{E}_n$, find $(V_1, \tilde{V}) \in V_{0h}^n$, such that for all $(W_1, \hat{W}) \in W_{0h}^n$, the following relation is satisfied:

$$\begin{aligned}
B_n(V_1, \tilde{V}, W_1, \hat{W}) := & \frac{1}{2}(\hat{A}_\ell \tilde{V}_{,\ell} - \hat{\mathcal{F}} V_1, \hat{W})_{\mathcal{E}_n} - \frac{1}{2}(\tilde{V}, \hat{A}_\ell \hat{W}_{,\ell} - \hat{\mathcal{F}} W_1)_{\mathcal{E}_n} \\
& + \frac{1}{2}(\hat{\mathcal{D}}\tilde{V}, W_1)_{\mathcal{E}_n} - \frac{1}{2}(V_1, \hat{\mathcal{D}}\hat{W})_{\mathcal{E}_n} + \frac{1}{\mathrm{Re}}(\hat{K}_{ij}\tilde{V}_{,j}, \hat{W}_{,i})_{\mathcal{E}_n} \\
& + \frac{1}{2}(\tilde{V}(t_{n+1}^-, \cdot), \hat{W}(t_{n+1}^-, \cdot))_{\hat{A}_0, \Omega(t_{n+1})} \\
& - \frac{1}{2}(\tilde{V}(t_n^+, \cdot), \hat{W}(t_n^+, \cdot))_{\hat{A}_0, \Omega(t_n)} + B_{jump}^n(\tilde{V}, \hat{W}) \\
& + \sum_{\mathcal{E}_n^e \in \mathcal{T}_h^n} \big\{ (\hat{\mathcal{L}}\tilde{V} - \hat{\mathcal{F}} V_1 - \hat{\mathcal{S}}, \hat{\tau}(\hat{\mathcal{L}}\hat{W} - \hat{\mathcal{F}} W_1 - \hat{\mathcal{S}}))_{\mathcal{E}_n^e} + \\
& \qquad (\hat{\mathcal{D}}\tilde{V}, \delta\hat{\mathcal{D}}\hat{W})_{\mathcal{E}_n^e} + (\hat{\mathcal{L}}\tilde{V} - \hat{\mathcal{F}} V_1 - \hat{\mathcal{S}}, \hat{\mathcal{D}}\hat{W}\sigma)_{\mathcal{E}_n^e} +
\end{aligned}$$

$$(\hat{\mathcal{D}}\tilde{V}\sigma, \hat{\mathcal{L}}\hat{W} - \hat{\mathcal{F}}W_1 - \hat{\mathcal{S}})_{\mathcal{E}_n^e}\big\}$$

$$= B_n(0, \ell(g_w), W_1, \hat{W}) + (\hat{S}, \hat{W})_{\mathcal{E}_n}, \tag{4.4.15}$$

where $\ell(g_w)$ denotes the trace lifting of g_w from the boundary $\partial\Omega$ to Ω, and the jump term is defined as:

$$B_{jump}^n(\tilde{V}, \hat{W}) = \int_{\Omega(t_n)} \hat{W}^T(t_n^+)\hat{A}_0 \left(\tilde{V}(t_n^+) - \tilde{V}(t_n^-)\right) \mathrm{d}\Omega. \tag{4.4.16}$$

Finally, summing over all space time slabs we obtain the weak formulation for the whole space-time domain:

Find $(V_1, \tilde{V}) \in V_{0h}$, such that for all $(W_1, \hat{W}) \in W_{0h}$, the following relation is satisfied:

$$B_N(V_1,\tilde{V}, W_1, \hat{W}) := \frac{1}{2}\sum_{n=0}^{N}\big\{(\hat{A}_\ell \tilde{V}_{,\ell} - \hat{\mathcal{F}}V_1, \hat{W})_{\mathcal{E}_n} - (\tilde{V}, \hat{A}_\ell \hat{W}_{,\ell} - \hat{\mathcal{F}}W_1)_{\mathcal{E}_n}\big\}$$

$$+ \frac{1}{2}\sum_{n=0}^{N}\big\{(\hat{\mathcal{D}}\tilde{V}, W_1)_{\mathcal{E}_n} - (V_1, \hat{\mathcal{D}}\hat{W})_{\mathcal{E}_n}\big\} + \frac{1}{\mathrm{Re}}\sum_{n=0}^{N}(\hat{K}_{ij}\tilde{V}_{,j}, \hat{W}_{,i})_{\mathcal{E}_n}$$

$$+ \frac{1}{2}\sum_{n=0}^{N}\big\{(\tilde{V}(t_{n+1}^-, \cdot), \hat{W}(t_{n+1}^-, \cdot))_{\hat{A}_0, \Omega(t_{n+1})} + (\tilde{V}(t_n^+, \cdot), \hat{W}(t_n^+, \cdot))_{\hat{A}_0, \Omega(t_n)}\big\}$$

$$- \sum_{n=1}^{N}(\tilde{V}(t_n^-), \hat{W}(t_n^+))_{\hat{A}_0, \Omega(t_n)}$$

$$+ \sum_{n=0}^{N}\sum_{\mathcal{E}_n^e \in \mathcal{T}_h^n}\big\{(\hat{\mathcal{L}}\tilde{V} - \hat{\mathcal{F}}V_1 - \hat{\mathcal{S}}, \hat{\tau}(\hat{\mathcal{L}}\hat{W} - \hat{\mathcal{F}}W_1 - \hat{\mathcal{S}}))_{\mathcal{E}_n^e} +$$

$$(\hat{\mathcal{D}}\tilde{V}, \delta\hat{\mathcal{D}}\hat{W})_{\mathcal{E}_n^e} + (\hat{\mathcal{L}}\tilde{V} - \hat{\mathcal{F}}V_1 - \hat{\mathcal{S}}, \hat{\mathcal{D}}\hat{W}\sigma)_{\mathcal{E}_n^e} +$$

$$(\hat{\mathcal{D}}\tilde{V}\sigma, \hat{\mathcal{L}}\hat{W} - \hat{\mathcal{F}}W_1 - \hat{\mathcal{S}})_{\mathcal{E}_n^e}\big\}$$

$$= B_N(0, \ell(g_w), W_1, \hat{W}) + \sum_{n=0}^{N}(\hat{S}, \hat{W})_{\mathcal{E}_n} + (\tilde{V}(t_0^-), \hat{W}(t_0^+))_{\hat{A}_0, \Omega(t_0)}. \tag{4.4.17}$$

In the next part of this we section we will prove that the bilinear form B_N with the class of stabilization operators given by Theorem 4.3.1, with coefficients given in Definition 4.4.1, is coercive in V_{0h}, hence (4.4.17) has a unique solution. Before we can prove this result, we first need some technical lemmas.

Lemma 4.4.2 *There exists a positive constant C_k, independent of the viscosity coefficient μ and element diameter h_e, such that*

$$C_k \sum_{\mathcal{E}_n^e \in \mathcal{T}_h^n} h_e^2 \|(\tilde{K}_{ij}V_{,j})_{,i}\|_{0,\mathcal{E}_n^e}^2 \leq \sum_{\mathcal{E}_n^e \in \mathcal{T}_h^n} \|\tilde{K}_{ij}V_{,j}\|_{0,\mathcal{E}_n^e}^2 \quad \textit{for all } V \in V_h^n. \tag{4.4.18}$$

Proof:
The proof of this lemma is a direct application of inverse estimates given in for instance [4]. Using Theorem 4.5.11 in [4] for $p = q = 2$, $l = 1$, $m = 0$, there exists $C = C(l, p, q, \rho)$ such that

$$\left(\sum_{e=1}^{n_{el(n)}} |v|^2_{H^1(\mathcal{E}_n^e)} \right)^{1/2} \leq \left(\sum_{e=1}^{n_{el(n)}} \|v\|^2_{H^1(\mathcal{E}_n^e)} \right)^{1/2} \leq Ch^{-1} \left(\sum_{e=1}^{n_{el(n)}} \|v\|^2_{L^2(\mathcal{E}_n^e)} \right)^{1/2} \tag{4.4.19}$$

Let $v = \tilde{K}_{ij} V_{,j}$, then (4.4.19) implies that

$$\sum_{e=1}^{n_{el(n)}} \|(\tilde{K}_{ij} V_{,j})_{,i}\|^2_{L^2(\mathcal{E}_n^e)} \leq C^2 h^{-2} \sum_{e=1}^{n_{el(n)}} \|\tilde{K}_{ij} V_{,j}\|^2_{L^2(\mathcal{E}_n^e)} \tag{4.4.20}$$

which is equivalent to (4.4.18) with $C_k = C^{-2}$. □

Lemma 4.4.3 *There exists a constant $\alpha^* = \alpha^*(|u|, T, \mu, \kappa) > 0$, such that*

$$\int_{\mathcal{E}_n} \tilde{V}^T_{,i} \hat{K}_{ij} \tilde{V}_{,j} \, \mathrm{d}\mathcal{E} - \alpha \|\hat{K}_{ij} \tilde{V}_{,j}\|^2_0 \geq 0 \tag{4.4.21}$$

for all $\alpha \in I_{\alpha^*}$, where $I_{\alpha^*} = \begin{cases} [0, \alpha^*] & \text{if } \alpha > 0 \\ [-\alpha^*, 0] & \text{if } \alpha < 0. \end{cases}$

Proof:
The expression in (4.4.21) is equivalent to

$$\int_{\mathcal{E}_n} \sum_{i=1}^{3} \left(\tilde{V}^T_{,i} \Big(\sum_{j=1}^{3} \hat{K}_{ij} \tilde{V}_{,j} \Big) - \alpha \Big(\sum_{m=1}^{3} \tilde{V}^T_{,m} \hat{K}^T_{im} \Big) \Big(\sum_{l=1}^{3} \hat{K}_{il} \tilde{V}_{,l} \Big) \right) \mathrm{d}\mathcal{E}. \tag{4.4.22}$$

Rearranging terms, we can rewrite (4.4.22) into matrix form as

$$\int_{\mathcal{E}_n} \left(\tilde{V}^T_{,1}, \tilde{V}^T_{,2}, \tilde{V}^T_{,3} \right) N(\alpha) \left(\tilde{V}^T_{,1}, \tilde{V}^T_{,2}, \tilde{V}^T_{,3} \right)^T \mathrm{d}\mathcal{E} \tag{4.4.23}$$

where the 12×12 matrix $N(\alpha)$ has the form

$$N(\alpha) = \left(\begin{array}{c|c|c} \hat{K}_{11} - \alpha \left(\hat{K}^T_{i1} \hat{K}_{i1} \right) & \hat{K}_{12} - \alpha \left(\hat{K}^T_{i1} \hat{K}_{i2} \right) & \hat{K}_{13} - \alpha \left(\hat{K}^T_{i1} \hat{K}_{i3} \right) \\ \hat{K}_{21} - \alpha \left(\hat{K}^T_{i2} \hat{K}_{i1} \right) & \hat{K}_{22} - \alpha \left(\hat{K}^T_{i2} \hat{K}_{i2} \right) & \hat{K}_{23} - \alpha \left(\hat{K}^T_{i2} \hat{K}_{i3} \right) \\ \hat{K}_{31} - \alpha \left(\hat{K}^T_{i3} \hat{K}_{i1} \right) & \hat{K}_{32} - \alpha \left(\hat{K}^T_{i3} \hat{K}_{i2} \right) & \hat{K}_{33} - \alpha \left(\hat{K}^T_{i3} \hat{K}_{i3} \right) \end{array} \right).$$

For the proof of this lemma it is sufficient to show that there exist $\alpha \in I_{\alpha^*}$ such that the matrix $N(\alpha)$ is positive-semidefinite for any given flow data. Let us write

$N(\alpha)$ in the form $N(\alpha) = -\alpha A + B$. Observe that $A = BB^T$. Using the property $\tilde{K}_{ij} = \tilde{K}_{ji}^T$ of the viscous flux Jacobian matrices for the entropy variables, it follows that $N(\alpha), A$ and B are symmetric matrices. Since B is symmetric, there exists a unique decomposition

$$B = R\Lambda R^T, \quad \text{with } RR^T = I, \tag{4.4.24}$$

where Λ is the diagonal matrix with eigenvalues of B, R the corresponding right eigenvector matrix and I the identity matrix. Since A is symmetric, a similar decomposition is valid also for A. On the other hand,

$$A = BB^T = R\Lambda R^T (R^T)^T \Lambda^T R^T = R\Lambda^2 R^T. \tag{4.4.25}$$

Using the uniqueness of such a decomposition for A, it follows that the matrices A and B have the same set of eigenvectors and, denoting by λ_A and λ_B the eigenvalues of the matrices A and B, respectively, the relation between the corresponding eigenvalues is $\lambda_A = \lambda_B^2$. The eigenvalues are functions of the flow variables $\mathcal{V} = \{|u|^2, T, \mu, \kappa\}$, but do not depend on the direction of the flow. Then, for any eigenvector of A and B, the following holds:

$$N(\alpha)v = (-\alpha A + B)v = (-\alpha\lambda_A + \lambda_B)v,$$

which means that v is also an eigenvector of $N(\alpha)$ and, if we denote the eigenvalues of $N(\alpha)$ by $\lambda_N(\alpha)$, they have the form

$$\lambda_N(\alpha) = -\alpha\lambda_A + \lambda_B = -\alpha\lambda_B^2 + \lambda_B.$$

One property of the symmetrizing variables is that the matrix B is positive-semidefinite, therefore $\lambda_B \geq 0$. The relation between the eigenvalues of A and B implies that A is also positive-semidefinite.

Consider the case $\alpha \geq 0$.

(a) Assume that $\lambda_B > 0$. Since

$$\lambda_N(\alpha = 0) = \lambda_B > 0 \quad \text{and} \quad \frac{\partial \lambda_N(\alpha)}{\partial \alpha} = -\lambda_B^2 < 0,$$

there exists an $\alpha^\star > 0$ and an interval $I = [0, \alpha^\star]$ such that $\lambda_N(\alpha) \geq 0$ for all $\alpha \in I$ and $\lambda_B \neq 0$. Moreover, we have $\alpha^* = 1/\lambda_{\max}$, where $\lambda_{\max} = \max\{\lambda \mid \lambda \in \text{sp}\,(B) \setminus \{0\}\}$, with $\text{sp}\,(B)$ denoting the spectrum of matrix B.

Consequently, we have found an α^*, which value only depends on the magnitude of velocity, temperature and viscosity coefficient, but not on the direction of the flow.

(b) Assume that $\lambda_B = 0$.

Then, the condition is trivially satisfied for all α.

The case $\alpha \leq 0$ is analogous. □

Before we state the main result of this section, we first make the following assumptions on the stabilization matrix.

Assumption 4.4.1 *The stabilization matrix $\tilde{\tau}$ and the fluid viscosity μ are element-wise constant.*

These assumptions are not essential for the coercivity proof, but remove unnecessary technical complications in the analysis. The next theorem shows that the Galerkin least squares discretization (4.4.17) results in a well posed problem with a unique solution.

Theorem 4.4.1 *Given the conditions stated in Assumption 4.4.1, and the stabilization matrix defined in Theorem 4.3.1, with coefficients given in Definition 4.4.1, then there exists a positive constant C, independent of $(V_1, \tilde{V})$, such that for all $(V_1, \tilde{V}) \in V_{0h}$ the following condition is satisfied*

$$B_N(\tilde{V}, V_1; \tilde{V}, V_1) \geq C \left|||V|||^2 , \tag{4.4.26}$$

where the norm $\left|||V|||^2$ is defined as

$$\begin{aligned}
\left|||V|||^2 = {} & \frac{1}{2}\|\tilde{V}(t_{N+1}^-, \cdot)\|^2_{\hat{A}_0, \Omega(t_{N+1}, \cdot)} + \frac{1}{2}\|\tilde{V}(t_0^+, \cdot)\|^2_{\hat{A}_0, \Omega(t_0)} \\
& + \frac{1}{2}\sum_{n=1}^{N} \|\tilde{V}(t_n^+, \cdot) - \tilde{V}(t_n^-, \cdot)\|^2_{\hat{A}_0, \Omega(t_n)} \\
& + \sum_{n=0}^{N} \sum_{\mathcal{E}_n^e \in \mathcal{T}_h} \big\{ \|V\|^2_{0, \mathcal{E}_n^e} + \|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0, \mathcal{E}_n^e} + \|\hat{\mathcal{D}}\tilde{V}\|^2_{0, \mathcal{E}_n^e} + \\
& \qquad \|\hat{\tau}^{1/2}(\hat{\mathcal{L}}^{inv}\tilde{V} - \hat{\mathcal{F}}V_1 - \hat{\mathcal{S}})\|^2_{0, \mathcal{E}_n^e} \big\}.
\end{aligned} \tag{4.4.27}$$

Remark 5. Note, since the polynomials are only linear in time the jumps across the space-time slabs at the different time levels in $|||\cdot|||$ provide sufficient conditions to specify also the time derivative completely for all $V \in V_{0h}$.

Proof:
Take $W = V$, the expression for $B_N(V_1, \tilde{V}; V_1, \tilde{V})$ given by (4.4.17) then is after some algebraic manipulations equivalent with

$$B_N(V_1, \tilde{V}; V_1, \tilde{V}) = \frac{1}{2}\|\tilde{V}(t_{N+1}^-, \cdot)\|^2_{\hat{A}_0, \Omega(t_{N+1})} + \frac{1}{2}\|\tilde{V}(t_0^+, \cdot)\|^2_{\hat{A}_0, \Omega(t_0)}$$

$$+\frac{1}{2}\sum_{n=1}^{N}\|\tilde{V}(t_n^+,\cdot)-\tilde{V}(t_n^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_n)}+\frac{1}{\mathrm{Re}}\sum_{n=0}^{N}(\hat{K}_{ij}\tilde{V}_{,j},\tilde{V}_{,i})_{\mathcal{E}_n}$$

$$+\sum_{n=0}^{N}\sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\big\{\|\delta^{1/2}\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}_n^e}+\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e}+$$

$$2(\hat{\mathcal{D}}\tilde{V}\sigma,\hat{\mathcal{L}}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})_{\mathcal{E}_n^e}\big\},$$

where $\hat{\tau}^{1/2}$ denotes the matrix square root of $\hat{\tau}$. Introduce the following estimate

$$2\left(\hat{\mathcal{D}}\tilde{V}\sigma,\hat{\mathcal{L}}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}}\right)_{\mathcal{E}_n^e}=2\left(\hat{\mathcal{D}}\tilde{V}\,\hat{\tau}^{-1/2}\sigma,\hat{\tau}^{1/2}(\hat{\mathcal{L}}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\right)_{\mathcal{E}_n^e}$$
$$\leq\epsilon_1\|\hat{\mathcal{D}}\tilde{V}\hat{\tau}^{-1/2}\sigma\|^2_{0,\mathcal{E}_n^e}+\frac{1}{\epsilon_1}\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e},$$

which is valid for all $\epsilon_1>0$, then we obtain

$$B_N(V_1,\tilde{V};V_1,\tilde{V})\geq\frac{1}{2}\|\tilde{V}(t_{N+1}^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_{N+1})}+\frac{1}{2}\|\tilde{V}(t_0^+,\cdot)\|^2_{\hat{A}_0,\Omega(t_0)}$$

$$+\frac{1}{2}\sum_{n=1}^{N}\|\tilde{V}(t_n^+,\cdot)-\tilde{V}(t_n^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_n)}+\frac{1}{\mathrm{Re}}\sum_{n=0}^{N}(\hat{K}_{ij}\tilde{V}_{,j},\tilde{V}_{,i})_{\mathcal{E}_n}$$

$$+\sum_{n=0}^{N}\sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\big\{\|\delta^{1/2}\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}_n^e}-\epsilon_1\|\hat{\mathcal{D}}\tilde{V}\hat{\tau}^{-1/2}\sigma\|^2_{0,\mathcal{E}_n^e}+$$

$$(1-\frac{1}{\epsilon_1})\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e}\big\}. \qquad (4.4.28)$$

The last norm in (4.4.28) can be further estimated by splitting the operator $\hat{\mathcal{L}}$ into the inviscid and viscous part:

$$\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e}=\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}^{inv}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e}$$

$$+\frac{1}{\mathrm{Re}^2}\|\hat{\tau}^{1/2}(\hat{K}_{ij}\tilde{V}_{,j})_{,i}\|^2_{0,\mathcal{E}_n^e}$$

$$-2(\hat{\mathcal{L}}^{inv}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}},\frac{1}{\mathrm{Re}}\hat{\tau}(\hat{K}_{ij}\tilde{V}_{,j})_{,i})_{\mathcal{E}_n^e}$$

$$\geq(1-\frac{1}{\epsilon_2})\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}^{inv}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e}$$
$$+(1-\epsilon_2)\frac{1}{\mathrm{Re}^2}\|\hat{\tau}^{1/2}(\hat{K}_{ij}\tilde{V}_{,j})_{,i}\|^2_{0,\mathcal{E}_n^e}, \qquad (4.4.29)$$

for all $\epsilon_2>0$. Furthermore, since we assume that $\tilde{\tau}$ is constant on each element, the second to last norm in (4.4.28) can be written as

$$\|\hat{\mathcal{D}}\tilde{V}\hat{\tau}^{-1/2}\sigma\|^2_{0,\mathcal{E}_n^e}=(\sigma^T\hat{\tau}^{-1}\sigma)\|\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}_n^e}. \qquad (4.4.30)$$

Note here that since $\hat{\tau}$ is positive definite, so is $\hat{\tau}^{-1}$, which means that $\sigma^T\hat{\tau}^{-1}\sigma > 0$ for all $\sigma \in \mathbb{R}^4 \setminus \{0\}$. Using (4.4.29), (4.4.30) and $\delta > 0$, we obtain

$$\begin{aligned}
B_N(V_1,\tilde{V};V_1,\tilde{V}) \geq\ & \frac{1}{2}\|\tilde{V}(t_{N+1}^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_{N+1})} + \frac{1}{2}\|\tilde{V}(t_0^+,\cdot)\|^2_{\hat{A}_0,\Omega(t_0)} \\
&+ \frac{1}{2}\sum_{n=1}^{N}\|\tilde{V}(t_n^+,\cdot)-\tilde{V}(t_n^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_n)} + \frac{1}{\mathrm{Re}}\sum_{n=0}^{N}(\hat{K}_{ij}\tilde{V}_{,j},\tilde{V}_{,i})_{\mathcal{E}_n} \\
&+ \sum_{n=0}^{N}\sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\big\{(\delta-\epsilon_1(\sigma^T\hat{\tau}^{-1}\sigma))\|\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}_n^e} + \\
&\qquad (1-\frac{1}{\epsilon_1})(1-\frac{1}{\epsilon_2})\|\hat{\tau}^{\frac{1}{2}}(\hat{\mathcal{L}}^{inv}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e} + \\
&\qquad (1-\frac{1}{\epsilon_1})(1-\epsilon_2)\frac{1}{\mathrm{Re}^2}\|\hat{\tau}^{1/2}(\hat{K}_{ij}\tilde{V}_{,j})_{,i}\|^2_{0,\mathcal{E}_n^e}\big\}.
\end{aligned}$$

Since $\tilde{\tau}$ is element-wise constant, using the norm of $\hat{\tau}$, which is defined as

$$\|\hat{\tau}\| = \max_i\{\lambda_i : \lambda_i \in \mathrm{sp}\,(\hat{\tau})\} = c_{\max}\tau_m,$$

in combination with Lemma 4.4.2 and the bound on the stabilization parameter τ_m (4.4.5), we obtain

$$\begin{aligned}
\sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\|\hat{\tau}^{1/2}(\hat{K}_{ij}\tilde{V}_{,j})_{,i}\|^2_{0,\mathcal{E}_n^e} &\leq \sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\frac{c_{\max}m_k}{2\mu C_k}\|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n^e} \\
&\leq \frac{c_{\max}}{2\mu}\sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n^e},
\end{aligned}$$

where in the last inequality we used the definition of m_k (4.4.1) and the assumption that μ is element-wise constant. Summarizing, we obtain the estimate

$$\begin{aligned}
B_N(V_1,\tilde{V};V_1,\tilde{V}) \geq\ & \frac{1}{2}\|\tilde{V}(t_{N+1}^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_{N+1})} + \frac{1}{2}\|\tilde{V}(t_0^+,\cdot)\|^2_{\hat{A}_0,\Omega(t_0)} \\
&+ \frac{1}{2}\sum_{n=1}^{N}\|\tilde{V}(t_n^+,\cdot)-\tilde{V}(t_n^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_n)} + \\
&+ \sum_{n=0}^{N}\sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\big\{\big(\delta-\epsilon_1(\sigma^T\hat{\tau}^{-1}\sigma)\big)\|\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}_n^e} \\
&+ (1-\frac{1}{\epsilon_1})(1-\frac{1}{\epsilon_2})\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}^{inv}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e} \\
&+ (1-\frac{1}{\epsilon_1})(1-\epsilon_2)\frac{c_{\max}}{2\mu\mathrm{Re}^2}\|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n^e} + \frac{1}{\mathrm{Re}}(\hat{K}_{ij}\tilde{V}_{,j},\tilde{V}_{,i})_{\mathcal{E}_n^e}\big\}.
\end{aligned} \tag{4.4.31}$$

Define the coefficient α as

$$\alpha = (1 - \frac{1}{\epsilon_1})(\epsilon_2 - 1)\frac{c_{\max}}{2\mu \mathrm{Re}}, \tag{4.4.32}$$

then we can always choose $\epsilon_1 > 1$ and $\epsilon_2 > 1$, but sufficiently close to one, such that α is in the interval $0 < \alpha < \alpha^*$. Applying Lemma 4.4.3, we obtain with $\alpha = \epsilon_3 \alpha^*$ the relation

$$\begin{aligned}
&(\hat{K}_{ij}\tilde{V}_{,j}, \tilde{V}_{,i})_{\mathcal{E}_n} - \alpha \|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n} \\
&\qquad = \Big((\hat{K}_{ij}\tilde{V}_{,j}, \tilde{V}_{,i})_{\mathcal{E}_n} - \alpha^* \|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n} \Big) + \alpha^*(1-\epsilon_3)\|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n} \\
&\qquad \geq \alpha^*(1-\epsilon_3)\|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n},
\end{aligned}$$

for any $0 < \epsilon_3 < 1$, where in the last inequality we used (4.4.21) with $\alpha = \alpha^*$.

For the remaining part of the coercivity proof we need to show that there exists an $\epsilon_1 > 1$ such that the condition

$$\delta - \epsilon_1(\sigma^T \hat{\tau}^{-1}\sigma) > 0 \tag{4.4.33}$$

is satisfied. In order to investigate (4.4.33), we consider the following function of ϵ_1 : $f(\epsilon_1) = \delta - \epsilon_1(\sigma^T \hat{\tau}^{-1}\sigma)$, where $\delta > 0$ and $\sigma^T \hat{\tau}^{-1}\sigma > 0$. Observe that

$$f(\epsilon_1 = 1) = \frac{1}{\overline{\rho}\overline{T}}\tau_c - \omega(\omega+1)\frac{|u|^2}{\overline{T}}\tau_m. \tag{4.4.34}$$

Note that $f(\epsilon_1 = 1) > 0$ is always satisfied since it is equivalent to condition (4.3.24) on the positive definiteness of the stabilization operator $\tilde{\tau}$. Consequently, there exists $1 < \epsilon_1 < \epsilon_1^*$ such that (4.4.33) is satisfied. Since in order to satisfy the condition $0 < \alpha < \alpha^*$, with α given in (4.4.32), we only need a value $\epsilon_1 > 1$ arbitrary close to one, hence we can always find an ϵ_1 which satisfies both conditions.

Combining all terms, we obtain that for each element there exist $\epsilon_1 > 1$, $\epsilon_2 > 1$ and $0 < \epsilon_3 < 1$ such that

$$\begin{aligned}
B_N(V_1, \tilde{V}; V_1, \tilde{V}) \geq\ & \frac{1}{2}\|\tilde{V}(t^-_{N+1}, \cdot)\|^2_{\hat{A}_0, \Omega(t_{N+1})} + \frac{1}{2}\|\tilde{V}(t^+_0, \cdot)\|^2_{\hat{A}_0, \Omega(t_0)} \\
&+ \frac{1}{2}\sum_{n=1}^{N} \|\tilde{V}(t^+_n, \cdot) - \tilde{V}(t^-_n, \cdot)\|^2_{\hat{A}_0, \Omega(t_n)} \\
&+ \sum_{n=0}^{N} \sum_{\mathcal{E}^e_n \in \mathcal{T}^n_h} \Big\{ \frac{(1-\epsilon_3)\alpha^*}{\mathrm{Re}} \|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n} \\
&+ \big(\delta - \epsilon_1(\sigma^T \hat{\tau}^{-1}\sigma)\big) \|\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}^e_n}
\end{aligned}$$

$$+ (1 - \frac{1}{\epsilon_1})(1 - \frac{1}{\epsilon_2})\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}^{inv}\tilde{V} - \hat{\mathcal{F}}V_1 - \hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e}\Big\}. \qquad (4.4.35)$$

The last step in the proof is to apply the Poincaré-Friedrichs inequality for V_1, which applies since $\int_{\mathcal{E}_n} V_1 \mathrm{d}\mathcal{E} = 0$, and to use the Poincaré inequality for $\tilde{V}$, which is valid since Ω is bounded and $\tilde{V} = 0$ at $\partial\Omega$. Combining these inequalities results in the estimate

$$\|V\|_{0,\mathcal{E}_n} \leq C_{\mathcal{P}}\|\nabla V\|_{0,\mathcal{E}_n},$$

with $C_{\mathcal{P}}$ the Poincaré constant. Introduce the coefficient

$$C = \min_{\mathcal{E}_n^e \in \mathcal{T}_h^n}\left\{\min\left\{1, \frac{(1-\epsilon_3)\alpha^*}{\mathrm{Re}}, \delta - \epsilon_1(\sigma^T\hat{\tau}^{-1}\sigma), (1-\frac{1}{\epsilon_1})(1-\frac{1}{\epsilon_2})\right\}\right\} > 0,$$

then we can further evaluate (4.4.35) into

$$\frac{C}{C_{\mathcal{P}}^2}\sum_{n=0}^{N}\|V\|^2_{0,\mathcal{E}_n} \leq C\,|||V|||^2_* \leq B_N(V_1, \tilde{V}; V_1, \tilde{V}).$$

where

$$\begin{aligned}
|||V|||^2_* = {} & \frac{1}{2}\|\tilde{V}(t^-_{N+1},\cdot)\|^2_{\hat{A}_0,\Omega(t_{N+1})} + \frac{1}{2}\|\tilde{V}(t^+_0,\cdot)\|^2_{\hat{A}_0,\Omega(t_0)} \\
& + \frac{1}{2}\sum_{n=1}^{N}\|\tilde{V}(t^+_n,\cdot) - \tilde{V}(t^-_n,\cdot)\|^2_{\hat{A}_0,\Omega(t_n)} \\
& + \sum_{n=0}^{N}\sum_{\mathcal{E}_n^e \in \mathcal{T}_h^n}\Big\{\|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n} + \|\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}_n^e} \\
& + \|\hat{\tau}^{1/2}(\hat{\mathcal{L}}^{inv}\tilde{V} - \hat{\mathcal{F}}V_1 - \hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e}\Big\} \qquad (4.4.36)
\end{aligned}$$

Adding $\frac{C}{C_{\mathcal{P}}^2}|||V|||^2_*$ to the inequalities finally results in

$$\frac{C}{1+C_{\mathcal{P}}^2}|||V|||^2 \leq B_N(V_1, \tilde{V}; V_1, \tilde{V}).$$

□

The following corollary shows the importance of the viscous contribution of $\hat{\mathcal{L}}\hat{W}$ in the least-squares operator for small Reynolds numbers.

Corollary 4.4.1 *The coercivity of the Galerkin least squares method (4.4.17), with $\hat{\mathcal{L}}\hat{W}$ replaced by $\hat{\mathcal{L}}^{inv}\hat{W}$, can only be guaranteed for Reynolds numbers $Re \gg 1$.*

Proof:
Similar as in Theorem 4.4.1, we find the following lower bound

$$
\begin{aligned}
B_N^*(V_1,\tilde{V};V_1,\tilde{V}) \geq\ & \frac{1}{2}\|\tilde{V}(t_{N+1}^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_{N+1})} + \frac{1}{2}\|\tilde{V}(t_0^+,\cdot)\|^2_{\hat{A}_0,\Omega(t_0)} \\
& + \frac{1}{2}\sum_{n=1}^{N}\|\tilde{V}(t_n^+,\cdot)-\tilde{V}(t_n^-,\cdot)\|^2_{\hat{A}_0,\Omega(t_n)} + \|V\|^2_{0,\mathcal{E}} \\
& + \sum_{n=0}^{N}\sum_{\mathcal{E}_n^e\in\mathcal{T}_h^n}\Big\{\big(\delta-(\epsilon_1+\frac{1}{2\epsilon_3})(\sigma^T\hat{\tau}^{-1}\sigma)\big)\|\hat{\mathcal{D}}\tilde{V}\|^2_{0,\mathcal{E}_n^e} \\
& + (1-\frac{1}{\epsilon_1}-\frac{1}{2\epsilon_2})\|\hat{\tau}^{1/2}(\hat{\mathcal{L}}^{inv}\tilde{V}-\hat{\mathcal{F}}V_1-\hat{\mathcal{S}})\|^2_{0,\mathcal{E}_n^e} \\
& + \frac{1}{\mathrm{Re}}(\hat{K}_{ij}\tilde{V}_{,j},\tilde{V}_{,i})_{\mathcal{E}_n} - (\epsilon_2+\epsilon_3)\frac{c_{\max}}{4\mu\mathrm{Re}^2}\|\hat{K}_{ij}\tilde{V}_{,j}\|^2_{0,\mathcal{E}_n^e}\Big\}. \qquad (4.4.37)
\end{aligned}
$$

For all $\epsilon_1 > 1$, but sufficiently close to one, $\epsilon_2 > \epsilon_1/(2(\epsilon_1-1))$ and $\epsilon_3 > 1/(2(1-\epsilon_1+\epsilon^*)) > 0$, with small $\epsilon^* > 0$, we can ensure that the first two coefficients depending on ϵ_1, ϵ_2 and ϵ_3 on the right-hand side of (4.4.37) are positive. It is, however, in general not possible to choose ϵ_2 and ϵ_3 such that the condition

$$(\epsilon_2+\epsilon_3)\frac{c_{\max}}{4\mu\mathrm{Re}} \leq \alpha^*$$

is also satisfied, which is necessary in order to apply Lemma 4.4.3. This is only possible if the Reynolds number Re is sufficiently larger than one. □

4.5 Concluding remarks

In this chapter we have derived a class of stabilization operators suitable for the incompressible limit of the symmetrized formulation for the Navier-Stokes equations given in [23]. This set of equations consists of the incompressible Navier-Stokes equations and the heat equation. We derived a class of dimensionally consistent stabilization operators and give conditions which guarantee the positive definiteness of the stabilization matrix and the coercivity of the Galerkin least squares finite element discretization for the linearized problem. These are necessary conditions to ensure that the numerical discretization is stable and results in a unique solution. The analysis also shows that only for large Reynolds numbers it is possible to neglect the viscous operator acting on the test functions in the stabilization operator.

Further research will be conducted towards the actual performance of the stabilization operators discussed in this chapter and we will also report in the next chapter on our research on stabilization operators for weakly compressible flows.

Chapter 5

Construction of stabilization operators for weakly compressible flows

In this chapter we discuss several aspects of the design of stabilization operators for a Galerkin least-squares finite element method suitable for both the compressible and incompressible Navier-Stokes equations, as we described also in [45]. As mentioned in Chapter 3, there is not much in common between the large variety of numerical schemes developed for the compressible and incompressible Navier-Stokes equations. The concept of symmetrized equations using entropy variables is, however, a good starting point towards a unified formulation which is valid for both types of flows, see [2]. Symmetrized equations using entropy variables in compressible flow, have been investigated by Godunov [18], Mock [43], Harten [21], Hughes et al. [30], Dutt [12] and Johnson et al. [37]. The use of the symmetrized compressible Navier-Stokes equations employing entropy variables results in a global entropy stability which is automatically inherited by the numerical discretization, see for instance Shakib et al. [51]. This is not true when for instance conservative or primitive variables are used. The concept of symmetrization is also important for incompressible flows which we discussed in Chapter 3. For a detailed analysis of the entropy stability of the symmetrized Navier-Stokes equations and the choice of variables, see Barth [3] and Hauke and Hughes [23]. We will summarize the main results on entropy stability at the end of this chapter.

The Galerkin method applied to the compressible Navier-Stokes equations lacks stability and spurious oscillations occur, generated for instance by unresolved internal and boundary layers. To improve the stability of the method, while maintaining the order of accuracy, a least-squares operator is added to the basic Galerkin formulation. The Galerkin least-squares method is a linear method, therefore, produces oscillatory

approximations to discontinuities due to Godunov's theorem, see [54]. For non-smooth solutions, there is therefore also a need for a so-called discontinuity-capturing operator to overcome oscillations around discontinuities. This term, introduced in [29], provides additional control over the gradients in the discrete solution and increases the robustness of the method. The discontinuity-capturing operator is, however, not part of this study. For more details we refer to [29], [51].

The concept of a stabilization operator can also be used for incompressible flows, and, as discussed in detail in Chapter 4, eliminates the complications of designing elements which satisfy the inf-sup stability condition.

The development of stabilized methods, starting from it's first applications till recent results, as well as the different types of stabilization operators are described in detail in the introduction of this thesis.

This chapter addresses two issues. The first one is the design of the stabilization operator in the Galerkin least-squares finite element method, which is critical for the accuracy and stability of the discretization. In Chapter 4 a stabilization operator is proposed for the Galerkin least-squares discretization of the incompressible Navier-Stokes equations using dimensional analysis and a careful verification of conditions ensuring the positive definiteness of the stabilization matrix. In the construction of the stabilization matrix for the compressible Navier-Stokes equations we use the same concept as used in the incompressible case. This is, therefore, another example of ideas that can be used for both types of flows.

We start the consistent mathematical derivation of the stabilization operator using the primitive variables in the construction of the stabilization matrix and apply dimensional analysis to determine its dependence on the flow variables. Next, this systematic derivation of the stabilization operator for the primitive variables is extended to the set of entropy variables. In [23], Hauke and Hughes demonstrated that using either the primitive variables employing pressure or the entropy variables, the same formulation can be used to compute both compressible and incompressible flows. The difficulty of this unified approach is to design a stabilization matrix which is valid for both type of flows. In general, the compressible stabilization matrix is not well defined in the incompressible limit and conversely, the incompressible stabilization is not effective for transonic and supersonic flows. In the same paper, the authors give a first attempt to define a stabilization matrix suitable for both compressible and incompressible flows. This is however, based on numerical experiments and the authors emphasize the need of further research to define a suitable stabilization matrix. It is, therefore very challenging to give a unified formulation of stabilization operators, valid for both compressible and incompressible flows. We will demonstrate that the consistent construction of stabilization operators discussed in this chapter is applicable to both flow regimes. The resulting stabilization matrix is also well defined in the incompressible limit since it is identical to the stabilization matrix we obtained for

incompressible flows in Chapter 4 of this thesis. Note however, that our definition of the unified stabilization matrix is restricted to weakly compressible flows.

The second topic of this chapter is the analysis of the proposed stabilization operator in order to ensure non-linear stability. The Galerkin least-squares method using the symmetrized form of the equations satisfies the Clausius-Duhem inequality or entropy condition, which results in a non-linear stability condition, as discussed in [3] and [50]. Therefore, we give necessary and sufficient conditions on the positive definiteness of the stabilization operator. These conditions provide additional information on the admissible stabilization operators.

5.1 The Galerkin least-squares variational formulation

Consider the compressible Navier-Stokes equations in a time-dependent flow domain $\Omega(t)$. Let $Y : \mathcal{E} \mapsto \mathbb{R}^5$ denote the vector of primitive variables $(p, u_1, u_2, u_3, T)^T$ and $\Phi : \mathbb{R}^5 \mapsto \mathbb{R}^{5\times 4}$ the flux tensor defined in (4.1.1), with the flux vector in the ℓth coordinate direction F_ℓ, $(\ell = 0, \ldots, 3)$ given by the columns of Φ. Using these notations, the compressible Navier-Stokes equations can be written in conservative form as

$$F_\ell(Y(x))_{,\ell} - (K_{ij}(Y(x))Y_{,j})_{,i} = 0, \quad x \in \mathcal{E}, \tag{5.1.1}$$

where $K_{ij} \in \mathbb{R}^{5\times 5}$ for $i, j = 1, 2, 3$ denote the viscous flux Jacobian matrices and the summation convention is used on repeated indices. In [23], Hauke and Hughes demonstrated that using either the variables Y or the entropy variables V, defined in (3.2.2), the incompressible limit of the Navier-Stokes equations is well defined.

Introducing the set of entropy variables V into the quasi-linear form (5.1.1), we obtain the symmetrized form of the compressible Navier-Stokes equations (see for details Chapter 3) as:

$$\tilde{A}_0(V)V_{,t} + \tilde{A}_i(V)V_{,i} = (\tilde{K}_{ij}(V)V_{,j})_{,i},$$

with $\tilde{A}_0$, $\tilde{A}_i$ and $\tilde{K}_{ij}$ given in Appendix B.1. Note that the coefficient matrices $\tilde{A}_\ell$ are expressed in terms of the volume expansivity α_p, isothermal compressibility β_T and specific heat at constant pressure c_p, whereas the viscous flux Jacobian matrices do not depend on these compressibility parameters.

Before presenting the Galerkin least-squares method, we need to introduce some notations. We use the same notations for the partitioning of the space-time domain as in Section 4.2 and for completeness, we briefly summarize them here.

Consider a partitioning of the time interval $I = (t_0, t_{N+1})$ using the time levels $t_0 < t_1 < \ldots < t_{N+1}$. We denote by $I_n = (t_n, t_{n+1})$ the nth time interval and define a space-time slab as $\mathcal{E}_n = \mathcal{E} \cap I_n$. Each space-time slab $\mathcal{E}_n$ is bounded by the hypersurfaces

$\Omega(t_n)$, $\Omega(t_{n+1})$ and $\mathcal{Q}_n = \partial\mathcal{E}_n \setminus (\Omega(t_n) \cup \Omega(t_{n+1}))$. In each space-time slab $\mathcal{E}_n$ we define a partition $\mathcal{T}_h^n$ into $(n_e)_n$ non-overlapping elements $\mathcal{E}_n^e$. The space-time elements $\mathcal{E}_n^e$ are obtained by splitting the spatial domain $\Omega(t_n)$ into a set of non-overlapping elements Ω_n^e and connecting them with a mapping Φ_t^n to the elements $\Omega_{n+1}^e \subset \Omega(t_{n+1})$ at time t_{n+1}.

Since we are working with two different sets of variables, the corresponding finite element spaces need to be defined. The trial function spaces for entropy and primitive variables in each space-time slab $\mathcal{E}_n$ are denoted by $\mathcal{S}_V^n$ and $\mathcal{S}_Y^n$, respectively, and the test function space by $\mathcal{W}_V^n$ and $\mathcal{W}_Y^n$. Their elements are assumed to be $\mathcal{C}^0$ continuous within each space-time slab, but discontinuous across the interfaces of the space-time slabs, namely at times $t_1, t_2, \ldots, t_{N-1}$. The finite element spaces are now defined as:

$$\begin{aligned}\mathcal{S}_V^n = \big\{ & V \in C^0(\mathcal{E}_n)^5 : V \mid_{\mathcal{E}_n^e} \circ G_n^e \in \Big(\hat{\mathcal{P}}_1(0,1) \otimes \hat{\mathcal{P}}_k(\hat{\Omega})\Big)^5, \\ & \forall\, \mathcal{E}_n^e \in \mathcal{T}_h^n,\ q_1(V) = \bar{q}_1 \text{ on } \mathcal{Q}_n \big\}\end{aligned}$$

$$\begin{aligned}\mathcal{W}_V^n = \big\{ & W \in C^0(\mathcal{E}_n)^5 : W \mid_{\mathcal{E}_n^e} \circ G_n^e \in \Big(\hat{\mathcal{P}}_1(0,1) \otimes \hat{\mathcal{P}}_k(\hat{\Omega})\Big)^5, \\ & \forall\, \mathcal{E}_n^e \in \mathcal{T}_h^n,\ q_2(W) = \bar{q}_2 \text{ on } \mathcal{Q}_n \big\},\end{aligned}$$

where G_n^e denotes the mapping from the space-time reference element $(0,1) \times \hat{\Omega}$, with $\hat{\Omega}$ the reference element in R^3, to the element in physical space $\mathcal{E}_n^e$ and $\hat{\mathcal{P}}_k$ represent kth-order polynomials. Further, $q_1 : \mathcal{E}^5 \to \mathbb{R}^5$ are the (nonlinear) boundary conditions, with a similar expression for $q_2 : \mathcal{E}^5 \to \mathbb{R}^5$, and $\bar{q}_1, \bar{q}_2 \in \mathbb{R}^5$ are the prescribed boundary conditions. When the finite element spaces are defined on the whole space-time domain then the superscript n is omitted. The finite element spaces for the primitive variables can be defined analogously.

Recall the Galerkin least-squares variational formulation for the compressible Navier-Stokes equations in terms of the entropy variables:

Find $V \in \mathcal{S}_V$ such that for all $W \in \mathcal{W}_V$ the following relation is satisfied:

$$\begin{aligned}&\sum_{n=0}^{N} \Big\{ \int_{\mathcal{E}_n} \Big(-W_{,0} \cdot F_0(V) - W_{,i} \cdot F_i(V) + W_{,i} \cdot \tilde{K}_{ij} V_{,j} \Big) \mathrm{d}\mathcal{E} \\ &+ \int_{\Omega(t_{n+1})} W(t_{n+1}^-) \cdot F_0(V(t_{n+1}^-)) \mathrm{d}\Omega - \int_{\Omega(t_n)} W(t_n^+) \cdot F_0(V(t_n^-)) \mathrm{d}\Omega \\ &+ \sum_{e=1}^{(n_{el})_n} \int_{\mathcal{E}_n^e} (\mathcal{L}_V^T W) \cdot \tilde{\tau} (\mathcal{L}_V V) \mathrm{d}\mathcal{E} \\ &+ \int_{\mathcal{Q}_n} W \cdot \Big(F_i(V) - \tilde{K}_{ij} V_{,j} \Big) \bar{n}_i \, \mathrm{d}\mathcal{Q} - \int_{\mathcal{Q}_n} \bar{n} \cdot v(W \cdot F_0(V)) \mathrm{d}\mathcal{Q} \Big\} = 0, \end{aligned} \tag{5.1.2}$$

where n is the unit outward space-time normal vector at the boundary $\mathcal{Q}_n$ and $\bar{n}$ its the spatial component. The compressible Navier-Stokes differential operator for entropy variables is given by

$$\mathcal{L}_V = \tilde{A}_0(V)\frac{\partial}{\partial x_0} + \tilde{A}_i(V)\frac{\partial}{\partial x_i} - \frac{\partial}{\partial x_i}\left(\tilde{K}_{ij}(V)\frac{\partial}{\partial x_j}\right).$$

Similarly, for the primitive variables the weak formulation is given by:

Find $Y \in \mathcal{S}_Y$ such that for all $W \in \mathcal{W}_Y$ the following relation is satisfied:

$$\begin{aligned}
&\sum_{n=0}^{N}\Bigg\{ \int_{\mathcal{E}_n}\Big(-W_{,0}\cdot F_0(Y) - W_{,i}\cdot F_i(Y) + W_{,i}\cdot K_{ij}(Y)Y_{,j}\Big)\mathrm{d}\mathcal{E} \\
&+ \int_{\Omega(t_{n+1})} W(t_{n+1}^-)\cdot F_0(Y(t_{n+1}^-))\mathrm{d}\Omega - \int_{\Omega(t_n)} W(t_n^+)\cdot F_0(Y(t_n^-))\mathrm{d}\Omega \\
&+ \sum_{e=1}^{(n_{el})_n}\int_{\mathcal{E}_n^e}(\mathcal{L}_Y^T W)\cdot\tau_Y(\mathcal{L}_Y Y)\mathrm{d}\mathcal{E} \\
&+ \int_{\mathcal{Q}_n} W\cdot\Big(F_i(Y) - K_{ij}(Y)Y_{,j}\Big)\bar{n}_i\ \mathrm{d}\mathcal{Q} - \int_{\mathcal{Q}_n}\bar{n}\cdot v(W\cdot F_0(Y))\mathrm{d}\mathcal{Q}\Bigg\} = 0. \qquad (5.1.3)
\end{aligned}$$

The compressible Navier-Stokes equations written in terms of the primitive variables Y have the form

$$\mathcal{L}_Y Y = A_0(Y)Y_{,0} + A_i(Y)Y_{,i} - (K_{ij}(Y)Y_{,j})_{,i} = 0, \qquad (5.1.4)$$

with the coefficient matrices $A_\ell(Y)$ and $K_{ij}(Y)$ given in Appendix B.2. Note that the Jacobian matrices for primitive variables $A_\ell(Y)$ can also be expressed using the compressibility parameters α_p and β_T and the specific heat c_p, and also for this set of variables the viscous flux Jacobian matrices $K_{ij}(Y)$ do not depend on these parameters. As discussed in earlier sections, this formulation is also suitable for the incompressible limit.

The transposed of the differential operator for primitive variables is defined by

$$\mathcal{L}_Y^T = A_\ell^T(Y)\frac{\partial}{\partial x_\ell} - \frac{\partial}{\partial x_i}\left(K_{ij}^T(Y)\frac{\partial}{\partial x_j}\right) \quad \text{for } \ell = 0,\ldots,3,\ i,j = 1,2,3.$$

Note that when entropy variables are used, $\mathcal{L}_V = \mathcal{L}_V^T$ because of the symmetry of the coefficient matrices. The first and the last two integrals in (5.1.3) result from the partial integration of the Galerkin term expressed as a function of the variables Y. The jump term remains unchanged through the change of variables, with the fluxes written in terms of the primitive variables Y. The weak formulations for primitive and entropy variables can be transformed one into the other which implies that the stabilization parameters for the two sets of variables are related as

$$\tau_Y = Y_{,V}\tilde{\tau}.$$

5.2 Dimensional analysis

In order to derive a dimensionally consistent stabilization operator in the Galerkin least-squares discretization of the incompressible Navier-Stokes equations, we introduced in Section 4.3 the concept "*a* scales like *b*." Using the definitions of Section 4.3, and following the same procedure as used for incompressible flows, we will derive now a dimensionally consistent stabilization matrix suitable for both compressible and incompressible flows. The main steps in the construction are:

- Dimensional analysis of the stabilization matrix τ_Y related to the primitive variables $Y = (p, u_1, u_2, u_3, T)^T$.
- The stabilization operator τ_Y is related to the stabilization operator for the entropy variables $\tilde{\tau}$ through the transformation

$$\tilde{\tau} = V_{,Y} \tau_Y . \tag{5.2.1}$$

Our first goal is to derive a dimensionally consistent stabilization operator for the compressible Navier-Stokes equations in primitive variables. We first consider the equations $\mathcal{L}_Y Y = 0$, with $Y = (p, u_1, u_2, u_3, T)^T$ the set of primitive variables.

Consider the canonical set $\mathcal{V} = \{u, T, \rho, l\}$, and the set of reference values $r(\mathcal{V}) = \{U, \Theta, R, L\}$, with U, Θ, R and L the reference values for velocity, temperature, density and length, respectively, as defined in Section 4.3. Using the definition of equivalence classes introduced in Section 4.3, the coefficients of the volume expansivity α_p and isothermal compressibility β_T, obey the following scaling relations

$$[\alpha_p]_{\mathcal{V}} = \frac{1}{\Theta}, \quad [\beta_T]_{\mathcal{V}} = \frac{1}{RU^2}. \tag{5.2.2}$$

Hence, using the definition of the various vectors and matrices given in the Appendix B.2, the following dimensional equivalence is valid

$$[Y_{,0}]_{\mathcal{V}} = \frac{U}{L} \begin{pmatrix} RU^2 \\ U \\ U \\ U \\ \Theta \end{pmatrix}, \; [Y_{,i}]_{\mathcal{V}} = \frac{U}{L} \begin{pmatrix} RU \\ 1 \\ 1 \\ 1 \\ \Theta/U \end{pmatrix},$$

$$[A_0(Y)]_{\mathcal{V}} = \begin{pmatrix} 1/U^2 & 0 & 0 & 0 & R/\Theta \\ 1/U & R & 0 & 0 & RU/\Theta \\ 1/U & 0 & R & 0 & RU/\Theta \\ 1/U & 0 & 0 & R & RU/\Theta \\ 1 & RU & RU & RU & RU^2/\Theta \end{pmatrix},$$

$$[A_i(Y)]_{\mathcal{V}} = \begin{pmatrix} 1/U & \delta_{1i}R & \delta_{2i}R & \delta_{3i}R & RU/\Theta \\ 1 & RU & \delta_{2i}RU & \delta_{3i}RU & RU^2/\Theta \\ 1 & \delta_{1i}RU & RU & \delta_{3i}RU & RU^2/\Theta \\ 1 & \delta_{1i}RU & \delta_{2i}RU & RU & RU^2/\Theta \\ U & RU^2 & RU^2 & RU^2 & RU^3/\Theta \end{pmatrix},$$

where δ_{ij} is the Kronecker delta symbol. For the viscosity coefficient matrices we obtain for $i = j$:

$$[K_{ii}(Y)]_{\mathcal{V}} = L \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & RU & 0 & 0 & 0 \\ 0 & 0 & RU & 0 & 0 \\ 0 & 0 & 0 & RU & 0 \\ 0 & RU^2 & RU^2 & RU^2 & RU^3/\Theta \end{pmatrix},$$

and for $i \neq j$ we have

$$[K_{ij}(Y)]_{\mathcal{V}} = L \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{12}RU & a_{13}RU & 0 \\ 0 & a_{21}RU & 0 & a_{23}RU & 0 \\ 0 & a_{31}RU & a_{32}RU & 0 & 0 \\ 0 & b_{11}RU^2 & b_{22}RU^2 & b_{33}RU^2 & 0 \end{pmatrix},$$

with $i, j = 1, 2, 3$. The coefficients in the matrices $K_{ij}(Y)$ are defined as

$$a_{kl} = \begin{cases} 1 \text{ if } (k = i \wedge l = j) \vee (k = j \wedge l = i) \\ 0 \text{ otherwise} \end{cases} \qquad b_{kk} = \begin{cases} 1 \text{ if } k = i \vee k = j \\ 0 \text{ otherwise,} \end{cases}$$

for $k, l = 1, 2, 3$. By dimensional consistency, we can add the various contributions and obtain the following dimensional equivalence for the Navier-Stokes equations

$$[\mathcal{L}_Y Y]_{\mathcal{V}} = [A_0(Y)Y_{,0} + A_i(Y)Y_{,i} - (K_{ij}(Y)Y_{,j})_{,i}]_{\mathcal{V}} = \frac{RU}{L}(1, U, U, U, U^2)^T. \quad (5.2.3)$$

Requirement 5.2.1 *Construct a stabilized finite element method, which satisfies the following requirements:*

(a) The method admits discrete solutions Y^h with the same dimension as the solution Y of (5.1.4).

(b) Elementwise the least-squares stabilization operator is dimensionally equivalent with the Galerkin operator.

Similar assumptions are made in [58], where a scaling analysis is performed to determine the appropriate low Mach number behavior of the stabilization matrix. These requirements imply

$$\left[(\mathcal{L}_Y^T W) \cdot \tau_Y(\mathcal{L}_Y Y)\right]_{\mathcal{V}} = [W \cdot (\mathcal{L}_Y Y)]_{\mathcal{V}}, \quad \forall W \in W_h^n,$$

which is equivalent with

$$\frac{1}{L}\left(\left(U[A_0^T(Y)]_{\mathcal{V}}+[A_i^T(Y)]_{\mathcal{V}}-\frac{1}{L}[K_{ij}^T(Y)]_{\mathcal{V}}\right)[W]_{\mathcal{V}}\right)^T[\tau_Y(\mathcal{L}_Y Y)]_{\mathcal{V}}=[W^T]_{\mathcal{V}}\,[(\mathcal{L}_Y Y)]_{\mathcal{V}} \tag{5.2.4}$$

where the scales U and $1/L$ originate from the derivatives of the test function. Note that using $[\mu]_{\mathcal{V}}=RUL$ and $[\kappa]_{\mathcal{V}}=RU^3L/\Theta$, it follows that $A_i(Y)\sim_{\mathcal{V}}(1/L)K_{ij}(Y)$ for all $i,j=1,2,3$, hence the addition and subtraction are well defined in (5.2.4). Since the test functions are arbitrary, equation (5.2.4) is equivalent to

$$\frac{1}{L}\left(U[A_0(Y)]_{\mathcal{V}}+[A_i(Y)]_{\mathcal{V}}-\frac{1}{L}[K_{ij}(Y)]_{\mathcal{V}}\right)[\tau_Y]_{\mathcal{V}}\,[(\mathcal{L}_Y Y)]_{\mathcal{V}}=[(\mathcal{L}_Y Y)]_{\mathcal{V}}\,. \tag{5.2.5}$$

Therefore, requirements (a) and (b) provide an additional condition on the components of the stabilization matrix τ_Y, i.e., (5.2.5) is equivalent to:

$$\frac{1}{L}\begin{pmatrix}1/U & R & R & R & RU/\Theta\\ 1 & RU & RU & RU & RU^2/\Theta\\ 1 & RU & RU & RU & RU^2/\Theta\\ 1 & RU & RU & RU & RU^2/\Theta\\ U & RU^2 & RU^2 & RU^2 & RU^3/\Theta\end{pmatrix}[\tau_Y]_{\mathcal{V}}\,[(\mathcal{L}_Y Y)]_{\mathcal{V}}=[(\mathcal{L}_Y Y)]_{\mathcal{V}}\,. \tag{5.2.6}$$

Consider an arbitrary nonsingular matrix M represented by the matrix in (5.2.6), that is

$$[M]_{\mathcal{V}}=\frac{1}{L}\begin{pmatrix}1/U & R & R & R & RU/\Theta\\ 1 & RU & RU & RU & RU^2/\Theta\\ 1 & RU & RU & RU & RU^2/\Theta\\ 1 & RU & RU & RU & RU^2/\Theta\\ U & RU^2 & RU^2 & RU^2 & RU^3/\Theta\end{pmatrix}.$$

Note that such nonsingular matrix always exist since the constants in the matrix are arbitrary and each class is closed to linear combinations. We can, therefore write (5.2.6) as

$$[M\tau_Y]_{\mathcal{V}}\,[(\mathcal{L}_Y Y)]_{\mathcal{V}}=[(\mathcal{L}_Y Y)]_{\mathcal{V}}\,, \tag{5.2.7}$$

where we used the property $[M]_{\mathcal{V}}[\tau_Y]_{\mathcal{V}}=[M\tau_Y]_{\mathcal{V}}$. The scaling relation (5.2.7) shows that a suitable stabilization matrix τ_Y is dimensionally equivalent to the inverse of the nonsingular matrix M, i.e.,

$$[M^{-1}]_{\mathcal{V}}=[\tau_Y]_{\mathcal{V}}. \tag{5.2.8}$$

Therefore, we obtain that the stabilization matrix τ_Y has the following scaling property

$$[\tau_Y]_{\mathcal{V}} = L \begin{pmatrix} U & 1 & 1 & 1 & 1/U \\ 1/R & 1/(RU) & 1/(RU) & 1/(RU) & 1/(RU^2) \\ 1/R & 1/(RU) & 1/(RU) & 1/(RU) & 1/(RU^2) \\ 1/R & 1/(RU) & 1/(RU) & 1/(RU) & 1/(RU^2) \\ \Theta/(RU) & \Theta/(RU^2) & \Theta/(RU^2) & \Theta/(RU^2) & \Theta/(RU^3) \end{pmatrix},$$

which defines the structure of a dimensionally consistent stabilization matrix τ_Y. It is important to note that we did not require that $\tau_Y = M^{-1}$. Moreover, it is straightforward to see that using (5.2.8) we obtain

$$[M]_{\mathcal{V}}[\tau_Y]_{\mathcal{V}} = \begin{pmatrix} 1 & 1/U & 1/U & 1/U & 1/U^2 \\ U & 1 & 1 & 1 & 1/U \\ U & 1 & 1 & 1 & 1/U \\ U & 1 & 1 & 1 & 1/U \\ U^2 & U & U & U & 1 \end{pmatrix},$$

which satisfies condition (5.2.7). Summarizing, we write the general form of the stabilization matrix in primitive variables, indicating the dimension of the entries in the matrix, as

$$\tau_Y = L \left(\begin{array}{c|ccc|c} c_{11}U & c_{12} & c_{13} & c_{14} & \frac{c_{15}}{U} \\ \hline \frac{c_{21}}{R} & \frac{c_{22}}{RU} & \frac{c_{23}}{RU} & \frac{c_{24}}{RU} & \frac{c_{25}}{RU^2} \\ \frac{c_{31}}{R} & \frac{c_{32}}{RU} & \frac{c_{33}}{RU} & \frac{c_{34}}{RU} & \frac{c_{35}}{RU^2} \\ \frac{c_{41}}{R} & \frac{c_{42}}{RU} & \frac{c_{43}}{RU} & \frac{c_{44}}{RU} & \frac{c_{45}}{RU^2} \\ \hline \frac{c_{51}\Theta}{RU} & \frac{c_{52}\Theta}{RU^2} & \frac{c_{53}\Theta}{RU^2} & \frac{c_{54}\Theta}{RU^2} & \frac{c_{55}\Theta}{RU^3} \end{array}\right), \tag{5.2.9}$$

where c_{ij} are functions of the dimensionless variables and R, U, Θ, L the reference density, velocity, temperature and length, respectively. In this form the matrix still has 25 unknowns which need to be specified. In the remaining part of this chapter we will specify the coefficients c_{ij} in the stabilization matrix τ_Y (5.2.9) such that it is suitable for both compressible and incompressible flow. The first step towards this is a discussion on the asymptotic behavior of the stabilization matrix τ_Y in the incompressible limit, given in the next section.

5.3 The asymptotic behavior of the stabilization matrix in the incompressible limit

Recall the following thermodynamic relation between the specific heats, valid for any type of material:

$$c_p - c_v = \frac{\alpha_p^2 v T}{\beta_T}. \tag{5.3.1}$$

Use the dimensional identity

$$[c_p - c_v]_{\mathcal{V}} = \frac{U^2}{\Theta} = \left[\frac{\alpha_p^2 v T}{\beta_T}\right]_{\mathcal{V}},$$

and consider (5.3.1) in the dimensionless form. In the incompressible limit $c_p = c_v$, as we discussed in Section 2.2.1, therefore, for $\alpha_p = 0$ and $\beta_T = 0$, we have

$$\frac{\alpha_p^2 v T}{\beta_T} = 0$$

which is only possible, since v and T are bounded, when $\alpha_p^2 = O\left(\beta_T^{1+\epsilon}\right)$ for $\epsilon > 0$, as $\beta_T \to 0$. Equivalently,

$$\alpha_p = O(\beta_T^{\delta}), \quad \text{with } \delta = \frac{1}{2} + \frac{\epsilon}{2},\ \epsilon > 0 \text{ as } \beta_T \to 0. \tag{5.3.2}$$

Additional to Requirement 5.2.1, we add the following condition:

Requirement 5.3.1 *The asymptotic behavior of the Galerkin and least-squares stabilization operator are identical in the incompressible limit, that is $\forall W \in W_h^n$ the following is satisfied*

$$(\mathcal{L}_Y^T W) \cdot \tau_Y (\mathcal{L}_Y Y) \sim W \cdot (\mathcal{L}_Y Y), \quad \textit{as } \alpha_p \to 0,\ \beta_T \to 0,$$

where the relation $\sim$ denotes asymptotic behavior.

Since the test function is arbitrary and the flux Jacobian matrices $A_i(Y)$ and the Navier-Stokes equations $\mathcal{L}_Y Y$ posses the following properties, respectively,

$$A_i(Y) = \begin{pmatrix} O(\beta_T) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(1) \end{pmatrix}, \quad \mathcal{L}_Y Y = \begin{pmatrix} O(1) \\ O(1) \\ O(1) \\ O(1) \\ O(1) \end{pmatrix}$$

as $\alpha_p \to 0$ and $\beta_T \to 0$, we obtain that Requirement 5.3.1 implies that a suitable stabilization matrix should have the following asymptotic behavior

$$\tau_Y = \begin{pmatrix} O(1) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(\alpha_p) \\ O(1) & O(1) & O(1) & O(1) & O(1) \end{pmatrix} \quad \text{as } \alpha_p \to 0,\ \beta_T \to 0. \tag{5.3.3}$$

The asymptotic behavior can also be expressed using β_T by introducing (5.3.2) into (5.3.3), and we obtain that the stabilization matrix for nearly incompressible flow must behave like

$$\tau_Y = \begin{pmatrix} O(1) & O(1) & O(1) & O(1) & O(\beta_T^\delta) \\ O(1) & O(1) & O(1) & O(1) & O(\beta_T^\delta) \\ O(1) & O(1) & O(1) & O(1) & O(\beta_T^\delta) \\ O(1) & O(1) & O(1) & O(1) & O(\beta_T^\delta) \\ O(1) & O(1) & O(1) & O(1) & O(1) \end{pmatrix} \quad \text{as } \beta_T \to 0,$$

with $\delta > 1/2$. The asymptotic behavior of τ_Y as $\alpha_p \to 0$, obtained in (5.3.3) will be used in the next section to give an explicit form of the stabilization matrices τ_Y and $\tilde{\tau}$.

5.4 Construction of the stabilization matrix

The dimensionally consistent stabilization matrix obtained in (5.2.9), in combination with the analysis on its asymptotic behavior in the incompressible limit results in the following class of stabilization operators.

Theorem 5.4.1 *A class of dimensionally consistent stabilization matrices τ_Y for primitive variables with a well defined incompressible limit as $\alpha_p \to 0$ can be stated as*

$$\tau_Y = \left(\begin{array}{c|c|c} \tau_c & \rho(\omega+1)\left(\tau_m + (h-k)\alpha_p\tau_e\right)u^T & 2\rho\alpha_p\tau_e k \\ \omega\left(\tau_m + (h-k)\alpha_p\tau_e\right)u & I_{3\times 3}\tau_m & \alpha_p\tau_e u \\ -(h-k)\tau_e & (\alpha_p T - 1)\tau_e u^T & \tau_e \end{array}\right) \tag{5.4.1}$$

where $u = (u_1, u_2, u_3)^T$ is the velocity vector, h the specific enthalpy, $k = |u|^2/2$, $\omega \in \mathbb{R}$ and $\tau_c, \tau_m, \tau_e \in \mathbb{R}^+$.

Proof:
Let us write

$$\frac{c_{25}}{RU^2} = \tau_r u_1, \quad \frac{c_{35}}{RU^2} = \tau_r u_2, \quad \frac{c_{45}}{RU^2} = \tau_r u_3 \tag{5.4.2}$$

for some $\tau_r \in P(S)$. This relation implies that $[\tau_r]_{\mathcal{V}} = 1/(RU^3)$. Using the symmetry of $\tilde{\tau}$ and (5.2.1), we obtain the following relations for the coefficients c_{ij} in (5.2.9):

$$c_{23} = c_{32},\ c_{24} = c_{42},\ c_{34} = c_{43}, \tag{5.4.3}$$

$$\begin{aligned}
c_{12} &= \frac{\rho(c_{22}u_1 + c_{23}u_2 + c_{24}u_3 + c_{21}U)}{RU} + \rho u_1 \tau_r (h+k) - \frac{c_{15}u_1}{U},\\
c_{13} &= \frac{\rho(c_{23}u_1 + c_{33}u_2 + c_{34}u_3 + c_{31}U)}{RU} + \rho u_2 \tau_r (h+k) - \frac{c_{15}u_2}{U},\\
c_{14} &= \frac{\rho(c_{24}u_1 + c_{34}u_2 + c_{44}u_3 + c_{41}U)}{RU} + \rho u_3 \tau_r (h+k) - \frac{c_{15}u_3}{U},\\
c_{51} &= -\frac{(h-k)c_{55}}{U^2} - \frac{\tau_r RUT|u|^2}{\Theta} + \frac{RTc_{15}}{\rho\Theta},\\
c_{52} &= -\frac{u_1 c_{55}}{U} + \frac{\tau_r RU^2 T u_1}{\Theta},\\
c_{53} &= -\frac{u_2 c_{55}}{U} + \frac{\tau_r RU^2 T u_2}{\Theta},\\
c_{54} &= -\frac{u_3 c_{55}}{U} + \frac{\tau_r RU^2 T u_3}{\Theta}.
\end{aligned}$$

Since $[k]_{\mathcal{V}} = [h]_{\mathcal{V}} = U^2$, all operations in the above relations are valid. Consider now the middle 3×3 block in (5.2.9), which corresponds to the three momentum equations. Relation (5.4.3) implies that this block is symmetric. Moreover, this block must be rotational invariant, which together with it's symmetry, implies that it is a constant times the identity matrix. The coefficients must therefore satisfy the relation $c_{22} = c_{33} = c_{44} = c$ and $c_{23} = c_{32} = c_{24} = c_{42} = c_{34} = c_{43} = 0$. For simplicity we introduce the following notation for the diagonal entries in τ_Y,

$$\tau_c := c_{11} UL, \quad \tau_m := \frac{cL}{RU}, \quad \tau_e := \frac{c_{55} L\Theta}{RU^3}. \tag{5.4.4}$$

Then, we can write

$$c_{51} = -\frac{RU(h-k)}{L\Theta}\tau_e - \frac{\tau_r RUT|u|^2}{\Theta} + \frac{RTc_{15}}{\rho\Theta}, \quad c_{52} = \frac{RU^2(\tau_r LT - \tau_e)u_1}{L\Theta},$$

$$c_{53} = \frac{RU^2(\tau_r LT - \tau_e)u_2}{L\Theta}, \quad c_{54} = \frac{RU^2(\tau_r LT - \tau_e)u_3}{L\Theta}$$

and obtain the relations

$$c_{12} = \frac{\rho u_1}{L}\tau_m + \frac{\rho}{R}c_{21} + \rho u_1(h+k)\tau_r - \frac{c_{15}u_1}{U}, \tag{5.4.5}$$

$$c_{13} = \frac{\rho u_2}{L}\tau_m + \frac{\rho}{R}c_{31} + \rho u_2(h+k)\tau_r - \frac{c_{15}u_2}{U}, \tag{5.4.6}$$

$$c_{14} = \frac{\rho u_3}{L}\tau_m + \frac{\rho}{R}c_{41} + \rho u_3(h+k)\tau_r - \frac{c_{15}u_3}{U}. \tag{5.4.7}$$

Since $\tau_m \neq 0$, it follows from (5.4.5-5.4.7) that there are at least three additional non-vanishing entries in the matrix τ_Y. The vector composed of $(c_{12}, c_{13}, c_{14})^T$, is multiplied in the least-squares operator with the momentum equations, which are rotational invariant, and this implies that $(c_{12}, c_{13}, c_{14})^T$ must also be rotational invariant. We can therefore write $(c_{12}, c_{13}, c_{14})^T = \eta u$, with $u = (u_1, u_2, u_3)^T$ and the scalar $[\eta]_{\mathcal{V}} = 1/U$. Using (5.4.5-5.4.7) we obtain the following relations

$$\frac{\rho}{R}(c_{21}, c_{31}, c_{41})^T = \left(\eta - \left(\frac{\rho\tau_m}{L} + \rho(h+k)\tau_r - \frac{c_{15}}{U}\right)\right) u.$$

Since $[\eta]_{\mathcal{V}} = [\frac{\rho}{L}\tau_m]_{\mathcal{V}}$, we can choose

$$\eta = (\omega+1)\left(\frac{\rho\tau_m}{L} + \rho(h+k)\tau_r - \frac{c_{15}}{U}\right)$$

with $\omega \in \mathbb{R}$. Therefore,

$$(c_{12}, c_{13}, c_{14})^T = (\omega+1)\left(\frac{\rho\tau_m}{L} + \rho(h+k)\tau_r - \frac{c_{15}}{U}\right) u,$$

$$\frac{\rho}{R}(c_{21}, c_{31}, c_{41})^T = \omega\left(\frac{\rho\tau_m}{L} + \rho(h+k)\tau_r - \frac{c_{15}}{U}\right) u.$$

For notational simplicity let us choose

$$\frac{c_{15}L}{U} = \rho f, \quad \tau_r L = g. \tag{5.4.8}$$

Therefore,

$$[f]_{\mathcal{V}} = \frac{L}{RU}, \quad [g]_{\mathcal{V}} = \frac{L}{RU^3} \tag{5.4.9}$$

and introducing (5.4.8) into the entries of τ_Y in (5.2.9), we obtain

$$L(c_{12}, c_{13}, c_{14})^T = (\omega+1)\left(\rho\tau_m + \rho(h+k)g - \rho f\right) u,$$

$$\frac{L}{R}(c_{21}, c_{31}, c_{41})^T = \omega\left(\tau_m + (h+k)g - f\right) u.$$

Furthermore,

$$\frac{c_{51}L\Theta}{RU} = -(h-k)\tau_e - \tau_r L|u|^2 T + \frac{c_{15}LT}{\rho U} = -(h-k)\tau_e - g|u|^2 T + fT$$

$$\frac{c_{5i}L\Theta}{RU^2} = (\tau_r LT - \tau_e)u_{i-1} = (gT - \tau_e)u_{i-1}, \quad i = 2, 3, 4.$$

Inserting all relations for the constants $c_{ij},\ i,j=1,\ldots,5$ into (5.2.9), we obtain the general form of the stabilization matrix (5.4.1).

In order to obtain the proper asymptotic behavior in the incompressible limit we define f and g as follows

$$f=\alpha_p\tau_e|u|^2,\quad g=\alpha_p\tau_e.$$

This definition satisfies the scaling requirement (5.4.8). Furthermore, the analysis of the asymptotic behavior of τ_Y in (5.3.3) implies that in their dimensionless form, f and g should have the following asymptotic behavior in the incompressible limit

$$f=O(\alpha_p),\quad g=O(\alpha_p),\quad \text{when } \alpha_p\to 0,$$

which together with the scaling argument motivates their definition. This ensures that a well posed stabilization operator is obtained in the incompressible limit, which reduces to the stabilization operator (4.3.13) for the incompressible Navier-Stokes equations discussed in Chapter 4. □

Theorem 5.4.2 *The stabilization matrix $\tilde{\tau}$ related to the entropy variables is composed of*

$$\begin{aligned}
\tilde{\tau}_{11} &= \frac{1}{\rho T}\tau_c-\frac{\omega|u|^2}{T}\tau_m+\left(\frac{h-k}{T}\right)^2\tau_e-\frac{\omega|u|^2(h-k)\tau_e}{T}\alpha_p,\\
\tilde{\tau}_{1i+1} &= \frac{\omega u_i}{T}\tau_m+\frac{(h-k)u_i\tau_e}{T}\left(\frac{1}{T}+\omega\alpha_p\right)\quad i=1,2,3,\\
\tilde{\tau}_{15} &= -\frac{(h-k)\tau_e}{T^2}\\
\tilde{\tau}_{ii} &= \frac{\tau_m}{T}+\frac{u_i^2\tau_e}{T}\left(\frac{1}{T}-\alpha_p\right),\quad i=2,3,4,\\
\tilde{\tau}_{ij} &= \frac{u_{i-1}u_{j-1}\tau_e}{T}\left(\frac{1}{T}-\alpha_p\right),\quad i\neq j,\ i,j=2,3,4,\\
\tilde{\tau}_{i5} &= -\frac{u_{i-1}\tau_e}{T}(\frac{1}{T}-\alpha_p),\quad i=2,3,4,\\
\tilde{\tau}_{55} &= \frac{\tau_e}{T^2}.
\end{aligned}$$

Proof:
The stabilization operator $\tilde{\tau}$ can be obtained directly using the transformation (5.2.1), with τ_Y given by (5.4.1). It is straightforward to see that in the incompressible limit, the stabilization matrix for entropy variables reduces to the stabilization operator (4.3.14) obtained for the symmetrized incompressible Navier-Stokes equations. □

Our next aim is to show that the stabilization matrix $\tilde{\tau}$ in Theorem 5.4.2 is positive definite, which is a necessary and sufficient condition for entropy stability as will be discussed in Section 5.5. For this purpose we make the following assumption:

Assumption 5.4.1 *Assume there is a temperature range such that the relation*

$$\alpha_p \leq \frac{1}{T} \tag{5.4.10}$$

is valid.

This assumption is valid for many equations of state. Consider some examples of gas laws when α_p can be given analytically, as in Section 2.4.2.

- It is straightforward to see that (5.4.10) is satisfied for the **ideal gas law,** since $\alpha_p = 1/T$, and for the **co-volume equation of state**, since in this case

  $$\alpha_p = \frac{v-b}{Tv} < \frac{1}{T},$$

  since $b > 0s$.

- For the **van der Waals equation of state** we have

  $$\alpha_p = \frac{(v-b)v^2R}{v^3RT - 2a(v-b)^2}. \tag{5.4.11}$$

  Let us first analyze for what temperature range (5.4.10) is satisfied, with α_p given by (5.4.11). When $\alpha_p < 0$, then (5.4.10) is straightforward. Consider the case when $\alpha_p > 0$, which is satisfied, when $v - b > 0$ in (5.4.11) and in the temperature range

  $$T > \frac{2a(v-b)^2}{v^3R} = T_c. \tag{5.4.12}$$

  Condition (5.4.10) is equivalent with finding for which temperature range the following function has negative values

  $$d(T) = \alpha_p - \frac{1}{T} = \frac{-v^2bRT + 2a(v-b)^2}{T(v^3RT - 2a(v-b)^2)},$$

  which implies using (5.4.12) that

  $$d(T) < 0 \Longleftrightarrow T > \frac{2a(v-b)^2}{v^2Rb} = T^\star.$$

  Since $v > b$, it follows that $T_c < T^\star$, therefore, we can conclude that (5.4.10) is satisfied if $T > T^\star$.

- In Figure 5.1 some **measured values** of α_p for water at different temperatures are plotted as well as the function $1/T$, for the temperature range $243K \leq T \leq 373K$, showing that Assumption 5.4.1 is also valid for this case.

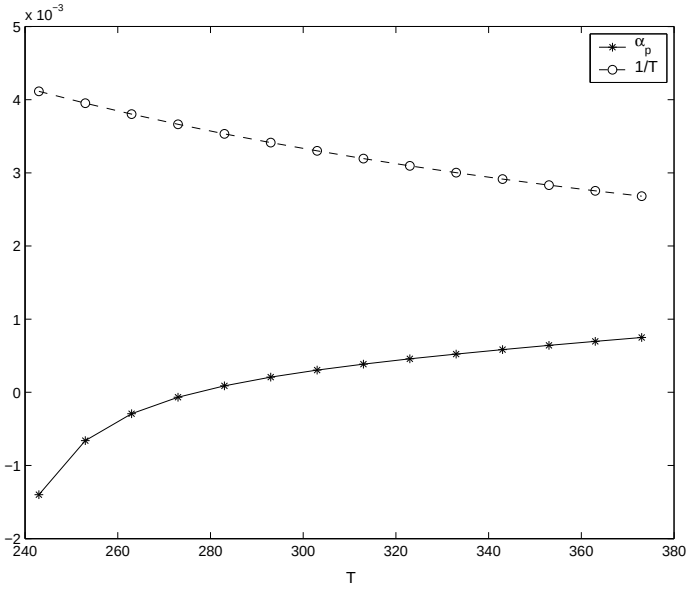

Figure 5.1: Measured values of α_p for water at different temperatures showing that α_p is less than $1/T$.

- Consider p and T as independent variables in a **weakly compressible** flow. As we discussed in Chapter 2 there are two intensive equations of state for the system. One of the equations of state relates density to temperature and pressure. Consider the total differential of the specific volume given by (2.3.1)

$$\mathrm{d}v = \alpha_p v \, \mathrm{d}T - \beta_T v \, \mathrm{d}p. \tag{5.4.13}$$

Since our interest is the range of weakly compressible flows, we assume pressure changes to be negligible. Then, (5.4.13) reduces to

$$\mathrm{d}\rho = -\rho\alpha_p \mathrm{d}T \tag{5.4.14}$$

where $\rho = 1/v$ is the density. The temperature can then be considered as a function of the density only and the following expansion of the temperature is valid

$$T = T_r + \left.\frac{\partial T}{\partial \rho}\right|_{\rho=\rho_r} (\rho - \rho_r) + O((\rho - \rho_r)^2),$$

where T_r and ρ_r are constant reference values of temperature and density. In case of weakly compressible flows the changes in density are small, therefore, we can neglect the higher order terms and obtain

$$T \approx T_r + \left.\frac{\partial T}{\partial \rho}\right|_{\rho=\rho_r} (\rho - \rho_r).$$

In combination with (5.4.14) this leads to an equation of state for a weakly compressible fluid in terms of the volume expansivity at the reference values, defined as

$$\alpha_{p_r} = -\frac{1}{\rho_r}\left(\frac{\partial \rho}{\partial T}\right)_p\Bigg|_{\rho=\rho_r},$$

viz.

$$T - T_r = -\frac{1}{\alpha_{p_r}\rho_r}(\rho - \rho_r).$$

Note that similar idea was used in [8] and [40] to obtain an equation of state for a slightly compressible fluid in terms of the bulk modulus. Therefore, we can express α_{p_r} as

$$\alpha_{p_r} = \frac{1-\rho/\rho_r}{T/T_r - 1}\,\frac{1}{T_r} < \frac{1}{T_r},$$

where in the last inequality we assumed (without loss of generality) that the reference values were chosen such that

$$\frac{1-\rho/\rho_r}{T/T_r - 1} \leq 1,$$

which is possible since the reference values are arbitrary. Hence, under the assumption that the fluid is slightly compressible it is possible to satisfy Assumption 5.4.1.

□

In Theorem 4.3.2 of Chapter 4, we gave necessary and sufficient conditions on the stabilization parameters τ_c, τ_m, τ_e and ω such that the stabilization matrix $\tilde{\tau}$ designed for incompressible flows is positive definite. In the next theorem we give similar conditions to guarantee the positive definiteness of the stabilization matrix $\tilde{\tau}$ for compressible flows stated in Theorem 5.4.2. Note, however, that we have an additional parameter α_p in the compressible stabilization matrix, which significantly complicates the analysis.

Theorem 5.4.3 *Under the condition stated in Assumption 5.4.1, the stabilization matrix $\tilde{\tau}$ given in Theorem 5.4.2 is positive definite if and only if the following conditions on the stabilization parameters τ_c, τ_m, τ_e and ω are satisfied*

$$\begin{cases} \tau_m > 0 \\ \tau_e > 0 \\ \tau_c > \mathcal{F}(\alpha_p, \omega) \end{cases} \tag{5.4.15}$$

where $\mathcal{F}(\alpha_p, \omega)$, with $\alpha_p, \omega \in \mathbb{R}$, is given as

$$\mathcal{F}(\alpha_p, \omega) = \frac{a(\alpha_p)\ \omega^2 + b(\alpha_p)\ \omega + c(\alpha_p)}{g(\alpha_p)} \tag{5.4.16}$$

with

$$
\begin{aligned}
a(\alpha_p) &= \rho|u|^2\left(2\tau_m + \alpha_p\tau_e(2h - |u|^2)\right)^2 \\
b(\alpha_p) &= 2\rho|u|^2\left(2\tau_m + \alpha_p\tau_e(2h - |u|^2)\right)\left(\tau_m + \alpha_p\tau_e(2h - |u|^2\alpha_p T)\right) \\
c(\alpha_p) &= \rho|u|^2\tau_e^2\alpha_p^2(2h - |u|^2)^2 \\
g(\alpha_p) &= 4\left(\tau_m + |u|^2\tau_e\alpha_p(1 - \alpha_p T)\right).
\end{aligned}
$$

Proof:
Assume that $\tilde{\tau}$ is positive definite. Then, all eigenvalues of $\tilde{\tau}$ are real and positive. Since τ_m is an eigenvalue of $\tilde{\tau}$, it follows that $\tau_m > 0$. From the positive definiteness of $\tilde{\tau}$ it follows that all it's principal submatrices are also positive definite, therefore $\tau_e > 0$. Moreover, all minor principals of $\tilde{\tau}$ are positive definite. Since $\tau_e > 0$ and $\tau_m > 0$, this implies the following system of five inequalities, which is obtained using MAPLE:

$$
\tau_c > f_i(\alpha_p, \omega), \quad \forall i \in \{1, \ldots, 5\}, \tag{5.4.17}
$$

where f_i are given functions of α_p and ω. The explicit form of $f_i(\alpha_p, \omega)$, for $i = 1, \ldots, 4$, can be obtained as:

$$
\begin{aligned}
f_1(\alpha_p, \omega) &= F(0), & f_2(\alpha_p, \omega) &= F(u_1^2), \\
f_3(\alpha_p, \omega) &= F(u_1^2 + u_2^2), & f_4(\alpha_p, \omega) &= F(|u|^2),
\end{aligned}
$$

using the following functional

$$
F(X) = \frac{\rho\left[\tau_e e^2(-\tau_m + \alpha_p\tau_e X) - 2e_1\left[\tau_e X(\alpha_p T|u|^2 - 2h) - \tau_m T|u|^2\right]\omega + e_1^2 TX\omega^2\right]}{4[\tau_m T + \tau_e X(1 - \alpha_p T)]}
$$

with $0 \leq X \leq |u|^2$, $e = |u|^2 - 2h$ and $e_1 = 2\tau_m - \alpha_p\tau_e e$. Note that the denominator of $F(X)$ is vanishing when

$$
\alpha_p = \frac{1}{T} + \frac{\tau_m}{\tau_e X} > \frac{1}{T},
$$

where in the last inequality we used that $\tau_m > 0$ and $\tau_e > 0$. This is, however, not possible when Assumption 5.4.1 applies because then $\alpha_p \leq 1/T$. Hence, $F(X)$ is well defined for all $\alpha_p \leq 1/T$.

Furthermore, we can write f_5 in the following form

$$
f_5(\alpha_p, \omega) = \frac{a(\alpha_p)\ \omega^2 + b(\alpha_p)\ \omega + c(\alpha_p)}{g(\alpha_p)} \tag{5.4.18}
$$

with

$$a(\alpha_p) = \rho|u|^2 \left(2\tau_m + \alpha_p\tau_e(2h - |u|^2)\right)^2 \geq 0$$
$$b(\alpha_p) = 2\rho|u|^2 \left(2\tau_m + \alpha_p\tau_e(2h - |u|^2)\right)\left(\tau_m + \alpha_p\tau_e(2h - |u|^2\alpha_p T)\right)$$
$$c(\alpha_p) = \rho|u|^2\tau_e^2\alpha_p^2(2h - |u|^2)^2 \geq 0$$
$$g(\alpha_p) = 4\left(\tau_m + |u|^2\tau_e\alpha_p(1 - \alpha_p T)\right).$$

Note that,

$$g(\alpha_p) = 0 \Longleftrightarrow \alpha_p^{p,m} = \frac{|u|\tau_e \pm \sqrt{|u|^2\tau_e^2 + 4\tau_m\tau_e T}}{2|u|\tau_e T},$$

where the superscripts p and m refer to the roots taken with $+$ and $-$ sign, respectively. Since $\tau_m > 0$ and $\tau_e > 0$, it follows that $\alpha_p^m < 0$ and $\alpha_p^p > 1/T$. Therefore, using Assumption 5.4.1 it follows that $g(\alpha_p) > 0$ for $\alpha_p^m < \alpha_p \leq 1/T$ and also f_5 is well defined in this range.

Next, we show now that the system of inequalities (5.4.17) is simultaneously satisfied if and only if the following inequality is satisfied

$$\tau_c > f_5(\alpha_p, \omega), \tag{5.4.19}$$

which results from the condition that the determinant of $\tilde{\tau}$ must be positive. If (5.4.17) holds, then (5.4.19) is obviously true. Reversely, assuming that (5.4.19) is valid, it is sufficient to show that all the following functions are positive:

$$g_i(\alpha_p, \omega) = f_{i+1}(\alpha_p, \omega) - f_i(\alpha_p, \omega), \quad i = 1, \ldots, 4, \tag{5.4.20}$$

since this implies

$$\tau_c > f_5 > \cdots > f_1.$$

Using the functional form of the functions f_i, we can write g_i as:

$$g_4(\alpha_p, \omega) = \frac{\rho\tau_e\ h_4^2(\alpha_p, \omega)}{4\left[\tau_m + \tau_e|u|^2\alpha_p(1 - \alpha_p T)\right]\left[\tau_m T + \tau_e|u|^2(1 - \alpha_p T)\right]}$$

$$g_3(\alpha_p, \omega) = \frac{\rho\tau_m u_3^2\ h_3^2(\alpha_p, \omega)}{4\left[\tau_m T + \tau_e|u|^2(1 - \alpha_p T)\right]\left[\tau_m T + \tau_e(u_1^2 + u_2^2)(1 - \alpha_p T)\right]}$$

$$g_2(\alpha_p, \omega) = \frac{\rho\tau_m u_2^2\ h_2^2(\alpha_p, \omega)}{4\left[\tau_m T + \tau_e u_1^2(1 - \alpha_p T)\right]\left[\tau_m T + \tau_e(u_1^2 + u_2^2)(1 - \alpha_p T)\right]}$$

$$g_1(\alpha_p, \omega) = \frac{\rho u_1^2\ h_1^2(\alpha_p, \omega)}{4T\left[\tau_m T + \tau_e u_1^2(1 - \alpha_p T)\right]}$$

where h_i, for $i = 1, \ldots, 4$ are given functions of α_p and ω. More specifically,

$$h_3(\alpha_p, \omega) = h_2(\alpha_p, \omega) = h_1(\alpha_p, \omega) = \tau_e(2h - |u|^2) + T\left(2\tau_m + \alpha_p\tau_e(2h - |u|^2)\right)\omega,$$

and

$$h_4(\alpha_p,\omega) = \tau_m(2h-|u|^2) - (1-\alpha_p T)|u|^2\left(2\tau_m + \alpha_p\tau_e(2h-|u|^2)\right)\omega.$$

Since $\tau_m > 0$ and $\tau_e > 0$, using Assumption 5.4.1 we can conclude that $g_i > 0$ for all $i = 1,\dots,4$, which implies that the system of inequalities (5.4.17) reduces to condition (5.4.19). This completes the proof of this theorem since $\mathcal{F}(\alpha_p,\omega) = f_5(\alpha_p,\omega)$ in (5.4.15).

The proof of the reverse statement of this lemma is straightforward, since (5.4.15) implies that the inequalities in (5.4.17) are valid, i.e., all minor principals of $\tilde{\tau}$ are positive definite, which is a sufficient condition for the positive definiteness of $\tilde{\tau}$.

□

Remark 5.4.1 *Consider the conditions for the positive definiteness of the stabilization matrix $\tilde{\tau}$ given in Theorem 5.4.3. In (5.4.16), $a(\alpha_p) = 0$ if and only if*

$$\alpha_p = \frac{2\tau_m}{(|u|^2-2h)\tau_e}.$$

Then, $\mathcal{F}(\alpha_p,\omega)$ is independent of α_p and ω and can be written as

$$\mathcal{F}(\alpha_p,\omega) = \frac{\rho|u|^2\left(2h-|u|^2\right)^2\tau_m\tau_e}{4\tau_e(h-|u|^2)^2 - |u|^2\left(|u|^2\tau_e + 4\tau_m T\right)}.$$

The class of stabilization matrices for the entropy variables obtained in Theorem 5.4.2 is dimensionally consistent and positive definite under the conditions of Theorem 5.4.3. Let us specify the coefficients τ_c, τ_m and τ_e as in Definition 4.4.1. The next lemma provides sufficient conditions such that the stabilization matrix $\tilde{\tau}$ satisfies the requirements of Theorem 5.4.3.

Lemma 5.4.1 *Under the condition stated in Assumption 5.4.1, using the definition of τ_c, τ_m and τ_e, given in Definition 4.4.1, the condition*

$$\tau_c > \mathcal{F}(\alpha_p,\omega), \tag{5.4.21}$$

with $\mathcal{F}$ given in Theorem 5.4.3, is equivalent to

$$f(\alpha_p,\omega) = \frac{a(\alpha_p)\ \omega^2 + b(\alpha_p)\ \omega + c(\alpha_p)}{g(\alpha_p)} < 0, \tag{5.4.22}$$

where

$$\begin{aligned}
a(\alpha_p) &= \left(2c_v + \alpha_p(2h-|u|^2)\right)^2\\
b(\alpha_p) &= 2\left(2c_v + \alpha_p(2h-|u|^2)\right)\left(c_v + \alpha_p(2h-|u|^2\alpha_p T)\right)\\
c(\alpha_p) &= -4c_v(c_v + \alpha_p|u|^2) + \left(4Tc_v|u|^2 + (2h-|u|^2)^2\right)\alpha_p^2\\
g(\alpha_p) &= 4c_v(c_v + |u|^2\alpha_p(1-\alpha_p T)).
\end{aligned}$$

Moreover, there exists an interval $(\alpha_p^*, 1/T]$ *of values of* α_p*, with* $\alpha_p^* \leq 0$*, such that* $\forall \alpha_p \in (\alpha_p^*, 1/T]$ *there is a non-empty interval for* ω *called* $I_\omega^{\alpha_p}$*, such that (5.4.22) is satisfied* $\forall \omega \in I_\omega^{\alpha_p}$.

Proof:
Inserting the definition of the stability parameters τ_c, τ_m and τ_e, given in Definition 4.4.1 into (5.4.21), it is straightforward to obtain inequality (5.4.22). Furthermore, the denominator $g(\alpha_p)$ of $f(\alpha_p, \omega)$ in (5.4.22) is vanishing if and only if

$$\alpha_p = \alpha_p^{p,m} = \frac{|u| \pm \sqrt{|u|^2 + 4c_v T}}{2|u|T}.$$

Since $\alpha_p^p > 1/T$ and $\alpha_p^m < 0$, it follows that $f(\alpha_p, \omega)$ is well defined for all $\alpha_p \leq 1/T$, except $\alpha_p \neq \alpha_p^m$. We consider three distinct cases.

Case 1. Assume that $\dfrac{a(\alpha_p)}{g(\alpha_p)} > 0$.

This case is only possible when $g(\alpha_p) > 0$, that is when $\alpha_p \in (\alpha_p^m, 1/T]$. The second statement of this lemma is now proven in two steps.

Step 1. First we give necessary and sufficient conditions for which inequality (5.4.22) is satisfied in a range $\omega \in I_\omega^{\alpha_p} \neq \emptyset$, for any $\alpha_p^m < \alpha_p \leq 1/T$. Let us fix $\alpha_p^m < \alpha_p \leq 1/T$ and consider $f(\alpha_p, \omega)$ in (5.4.22) only as a function of ω. Our aim is to obtain for any fixed $\alpha_p^m < \alpha_p \leq 1/T$ an interval $I_\omega^{\alpha_p} \neq \emptyset$ such that $f(\alpha_p, \omega) < 0$. Since $f(\alpha_p, \omega)$ is a second order polynomial in ω with the coefficient $a(\alpha_p) > 0$, it follows that

$$\frac{\partial f(\alpha_p, \omega)}{\partial \omega} = \frac{2a(\alpha_p)\ \omega + b(\alpha_p)}{g(\alpha_p)} = 0 \Longleftrightarrow \omega = \omega_{min} = -\frac{b(\alpha_p)}{2a(\alpha_p)}.$$

A necessary and sufficient condition for $I_\omega^{\alpha_p} \neq \emptyset$ is that

$$f(\alpha_p, \omega_{min}) = \frac{|u|^2 T\alpha_p^2 + (|u|^2 - 4h)\alpha_p - 5c_v}{4c_v} < 0. \tag{5.4.23}$$

Step 2. Next, let us vary α_p, in (5.4.23) and verify for what range of α_p the inequality is satisfied. We have now $f(\alpha_p, \omega_{min})$ a second order polynomial in α_p and

$$\frac{\partial f(\alpha_p, \omega_{min})}{\partial \alpha_p} = \frac{2|u|^2 T\alpha_p + |u|^2 - 4h}{4c_v} = 0 \Longleftrightarrow \alpha_p = \alpha_{p_{min}} = \frac{-|u|^2 + 4h}{2|u|^2 T}.$$

Since

$$f(\alpha_{p_{min}}, \omega_{min}) = -\frac{(-|u|^2 + 4h)^2 + 20c_v|u|^2 T}{16c_v|u|^2 T} < 0,$$

there is an interval (α_p^-, α_p^+), with

$$\alpha_p^\pm = \frac{-|u|^2 + 4h \pm \sqrt{(-|u|^2 + 4h)^2 + 20c_v|u|^2 T}}{2|u|^2 T},$$

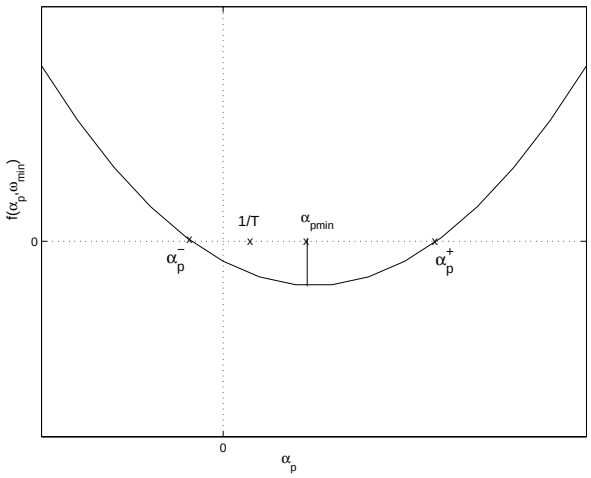

Figure 5.2: Illustration of $f(\alpha_p, \omega_{min})$ as a function of α_p.

such that $f(\alpha_p, \omega_{min}) < 0$ for all $\alpha_p \in (\alpha_p^-, \alpha_p^+)$. Since we need to satisfy the assumption $\alpha_p^m < \alpha_p \leq 1/T$, we have to further investigate the interval we have just found to satisfy (5.4.23). In Lemma 5.4.2 we prove that for sufficiently small α_p the following is valid

$$\alpha_{p_{min}} > \frac{1}{T}. \tag{5.4.24}$$

It is straightforward to see that $\alpha_p^- < 0$ and using (5.4.24), we have

$$\frac{1}{T} < \alpha_{p_{min}} < \alpha_p^+,$$

as illustrated in Figure 5.2. Consequently, we proved that

$$f(\alpha_p, \omega_{min}) < 0, \quad \forall \alpha_p \in (\alpha_p^-, 1/T].$$

Moreover,

$$f(\alpha_p, \omega_{min}) < 0, \quad \forall \alpha_p \in (\alpha_p^*, 1/T],$$

where $\alpha_p^* = \max\{\alpha_p^-, \alpha_p^m\} < 0$. Hence, we obtained the following result

$$\exists\ I_\omega^{\alpha_p} \neq \emptyset \quad \text{such that} \quad f(\alpha_p, \omega) < 0, \quad \forall \omega \in I_\omega^{\alpha_p} \quad \text{and} \quad \forall \alpha_p \in (\alpha_p^*, 1/T].$$

In Figure 5.3 we illustrated the function $f(\alpha_p, \omega)$ as a function of ω for given values of α_p, as marked on the plot.

Case 2. Assume that $\dfrac{a(\alpha_p)}{g(\alpha_p)} < 0$.

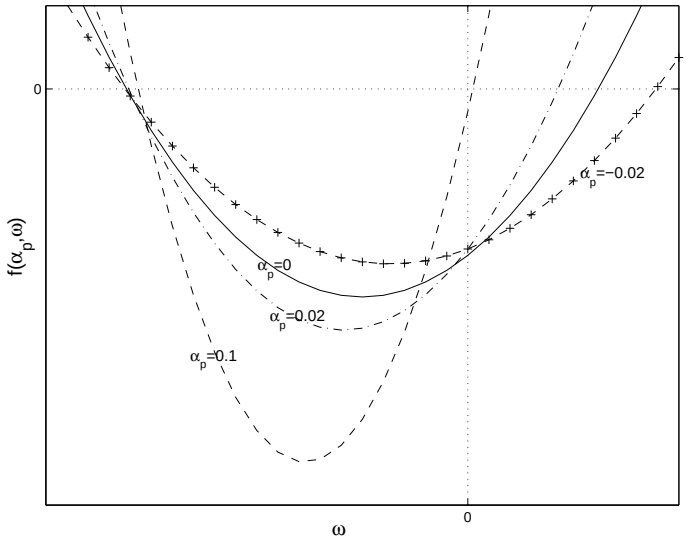

Figure 5.3: Illustration of $f(\alpha_p, \omega)$ as a function of ω for given values of α_p.

This case is only possible when $g(\alpha_p) < 0$. Since $\alpha_p^p > 1/T$, this condition is satisfied for $\alpha_p \in (-\infty, \alpha_p^m)$. This case is not interesting for us since it does not contain the incompressible limit, that is $\alpha_p = 0$.

Case 3. Consider the case when $a(\alpha_p) = 0$, which is equivalent to

$$\alpha_p = \alpha_p^0 = \frac{2c_v}{|u|^2 - 2h}. \tag{5.4.25}$$

In the next lemma we will show that for weakly compressible flows $\alpha_p^0 > 1/T$, therefore, this case we can omit from our analysis. □

Lemma 5.4.2 *For weakly compressible flows, or equivalently, for sufficiently small Eckert number, the followings hold*

$$\alpha_{p_{min}} > \frac{1}{T}, \tag{5.4.26}$$

$$\alpha_p^0 > \frac{1}{T}, \tag{5.4.27}$$

where $\alpha_{p_{min}}$ is the minimum of the function $f(\alpha_p, \omega_{min})$ in (5.4.23) and α_p^0 is defined in (5.4.25).

Proof:
Inserting the value of $\alpha_{p_{min}}$ into (5.4.26) we obtain

$$\frac{|u|^2}{h} < \frac{4}{3}. \tag{5.4.28}$$

The left hand side of this inequality is a dimensionless quantity and it is related to the reference values $\{\rho_r, |u|_r, L, \Delta T\}$ introduced in Section 3.5 in the following way:

$$\frac{|u^*|^2}{h^*}\frac{|u|_r^2}{c_{p_r}\Delta T} = \frac{|u^*|^2}{h^*}Ec < \frac{4}{3},$$

where the dimensionless variables are denoted by a star, c_{p_r} and the reference temperature difference $\Delta T = T_w - T_\infty$ are also defined in Section 3.5. For compressible flows, there is a close relation between the Mach number M and the Eckert number, that is

$$Ec = (\gamma - 1)M^2\frac{T_\infty}{\Delta T}.$$

In the context of the compressibility constraints discussed in this thesis, the Eckert number can be expressed as

$$Ec = (\gamma - 1)\rho_r\ |u|_r^2\ \beta_T\frac{T_\infty}{\Delta T}. \tag{5.4.29}$$

In the incompressible limit, $\beta_T \to 0$, the Eckert number is also decreasing, therefore, for weakly compressible flows (5.4.28) is satisfied.

Next we will show (5.4.27) indirectly. Assume that $\alpha_p^0 < 1/T$, which is leads to

$$2 < \frac{|u|^2}{c_v T + h} < \frac{|u|^2}{h} = \frac{|u^*|^2}{h^*}Ec. \tag{5.4.30}$$

Using the weakly compressible assumption and (5.4.29), the right hand side of (5.4.30) is decreasing as $\beta_T \to 0$, therefore, the inequality (5.4.30) will not hold, which leads to a contradiction. □

Remark 5.4.2 *Let us recall that in Lemma 4.4.1 of Chapter 4, we obtained that for* $\alpha_p = 0$

$$f(0,\omega) < 0 \quad \textit{if and only if} \quad \omega \in I_\omega^0 = \left(\frac{-1-\sqrt{5}}{2}, \frac{-1+\sqrt{5}}{2}\right).$$

Since f *in (5.4.22) is a second order polynomial in* ω*, we can write it as*

$$f(\alpha_p,\omega) = \frac{\left(2c_v + \alpha_p(2h - |u|^2)\right)^2}{4c_v(c_v + |u|^2\alpha_p(1-\alpha_p T))}(\omega - \omega^+(\alpha_p))(\omega - \omega^-(\alpha_p))$$

where ω^+ *and* ω^- *are the roots of this polynomial. When* α_p *is sufficiently small, we can write the following expansion of the roots*

$$\omega^\pm(\alpha_p) = \frac{-1 \pm \sqrt{5}}{2} \pm \frac{\sqrt{5}}{20c_v}\left((9 \mp \sqrt{5})|u|^2 - (6 \pm 2\sqrt{5})h\right)\alpha_p + O(\alpha_p^2). \tag{5.4.31}$$

5.5 Entropy stability

In this section we will show that the Galerkin least-squares finite element formulation of the compressible Navier-Stokes equations satisfies a global entropy stability. We will demonstrate that a necessary and sufficient condition for stability is the positive definiteness of the stabilization matrix $\tilde{\tau}$. In the previous section we constructed a class of stabilization operators and gave necessary and sufficient conditions for positive definiteness but restricted ourself only to weakly compressible flows. The stability analysis described in this section is, however, valid for general compressible flows if $\tilde{\tau}$ is positive definite.

Consider the compressible Navier-Stokes equations in conservative form

$$U_{,t} + F^a_{i,i} = F^d_{i,i} \ , \text{ for } i = 1, 2, 3, \tag{5.5.1}$$

where $U \in \mathbb{R}^5$ is the vector of conservative variables, and $F^a_i, F^d_i \in \mathbb{R}^5$ are, respectively, the advective and diffusive fluxes in the ith Cartesian coordinate direction, which are defined in (3.1.8). As discussed in Section 3.3.1, we need to complete (5.5.1) with the equations of state which we choose to be for an ideal gas

$$e = c_v T \quad \text{and} \quad p = (\gamma - 1)\rho e, \tag{5.5.2}$$

where $\gamma = c_p/c_v$. For smooth solutions, we can write the system (5.5.1) in the form

$$U_{,t} + A_i(U)U_{,i} = (K_{ij}(U)U_{,j})_{,i}, \tag{5.5.3}$$

where $A_i(U) = F_{i,U}$ and $K_{ij}(U)U_{,j} = F^d_i$. The flux Jacobian matrices for conservative variables are given for instance in [23].

5.5.1 The entropy function

In this section we recall the concept of an entropy function, the properties of entropy variables and some important theorems relating these two concepts.

The following definition shows the important link between hyperbolic systems and the concept of entropy.

Definition 5.5.1 *A scalar valued function $H = H(U)$ is called generalized entropy function for the system*

$$U_{,t} + A_i(U)U_{,i} = 0 \tag{5.5.4}$$

if the following two conditions are satisfied:

(1) H is a convex function.

(2) *There exists scalar-valued functions* $\sigma_i = \sigma_i(U),\ i = 1, 2, 3$, *called entropy fluxes such that*

$$H_{,U} A_i(U) = \sigma_{i,U}(U). \tag{5.5.5}$$

The following theorems, show the relationship between symmetric hyperbolic systems and generalized entropy functions.

Theorem 5.5.1 *(Godunov, [21], [30]) If a hyperbolic system can be symmetrized by introducing a change of variables, then a generalized entropy function and corresponding entropy fluxes exist for this system.*

Theorem 5.5.2 *(Mock, [21], [30]) A hyperbolic system of conservation laws possessing a generalized entropy function* H*, becomes symmetric hyperbolic under the change of variables*

$$V^T = \frac{\partial H}{\partial U}. \tag{5.5.6}$$

For the proof of the above theorems we refer to [21].

Theorem 5.5.3 *(Harten, [21],[30]) A family of generalized entropy functions for the Euler equations (5.5.4) is given by*

$$H(U) = -\rho g(s) \quad \textit{such that} \quad g' > 0,\ \frac{g''}{g'} < \frac{1}{\gamma}, \tag{5.5.7}$$

where s *is the thermodynamic entropy.*

Note that the last inequality in (5.5.7) ensures satisfaction of the convexity condition. The convexity of H is equivalent to the positive-definiteness of $\tilde{A}_0$, which can be seen from the relation

$$\frac{\partial^2 H}{\partial U^2} = \frac{\partial V}{\partial U} = \tilde{A}_0^{-1}.$$

The corresponding entropy fluxes are given by

$$\sigma_i = u_i H(U), \quad i = 1, 2, 3, \quad \text{where } u = (u_1, u_2, u_3)^T. \tag{5.5.8}$$

Given a general equation of state, or equivalently, the fundamental equation s, then using the entropy function H in (5.5.7) a new set of variables can be obtained by (5.5.6). As an example, consider the entropy function H in (5.5.7) in combination with the ideal gas equations of state, or equivalently, the fundamental equation for the entropy $s = \ln(p/\rho^\gamma) + s_0$. Then, a new set of variables is defined by (5.5.6) as

$$V = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \end{pmatrix} = \frac{g'}{e} \begin{pmatrix} e(\gamma - g/g') - \frac{1}{2}|u|^2 \\ u_1 \\ u_2 \\ u_3 \\ -1 \end{pmatrix}.$$

In [30] it is shown that symmetrization of the Navier-Stokes equations with heat conduction places additional requirements on the definition of the family of generalized entropy functions in (5.5.7). Since this is crucial for the stability analysis of the Navier-Stokes equations, we briefly illustrate this result. Note that in terms of the symmetrizing variables, the diffusive flux tensor $\tilde{K} = (\tilde{K}_{ij})$, $i, j = 1, 2, 3$, is symmetric and positive-semidefinite. The diffusive fluxes F_i^d are composed of the viscous fluxes F_i^v and the heat flux F_i^h. Since

$$\tilde{K}_{ij}^h V_{,j} = F_i^h = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \kappa T_{,i} \end{pmatrix}, \quad i = 1, 2, 3,$$

the only way the symmetry of $\tilde{K}$ can be maintained is if T depends only on V_5. Using the change of variables

$$T(V) = -\frac{g'}{c_v V_5},$$

this implies that g' is constant, therefore, g is an affine function of s. The conclusion is that for the Navier-Stokes equations, g must be an affine function of s.

Next, we give some examples of global entropy functions frequently used in the related literature.

Remark 5.5.1 *One simple case, considered in [50] is $H = -\rho(s - s_0)$, or the same function with the additive constant $s_0 = 0$ is discussed in [30]. An other example given in [3] assumes that $H = -\rho s/(\gamma - 1)$. The symmetrizing variables can then be obtained using (5.5.6), and the mappings $U \mapsto V$ and $V \mapsto U$ can be given.*

In the following lemma we summarize the properties of the entropy variables V, defined in Theorem 5.5.2.

Lemma 5.5.1 *Defining $H = -\rho g(s)$ where g is an affine function of s, the following identities are valid*

$$V \cdot U_{,t} = H_{,t}$$

$$V \cdot F_{i,i}(U) = (u_i H)_{,i}$$

$$V \cdot F_i^d = \frac{q_i}{c_v T}.$$

Proof:
These results follow directly from the definition of entropy variables and fluxes. In the last identity we used

$$V^T F_i^v = 0, \quad V^T F_i^h = \frac{q_i}{c_v T}. \tag{5.5.9}$$

□

Remark 5.5.2 *Given the entropy function $H = -\rho g(s)$, where g is an affine function of s, the present definition of the entropy variables enables us to derive the following important result*

$$\begin{aligned} 0 &= V^T \left(\tilde{A}_0 V_{,t} + \tilde{A}_i V_{,i} - (\tilde{K}_{ij} V_{,j})_{,i} \right) \\ &= H_{,t} + (u_i H)_{,i} + V_{,i}^T \tilde{K}_{ij} V_{,j} - (V^T \tilde{K}_{ij} V_{,j})_{,i} \\ &= -(\rho s)_{,t} - (\rho s u_i)_{,i} + V_{,i}^T \tilde{K}_{ij} V_{,j} - \left(\frac{q_i}{c_v T} \right)_{,i}, \end{aligned} \tag{5.5.10}$$

where we used the identities of Lemma 5.5.1.

Remark 5.5.3 *Equation (5.5.10) implies the entropy production inequality:*

$$(\rho s)_{,t} + (\rho s u_i)_{,i} + \left(\frac{q_i}{c_v T} \right)_{,i} = V_{,i}^T \tilde{K}_{ij} V_{,j} \geq 0. \tag{5.5.11}$$

The entropy production inequality (5.5.11) is of great importance in the finite element discretization of the Navier-Stokes equations, as we will see later in this section.

Remark 5.5.4 *Consider the variational formulation:*

Find $V \in \mathcal{S}_V$, such that for all $W \in \mathcal{W}_V$, the following is valid

$$0 = \int_{\mathcal{E}} W^T \left(\tilde{A}_0 V_{,t} + \tilde{A}_i V_{,i} - (\tilde{K}_{ij} V_{,j})_{,i} \right) \, \mathrm{d}\mathcal{E}. \tag{5.5.12}$$

Setting $W = V$, and using (5.5.11) we obtain that the discrete solution always satisfies the Clausius-Duhem inequality:

$$\int_{\mathcal{E}} \left((\rho s)_{,t} + (\rho s u_i)_{,i} + \left(\frac{q_i}{c_v T} \right)_{,i} \right) \mathrm{d}\mathcal{E} = \int_{\mathcal{E}} V_{,i}^T \tilde{K}_{ij} V_{,j} \mathrm{d}\mathcal{E} \geq 0. \tag{5.5.13}$$

It is well known that the Galerkin method without stabilization operator is not effective for non-smooth solutions and produces spurious solutions when the mesh under resolves the solution, in particular around discontinuities. We will discuss, therefore, the nonlinear stability of the Galerkin least-squares method in the next section in which the property (5.5.11) is a crucial component.

5.5.2 Nonlinear stability analysis

Recall the Galerkin least-squares variational formulation of the symmetrized compressible Navier-Stokes equations. Consider the trial and test function spaces defined in Section 5.1. For our nonlinear stability analysis, we use the following non-integrated-by-parts form of the weak formulation:

Find a $V \in \mathcal{S}_V^n$ such that for all $W \in \mathcal{W}_V^n$, the following relation is satisfied

$$\sum_{n=0}^{N} \{B_{Gal}(V,W) + B_{ls}(V,W) + B_{jump}(V,W)\} = 0, \tag{5.5.14}$$

where the first term in (5.5.14) is the Galerkin term,

$$B_{Gal}(V,W) = \int_{\mathcal{E}_n} W \cdot \left[U(V)_{,0} + F_i^a(U(V))_{,i} - F_i^d(U(V))_{,i}\right] \mathrm{d}\mathcal{E} \quad \text{for } i = 1,2,3,$$

the second term is the least-squares stabilization operator, defined as:

$$B_{ls}(V,W) = \sum_{e=1}^{(n_{el})_n} \int_{\mathcal{E}_n^e} (\mathcal{L}_V W) \cdot \tilde{\tau} (\mathcal{L}_V V) \mathrm{d}\mathcal{E},$$

with $\mathcal{L}_V$ the symmetrized Navier-Stokes operator:

$$\mathcal{L}_V = \tilde{A}_\ell \frac{\partial}{\partial x_\ell} - \frac{\partial}{\partial x_i}(\tilde{K}_{ij} \frac{\partial}{\partial x_j}) \quad \text{for } \ell = 0, \ldots, 3,\ i,j = 1,2,3$$

and the last term in (5.5.14) is the so-called jump term, defined as:

$$B_{jump}(V,W) = \int_{\Omega(t_n)} W(t_n^+) \cdot \left[(U(V(t_n^+)) - U(V(t_n^-))\right] \mathrm{d}\Omega$$

First, we need to state the following lemma, which was also used in [3].

Lemma 5.5.2 *(Time-discontinuous entropy production) The following entropy function jump identity holds across time slab boundaries*

$$\int_{\Omega(t_n)} \left([H]_{t_n^-}^{t_n^+} - V^T(t_n^+)\ [U]_{t_n^-}^{t_n^+} \right) \mathrm{d}\Omega + \left| \left| \left| [U]_{t_n^-}^{t_n^+} \right| \right| \right|^2_{\tilde{A}_0^{-1}, \Omega(t_n)} = 0 \tag{5.5.15}$$

where

$$\left| \left| \left| [U]_{t_n^-}^{t_n^+} \right| \right| \right|^2_{\tilde{A}_0^{-1}, \Omega(t_n)} = \int_{\Omega(t_n)} \int_0^1 (1-\theta) [U]_{t_n^-}^{t_n^+} \cdot \tilde{A}_0^{-1}(\bar{U}(\theta))\ [U]_{t_n^-}^{t_n^+}\ \mathrm{d}\theta\ \mathrm{d}\Omega \tag{5.5.16}$$

and $\bar{U}(\theta) = U(t_n^+) - \theta\ [U]_{t_n^-}^{t_n^+}$, *with* $[U]_{t_n^-}^{t_n^+} = U(t_n^+) - U(t_n^-)$.

Proof:
First note that the definition of the norm in (5.5.16) is valid since $\tilde{A}_0^{-1}$ is symmetric positive definite. Let us recall the Taylor formula for a function f with integral remainder. Assume that the function $f(z) \in C^2$ in some interval (a_1, b_1). Then for each a and $z \in (a_1, b_1)$ the following formula is true

$$f(z) = f(a) + f'(a)(z-a) + \int_a^z (z-v) f''(v)\ \mathrm{d}\, v. \tag{5.5.17}$$

Applying the Taylor formula for the scalar-value function $f = H(U)$ with

$$z = U(t_n^-), \quad a = U(t_n^+),$$

we obtain

$$H(U(t_n^-)) = H(U(t_n^+)) - \frac{\partial H}{\partial U}(U(t_n^+))[U]_{t_n^-}^{t_n^+} + \int_{U(t_n^+)}^{U(t_n^-)} \big(U(t_n^-) - v\big) \cdot \frac{\partial^2 H}{\partial U^2}(v)\ \mathrm{d}\, v. \tag{5.5.18}$$

Introducing the variable change $v = U(t_n^+) - \theta[U]_{t_n^-}^{t_n^+}$ and the identities $\frac{\partial H}{\partial U} = V^T$, $\frac{\partial^2 H}{\partial U^2} = \tilde{A}_0^{-1}$ into (5.5.18) and then integrate over $\Omega(t_n)$, we obtain (5.5.15). □

Theorem 5.5.4 *A global entropy stability of the Galerkin least-squares method for compressible flows in a time dependent flow domain is ensured with the use of entropy variables and the stabilization operator $\tilde{\tau}$ given in Theorem 5.4.2. The energy balance in given by*

$$\int_{\Omega(t_{N+1}^-)} H(t_{N+1}^-)\mathrm{d}\Omega + \sum_{n=0}^{N} \left(\|[U]_{t_n^-}^{t_n^+}\|^2_{\tilde{A}_0^{-1},\Omega(t_n)} + \sum_{e=1}^{(n_{el})_n} \|\mathcal{L}_V V\|^2_{\tilde{\tau},\mathcal{E}_n} + \|\nabla_{\bar{x}} V\|^2_{\tilde{K},\mathcal{E}_n} \right)$$

$$= \int_{\Omega(t_0^-)} H(t_0^-)\mathrm{d}\Omega + \sum_{n=0}^{N} \left(\int_{\mathcal{Q}_n} \left(-u_i H(U) + \frac{q_i}{c_v T} \right) \bar{n}_i\ \mathrm{d}\mathcal{Q} + \int_{\mathcal{Q}_n} \bar{n} \cdot v\, H(U(V))\mathrm{d}\mathcal{Q} \right) \tag{5.5.19}$$

Proof:
In order to construct the energy balance, set $W = V$ in (5.5.14). Consider first the advection part of the Galerkin term. Using the definition of the generalized entropy function and then applying Gauss theorem, we obtain

$$\int_{\mathcal{E}_n} V \cdot [U(V)_{,0} + F_i^a(U(V))_{,i}]\ \mathrm{d}\mathcal{E} = \int_{\mathcal{E}_n} H_{,U} \cdot [U(V)_{,0} + F_i^a(U(V))_{,i}]\ \mathrm{d}\mathcal{E} =$$

$$\int_{\mathcal{E}_n} (H_{,0} + \sigma_i(U(V))_{,i})\mathrm{d}\mathcal{E} = \int_{\partial\mathcal{E}_n} n_\ell\, \mathcal{F}_\ell\ \mathrm{d}(\partial\mathcal{E}),$$

where $n \in \mathbb{R}^4$ is the unit outward space-time normal vector at the boundary $\partial\mathcal{E}_n$ and $\mathcal{F} = (H, \sigma_1, \sigma_2, \sigma_3)^T$.

Next, we split the boundary integral into integrals over the different types of boundaries. The normal vectors at $\Omega(t_{n+1}^-)$ and $\Omega(t_n^+)$ are $(1,0,0,0)^T$ and $(-1,0,0,0)^T$, respectively. Hence, the boundary integral over the surfaces $\Omega(t_{n+1}^-)$ and $\Omega(t_n^+)$ is equal to

$$\int_{\Omega(t_{n+1})\cup\Omega(t_n)} n\cdot\mathcal{F}\,\mathrm{d}\Omega = \int_{\Omega(t_{n+1})} H(t_{n+1}^-)\mathrm{d}\Omega - \int_{\Omega(t_n)} H(t_n^+)\mathrm{d}\Omega.$$

By adding the jump term to the above integral and sum over all space-time slabs, we obtain

$$\begin{aligned}\mathcal{T}_1 &= \sum_{n=0}^{N}\left\{\int_{\Omega(t_{n+1})} H(t_{n+1}^-)\mathrm{d}\Omega - \int_{\Omega(t_n)} H(t_n^-)\mathrm{d}\Omega - \int_{\Omega(t_n)}\left([H]_{t_n^-}^{t_n^+} - V^T(t_n^+)[U]_{t_n^-}^{t_n^+}\right)\,\mathrm{d}\Omega\right\}\\ &= \int_{\Omega(t_{N+1})} H(t_{N+1}^-)\mathrm{d}\Omega - \int_{\Omega(t_0)} H(t_0^-)\mathrm{d}\Omega + \sum_{n=0}^{N}\left|\left|\left|[U]_{t_n^-}^{t_n^+}\right|\right|\right|^2_{\tilde{A}_0^{-1},\Omega(t_n)}\end{aligned}$$

where in the last equality we used Lemma 5.5.2.

For the remaining parts of the boundary, let $\bar{n}^i(\bar{x},t)\in\mathbb{R}^3$, $1\le i\le 6$, (since we use hexahedral elements), be the spatial component of the normal vector at the boundary $\mathcal{Q}_n$. The space-time normal vector n^i, $(1\le i\le 6)$ at the boundary $\mathcal{Q}_n$ is equal to

$$n^i = (-v\cdot\bar{n}^i, \bar{n}^i)^T,$$

with the grid velocity $v\in\mathbb{R}^3$ given by the relation $\Delta\bar{x}/\Delta t$. Therefore, the boundary integral over the boundary $\mathcal{Q}_n$ is equal to

$$\begin{aligned}\mathcal{T}_2 = \int_{\mathcal{Q}_n} n\cdot\mathcal{F}\mathrm{d}\mathcal{Q} &= \int_{\mathcal{Q}_n}\left[\bar{n}_i\sigma_i(U(V)) - \bar{n}\cdot v\,H(U(V))\right]\mathrm{d}\mathcal{Q}\\ &= \int_{\mathcal{Q}_n}\left[u_i H(U)\bar{n}_i - \bar{n}\cdot v\,H(U(V))\right]\mathrm{d}\mathcal{Q},\quad i=1,2,3,\end{aligned}$$

where we used that $\sigma_i = u_i H(U)$.

Using the relation $F_i^d(U(V)) = \tilde{K}_{ij}(V)V_{,j}$, the diffusive part of the Galerkin term can be written as

$$\begin{aligned}\mathcal{T}_3 = \int_{\mathcal{E}_n} V\cdot F_i^d(U(V))_{,i}\,\mathrm{d}\mathcal{E} &= \int_{\mathcal{E}_n} V\cdot(\tilde{K}_{ij}(V)V_{,j})_{,i}\,\mathrm{d}\mathcal{E}\\ &= \int_{\mathcal{E}_n}\left[\left(V^T\tilde{K}_{ij}(V)V_{,j}\right)_{,i} - V_{,i}^T\tilde{K}_{ij}V_{,j}\right]\mathrm{d}\mathcal{E}\\ &= \int_{\mathcal{Q}_n}\left(V^T\tilde{K}_{ij}(V)V_{,j}\right)\bar{n}_i\,\mathrm{d}\mathcal{Q} - \int_{\mathcal{E}_n} V_{,i}^T\tilde{K}_{ij}V_{,j}\,\mathrm{d}\mathcal{E}\\ &= \int_{\mathcal{Q}_n} V^T F_i^d\bar{n}_i\,\mathrm{d}\mathcal{Q} - \int_{\mathcal{E}_n} V_{,i}^T\tilde{K}_{ij}V_{,j}\,\mathrm{d}\mathcal{E}\\ &= \int_{\mathcal{Q}_n}\frac{q_i}{c_v T}\bar{n}_i\,\mathrm{d}\mathcal{Q} - \|\nabla_{\bar{x}}V\|^2_{\tilde{K},\mathcal{E}_n}\end{aligned}$$

where in the last equality we used (5.5.9), $\tilde{K} = (\tilde{K}_{ij})$ and the operator $\nabla_{\bar{x}}$ is the gradient operator with respect to the physical space coordinates. Note that since $\tilde{K}$ is symmetric positive-semidefinite, the second term in the last equality above is only a semi-norm.

Using the positive definiteness of the stabilization matrix $\tilde{\tau}$, we obtain the following norm for the least-squares operator

$$\mathcal{T}_4 = \sum_{e=1}^{(n_{el})_n} \int_{\mathcal{E}_n^e} (\mathcal{L}_V V) \cdot \tilde{\tau} (\mathcal{L}_V V) \mathrm{d}\mathcal{E} = \|\mathcal{L}_V V\|^2_{\tilde{\tau}, \mathcal{E}_n^e}.$$

Finally, adding $\mathcal{T}_2$, $\mathcal{T}_3$, $\mathcal{T}_4$ and summing over all space-time slabs, together with $\mathcal{T}_1$ completes the proof of this theorem. □

Remark 5.5.5 *The above stability result shows that the global entropy of the system at the final time is bounded by the initial entropy state provided that the boundary entropy produced via the boundary integrals on the right hand side of (5.5.19) is negative. This condition can be used to derive a set of well-posed boundary conditions, and in [12], Dutt determined boundary conditions which lead to energy decaying systems.*

Remark 5.5.6 *Furthermore, (5.5.19) also shows that the individual terms, that is the jump term, the least-squares operator and the gradient of the entropy function are bounded.*

Chapter 6

Galerkin least-squares finite element formulation

In this chapter we define the finite element variational method for the solution of the incompressible Navier-Stokes equations. The basis of our formulation is a time-discontinuous Galerkin least-squares method. This method is well suited to deal with time dependent flow domains with moving boundaries and dynamic meshes. The time-discontinuous Galerkin method results in an implicit discretization using space-time elements which automatically is globally conservative on deforming meshes and prevents the inaccuracies of data interpolation between the various meshes, see for instance Masud and Hughes [40] and van der Vegt and van der Ven [55]. We discuss this approach for the symmetrized Navier-Stokes equations, but we did not yet implement all aspects of the mesh deformation discussed in this thesis in the corresponding computer program.

Consider the incompressible Navier-Stokes equations and the heat equation in a time-dependent flow domain $\Omega(t)$. Since the flow domain boundary is moving and deforming in time, we will not make a separation between the space and time variables and consider directly the space $\mathbb{R}^{d+1}$, where d is the number of space dimensions. Assume that $d = 3$. Let $\mathcal{E} \subset \mathbb{R}^4$ be an open, bounded space-time domain. A point $x \in \mathbb{R}^4$ has coordinates (x_0, x_1, x_2, x_3), with $x_0 = t$ representing time, but we will also use the notation $(t, \bar{x}) \in \mathbb{R}^4$, with $\bar{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ the position vector at time t. The flow domain $\Omega(t) \subset \mathcal{E}$ at time t is defined as: $\Omega(t) = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid (t, x_1, x_2, x_3) \in \mathcal{E}\}$. The space-time domain boundary $\partial\mathcal{E}$ consists of the hypersurfaces $\Omega(t_0) = \{x \in \partial\mathcal{E} \mid x_0 = t_0\}$, $\Omega(t_{N+1}) = \{x \in \partial\mathcal{E} \mid x_0 = t_{N+1}\}$, and $\mathcal{Q} = \{x \in \partial\mathcal{E} \mid t_0 < x_0 < t_{N+1}\}$.

Let $Y : \mathcal{E} \mapsto \mathbb{R}^5$ denote the vector of primitive variables $(p, u_1, u_2, u_3, T)^T$ and $\mathcal{F} : \mathbb{R}^5 \mapsto \mathbb{R}^{5\times 4}$ denote the flux tensor, with the flux vector in the ℓth coordinate direction

F_ℓ ($\ell = 0, \ldots, 3$) given by the columns of $\mathcal{F}$, i.e.,

$$\mathcal{F} = \begin{pmatrix} \rho & \rho u_1 & \rho u_2 & \rho u_3 \\ \rho u_1 & \rho u_1^2 + p & \rho u_1 u_2 & \rho u_1 u_3 \\ \rho u_2 & \rho u_1 u_2 & \rho u_2^2 + p & \rho u_2 u_3 \\ \rho u_3 & \rho u_1 u_3 & \rho u_2 u_3 & \rho u_3^2 + p \\ \rho T & \rho u_1 T & \rho u_2 T & \rho u_3 T \end{pmatrix},$$

where ρ denotes the density, u_i the velocity component in the ith Cartesian coordinate direction, p the pressure and T the temperature. Using these notations, the incompressible Navier-Stokes equations can be written in conservative form as

$$F_\ell(Y(x))_{,\ell} + (K_{ij}(Y)Y_{,j})_{,i} = 0, \quad x \in \mathcal{E},$$

where $K_{ij} \in \mathbb{R}^{5\times 5}$ for $i, j = 1, 2, 3$ denote the viscous flux Jacobian matrices and the summation convention is used on repeated indices.

6.1 Geometry of space-time elements

Consider the partition of the time interval $I = (t_0, t_{N+1})$ using the time levels $t_0 < t_1 < \ldots < t_{N+1}$ and denote by $I_n = (t_n, t_{n+1})$ the nth time interval. A space-time slab is defined as $\mathcal{E}_n = \mathcal{E} \cap I_n$. In each space-time slab $\mathcal{E}_n$ we define a partition $\mathcal{T}_h^n$ into $(n_{el})_n$ non-overlapping hexahedral elements $\mathcal{E}_n^e$. The space-time elements $\mathcal{E}_n^e$ are obtained by splitting the spatial domain $\Omega(t_n)$ into a set of non-overlapping elements Ω_n^e and connecting them with the mapping φ_t^n to the elements $\Omega_{n+1}^e \subset \Omega(t_{n+1})$ at time t_{n+1}. At each time level t_n we use hexahedral elements to define the triangulation. The evolution of the flow domain during the time interval I_n is represented by the mapping:

$$\begin{aligned} \varphi_t^n : \Omega(t_n) &\to \Omega(t) \\ \bar{x} &\mapsto \varphi_t^n(\bar{x}), \quad t \in I_n. \end{aligned}$$

The mapping φ_t^n is assumed to be sufficiently smooth, orientation preserving and invertible. Each element Ω_n^e is related to the master element $\hat{\Omega} = [0, 1]^3$ through the mapping:

$$\begin{aligned} F_n^e : \hat{\Omega} &\to \Omega_n^e \\ \bar{\xi} &\mapsto \bar{x} = \sum_{i=1}^{8} x_i(\Omega_n^e)\chi_i(\bar{\xi}) \end{aligned}$$

where $x_i(\Omega_n^e) \in \mathbb{R}^3$, $1 \le i \le 8$, are the spatial coordinates of the vertices of the hexahedron Ω_n^e and $\chi_i(\bar{\xi})$ the tri-linear finite element shape functions for hexahedra,

with $\bar{\xi} = (\xi_1, \xi_2, \xi_3) \in \hat{\Omega}$. The elements Ω^e_{n+1} are obtained by moving the vertices of each hexahedron Ω^e_n with the mapping φ^n_t to their new position at time $t = t_{n+1}$. Therefore, we can define the mapping:

$$\begin{aligned} F^e_{n+1} : \hat{\Omega} &\to \Omega^e_{n+1} \\ \bar{\xi} &\mapsto \bar{x} = \sum_{i=1}^{8} \varphi^n_{t_{n+1}}(x_i(\Omega^e_n))\chi_i(\bar{\xi}). \end{aligned}$$

The space-time elements are now obtained by connecting the elements in $\Omega(t_n)$ and $\Omega(t_{n+1})$ by linear interpolation in time. This results in the following parametrization of the space-time elements $\mathcal{E}^e_n$:

$$\begin{aligned} G^e_n : \hat{\mathcal{E}} &\to \mathcal{E}^e_n \\ \xi &\mapsto (t, \bar{x}) = x \end{aligned} \tag{6.1.1}$$

with

$$t(\xi) = (1 - \xi_0)t_n + \xi_0 t_{n+1} \tag{6.1.2}$$

$$\bar{x}(\xi) = (1 - \xi_0)F^e_n(\bar{\xi}) + \xi_0 F^e_{n+1}(\bar{\xi}) \tag{6.1.3}$$

with $\xi \in \hat{\mathcal{E}}$ the computational coordinates in the master element $\hat{\mathcal{E}} = [0, 1]^4$. Since the functions $t(\xi)$ and $\bar{x}(\xi)$ are continuously differentiable with respect to ξ, we obtain

$$\begin{pmatrix} \mathrm{d}t \\ \\ \mathrm{d}\bar{x} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial t(\xi)}{\partial \xi_0} & \dfrac{\partial t(\xi)}{\partial \bar{\xi}} \\ \dfrac{\partial \bar{x}(\xi)}{\partial \xi_0} & \dfrac{\partial \bar{x}(\xi)}{\partial \bar{\xi}} \end{pmatrix} \begin{pmatrix} \mathrm{d}\xi_0 \\ \\ \mathrm{d}\bar{\xi} \end{pmatrix}$$

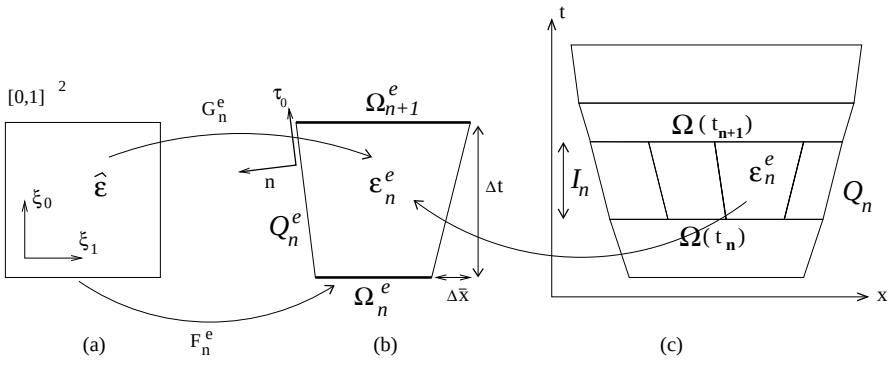

Figure 6.1: Illustration of the geometry of two-dimensional space-time elements in both computational and physical space

where the above $(d+1)\times(d+1)$ matrix is the Jacobian matrix of the transformation (6.1.1) and we denote it by $J_{\mathcal{E}}$. Define the matrix

$$J_{\bar{x}} = \begin{pmatrix} \frac{\partial x_1}{\partial \xi_1} & \cdots & \frac{\partial x_1}{\partial \xi_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial \xi_1} & \cdots & \frac{\partial x_d}{\partial \xi_d} \end{pmatrix}. \tag{6.1.4}$$

Using (6.1.2-6.1.3), we obtain the relations

$$\frac{\partial x_0}{\partial \xi_0} = t_{n+1} - t_n = \Delta t, \quad \text{and} \quad \frac{\partial x_0}{\partial \xi_i} = \frac{\partial t(\xi)}{\partial \xi_i} = 0,\ i = 1, \ldots, d, \tag{6.1.5}$$

thus $J_{\mathcal{E}}$ can be written as

$$J_{\mathcal{E}} = \begin{pmatrix} \frac{\partial x_0}{\partial \xi_0} & \cdots & \frac{\partial x_0}{\partial \xi_d} \\ \frac{\partial x_1}{\partial \xi_0} & \cdots & \frac{\partial x_1}{\partial \xi_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial \xi_0} & \cdots & \frac{\partial x_d}{\partial \xi_d} \end{pmatrix} = \begin{pmatrix} \Delta t & 0^T \\ \\ F^e_{n+1}(\bar{\xi}) - F^e_n(\bar{\xi}) & (1-\xi_0)\dfrac{\partial F^e_n(\bar{\xi})}{\partial \bar{\xi}} + \xi_0 \dfrac{\partial F^e_{n+1}(\bar{\xi})}{\partial \bar{\xi}} \end{pmatrix}$$

with the determinant given by

$$|J_{\mathcal{E}}| = \Delta t |J_{\bar{x}}| = \Delta t \det\left((1-\xi_0)\frac{\partial F^e_n(\bar{\xi})}{\partial \bar{\xi}} + \xi_0 \frac{\partial F^e_{n+1}(\bar{\xi})}{\partial \bar{\xi}} \right).$$

The inverse of the matrix $J_{\mathcal{E}}$ has the form

$$\mathcal{I} := J_{\mathcal{E}}^{-1} = \begin{pmatrix} \frac{\partial \xi_0}{\partial x_0} & \frac{\partial \xi_0}{\partial x_1} & \cdots & \frac{\partial \xi_0}{\partial x_d} \\ \frac{\partial \xi_1}{\partial x_0} & \frac{\partial \xi_1}{\partial x_1} & \cdots & \frac{\partial \xi_1}{\partial x_d} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial \xi_d}{\partial x_0} & \frac{\partial \xi_d}{\partial x_1} & \cdots & \frac{\partial \xi_d}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \frac{1}{\Delta t} & 0 & \cdots & 0 \\ \frac{\partial \xi_1}{\partial x_0} & \frac{\partial \xi_1}{\partial x_1} & \cdots & \frac{\partial \xi_1}{\partial x_d} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial \xi_d}{\partial x_0} & \frac{\partial \xi_d}{\partial x_1} & \cdots & \frac{\partial \xi_d}{\partial x_d} \end{pmatrix}.$$

We define the spatial and time submatrices of $J_{\mathcal{E}}^{-1}$ as

$$\bar{\mathcal{I}} := \begin{pmatrix} \frac{\partial \xi_0}{\partial x_1} & \cdots & \frac{\partial \xi_0}{\partial x_d} \\ \frac{\partial \xi_1}{\partial x_1} & \cdots & \frac{\partial \xi_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \xi_d}{\partial x_1} & \cdots & \frac{\partial \xi_d}{\partial x_d} \end{pmatrix} = \begin{pmatrix} 0 & \cdots & 0 \\ \frac{\partial \xi_1}{\partial x_1} & \cdots & \frac{\partial \xi_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \xi_d}{\partial x_1} & \cdots & \frac{\partial \xi_d}{\partial x_d} \end{pmatrix}, \quad \mathcal{I}_t := \begin{pmatrix} \frac{\partial \xi_0}{\partial x_0} \\ \frac{\partial \xi_1}{\partial x_0} \\ \cdots \\ \frac{\partial \xi_d}{\partial x_0} \end{pmatrix} = \begin{pmatrix} \frac{1}{\Delta t} \\ \frac{\partial \xi_1}{\partial x_0} \\ \cdots \\ \frac{\partial \xi_d}{\partial x_0} \end{pmatrix} \tag{6.1.6}$$

and

$$\mathcal{I}_{\bar{x}} := J_{\bar{x}}^{-1} = \begin{pmatrix} \frac{\partial \xi_1}{\partial x_1} & \cdots & \frac{\partial \xi_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \xi_d}{\partial x_1} & \cdots & \frac{\partial \xi_d}{\partial x_d} \end{pmatrix}, \quad \bar{\mathcal{I}}_t := \begin{pmatrix} \frac{\partial \xi_1}{\partial x_0} \\ \cdots \\ \frac{\partial \xi_d}{\partial x_0} \end{pmatrix} \tag{6.1.7}$$

respectively. We will frequently use the notations

$$\Omega(t_n^+) = \lim_{\epsilon \to 0} \Omega^e(t_n + \epsilon), \quad \Omega(t_{n+1}^-) = \lim_{\epsilon \to 0} \Omega^e(t_{n+1} - \epsilon)$$

to indicate that the mesh can change discontinuously at the time levels t_n and t_{n+1}. Each space-time slab $\mathcal{E}_n$ is bounded by the hypersurfaces $\Omega(t_n^+)$, $\Omega(t_{n+1}^-)$ and $\mathcal{Q}_n = \partial\mathcal{E}_n \setminus (\Omega(t_n^+) \cup \Omega(t_{n+1}^-))$. Similarly, each space-time element is bounded by the hypersurfaces $\Omega^e(t_n^+)$, $\Omega^e(t_{n+1}^-)$ and $\mathcal{Q}_n^e = \partial\mathcal{E}_n^e \setminus (\Omega^e(t_n^+) \cup \Omega^e(t_{n+1}^-))$.

6.2 Weak formulation of the incompressible Navier-Stokes equations

The time-discontinuous Galerkin method lacks stability. To improve the stability while maintaining accuracy, we add a least-squares operator to the basic Galerkin formulation.

The trial function space in each space-time slab $\mathcal{E}_n$ is denoted by V_h^n and the test function space by W_h^n. Their elements are assumed to be $\mathcal{C}^0$ continuous within each space-time slab, but discontinuous across the interfaces of the space-time slabs, namely at times $t_1, t_2, \ldots, t_{N-1}$. The *finite element spaces* are now defined as:

$$V_h^n = \Big\{ V \in H^1(\mathcal{E}_n)^5 : V\,|_{\mathcal{E}_n^e} \circ G_n^e \in \Big(\hat{\mathcal{P}}_1(0,1) \otimes \hat{\mathcal{P}}_k(\hat{\Omega})\Big)^5, \forall\, \mathcal{E}_n^e \in \mathcal{T}_h^n,$$
$$q_1(V) = \bar{q}_1 \text{ on } \mathcal{Q}_n \Big\}$$

$$W_h^n = \Big\{ W \in H^1(\mathcal{E}_n)^5 : W\,|_{\mathcal{E}_n^e} \circ G_n^e \in \Big(\hat{\mathcal{P}}_1(0,1) \otimes \hat{\mathcal{P}}_k(\hat{\Omega})\Big)^5, \ \forall\, \mathcal{E}_n^e \in \mathcal{T}_h^n,$$
$$q_2(W) = \bar{q}_2 \text{ on } \mathcal{Q}_n \Big\},$$

where G_n^e denotes the mapping from the space-time reference element $(0,1) \times \hat{\Omega}$, with $\hat{\Omega}$ the hexahedral reference element in R^3, to the element in physical space $\mathcal{E}_n^e$, and $\hat{\mathcal{P}}_k$ represent kth-order polynomials. Further, V_1, W_1 denote the first component of $V, W \in \mathbb{R}^5$, respectively, $q_1 : \mathcal{E}^5 \to \mathbb{R}^4$ are the (nonlinear) boundary conditions for the components V_2, V_3, V_4, and V_5 of V, with a similar expression for $q_2 : \mathcal{E}^5 \to \mathbb{R}^4$, and $\bar{q}_1, \bar{q}_2 \in \mathbb{R}^4$ are the prescribed boundary conditions. Note, not necessarily all components of V will have imposed boundary conditions, this depends on the type of boundary condition. When the finite element spaces are defined on the whole space time domain then the superscript n is omitted.

Let us recall the Galerkin least-squares variational formulation of the incompressible Navier-Stokes equations in terms of entropy variables:

Find a $V \in V_h$, such that for all $W \in W_h$, the following relation is satisfied

$$\sum_{n=0}^{N} \{B_{Gal}(V,W) + B_{LS}(V,W) + B_{jump}(V,W)\} = 0, \tag{6.2.1}$$

where the first term in (6.2.1) is the Galerkin term,

$$B_{Gal}(V,W) = \int_{\mathcal{E}_n} W \cdot \Big(F_\ell(V)_{,\ell} - (\tilde{K}_{ij} V_{,j})_{,i} \Big) \, \mathrm{d}\mathcal{E} \quad \text{for } \ \ell = 0, \ldots, 3, \ i,j = 1,2,3,$$

the second term is the least-squares stabilization operator, defined as:

$$B_{LS}(V,W) = \sum_{e=1}^{(n_{el})_n} \int_{\mathcal{E}_n^e} (\mathcal{L}_V W) \cdot \tilde{\tau} (\mathcal{L}_V V) \mathrm{d}\mathcal{E},$$

with $\mathcal{L}_V$ the symmetrized Navier-Stokes operator:

$$\mathcal{L}_V = \tilde{A}_\ell \frac{\partial}{\partial x_\ell} - \frac{\partial}{\partial x_i} (\tilde{K}_{ij} \frac{\partial}{\partial x_j}) \quad \text{for } \ell = 0, \ldots, 3, \ i,j = 1,2,3,$$

and the last term in (6.2.1) is the so-called jump term, defined as:

$$B_{jump}(V,W) = \int_{\Omega(t_n)} W(t_n^+) \cdot \left[F_0(V(t_n^+)) - F_0(V(t_n^-)) \right] \mathrm{d}\Omega. \tag{6.2.2}$$

The jump term is added to the weak formulation to ensure weak continuity between different space-time slabs. The stabilization operator is added in order to prevent numerical oscillations in regions with strong gradients which are not well represented on the computational mesh, and to ensure a unique solution satisfying the inf-sup condition, see for instance [48]. In the least-squares operator, the choice for the $\tilde{\tau}$ matrix is crucial, and is examined in detail in Chapter 4 of this thesis. This operator greatly influences the stability of the numerical scheme. The use of a stabilization operator avoids the need of using different order elements for pressure and velocity to meet the inf-sup stability condition, which is a common problem for incompressible flow.

6.3 Transformation of the space-time weak formulation into ALE form

In this section we will establish the relation between the Arbitrary Lagrangian Eulerian (ALE) and space-time weak formulation of the incompressible Navier-Stokes equations.

The weak formulation (6.2.1) can be transformed into an integrated-by-parts form using Gauss' theorem. If we introduce

$$W \cdot F_{\ell,\ell}(V) = (W \cdot F_\ell(V))_{,\ell} - W_{,\ell} \cdot F_\ell(V)$$

into the weak formulation (6.2.1) and apply Gauss' theorem, we obtain:

$$\int_{\mathcal{E}_n} (W\cdot F_\ell(V))_{,\ell}\,\mathrm{d}\mathcal{E} = \int_{\partial\mathcal{E}_n} n\cdot(W^T\mathcal{F}(V))\,\mathrm{d}\,(\partial\mathcal{E}) = \int_{\partial\mathcal{E}_n} n_\ell(W^T F_\ell(V))\,\mathrm{d}\,(\partial\mathcal{E}) \quad (6.3.1)$$

where $n\in\mathbb{R}^4$ is the unit outward space-time normal vector at the domain boundary $\partial\mathcal{E}_n$. The ALE form can now be obtained by calculating the space-time normal vector n.

Given the parametrization $(t,\bar{x}) = G_n^e(\xi)$ for the space-time element, the space-time normal vector n_e at the boundary surface is orthogonal to the tangential vectors τ_j, $j=0,\ldots,3$. The tangential vectors are defined as $\tau_j = \frac{\partial G_n^e}{\partial\xi_j}$, and are equal to:

$$\tau_0 = \begin{pmatrix} t_{n+1}-t_n \\ F_{n+1}^e(\bar{\xi}) - F_n^e(\bar{\xi}) \end{pmatrix} = \begin{pmatrix} \Delta t \\ \Delta\bar{x} \end{pmatrix}, \quad (6.3.2)$$

$$\tau_j = \begin{pmatrix} 0 \\ (1-\xi_0)\dfrac{\partial F_n^e(\bar{\xi})}{\partial\xi_j} + \xi_0\dfrac{\partial F_{n+1}^e(\bar{\xi})}{\partial\xi_j} \end{pmatrix}, \quad j=1,2,3. \quad (6.3.3)$$

First we split the boundary integral in (6.3.1) into integrals over different types of boundaries. The normal vectors at $\Omega(t_{n+1}^-)$ and $\Omega(t_n^+)$ are $(1,0,0,0)^T$ and $(-1,0,0,0)^T$, respectively. Hence, the boundary integral over the surfaces $\Omega(t_{n+1}^-)$ and $\Omega(t_n^+)$ is equal to

$$\int_{\Omega(t_{n+1})\cup\Omega(t_n)} n\cdot(W^T\mathcal{F}(V))\,\mathrm{d}\Omega = \int_{\Omega(t_{n+1})} W(t_{n+1}^-)F_0(V(t_{n+1}^-))\mathrm{d}\Omega - \int_{\Omega(t_n)} W(t_n^+)F_0(V(t_n^+))\mathrm{d}\Omega. \quad (6.3.4)$$

By adding the jump term (6.2.2) to (6.3.4) and sum over all space-time slabs, we obtain

$$\sum_{n=0}^{N}\left\{\int_{\Omega(t_{n+1})\cup\Omega(t_n)} n\cdot(W^T\mathcal{F}(V))\,\mathrm{d}\Omega + B_{jump}(V,W)\right\}$$
$$= \sum_{n=0}^{N}\left\{\int_{\Omega(t_{n+1})} W(t_{n+1}^-)F_0(V(t_{n+1}^-))\mathrm{d}\Omega - \int_{\Omega(t_n)} W(t_n^+)F_0(V(t_n^-))\mathrm{d}\Omega\right\}.$$

For the remaining parts of the boundary, let $\bar{n}^i(\bar{x},t)\in\mathbb{R}^3$, $1\le i\le 6$, be the spatial component of the normal vector at the boundary $\mathcal{Q}_n$. By definition, $\bar{n}^i$ for $1\le i\le 6$, is perpendicular to the tangential vectors

$$\bar{\tau}_j = (1-\xi_0)\frac{\partial F_n^e(\bar{\xi})}{\partial\xi_j} + \xi_0\frac{\partial F_{n+1}^e(\bar{\xi})}{\partial\xi_j}, \quad j=1,2,3.$$

Hence, the space-time normal vectors $n^i = (n_0, \bar{n}^i)$, $1 \leq i \leq 6$, are orthogonal to the tangential vectors defined in (6.3.2-6.3.3) if and only if the conditions

$$\Delta t \; n_0 + \Delta x \cdot \bar{n}^i = 0$$

are satisfied. The space-time normal vector n^i, $(1 \leq i \leq 6)$ at the boundary $\mathcal{Q}_n$ is equal to

$$n^i = (-v \cdot \bar{n}^i, \bar{n}^i)^T,$$

with the grid velocity $v \in \mathbb{R}^3$ given by the relation $\Delta\bar{x}/\Delta t$. Therefore, the boundary integral over the boundary $\mathcal{Q}_n$ is equal to

$$\int_{\mathcal{Q}_n} n \cdot (W^T \mathcal{F}) \mathrm{d}\mathcal{Q} = \int_{\mathcal{Q}_n} \left(\bar{n}_i (W^T F_i(V)) - \bar{n} \cdot v (W^T F_0(V)) \right) \mathrm{d}\mathcal{Q}$$

with $i = 1, 2, 3$. Finally, summing over all space-time slabs we obtain the weak formulation of the symmetrized Navier-Stokes equations in ALE form for the whole space-time domain:

Find $V \in V_h$, such that for all $W \in W_h$, the following relation is satisfied:

$$\sum_{n=0}^{N} \Big\{ \int_{\mathcal{E}_n} \Big(-W_{,0} \cdot F_0(V) - W_{,i} \cdot F_i(V) + W_{,i} \cdot \tilde{K}_{ij} V_{,j} \Big) \mathrm{d}\mathcal{E} \tag{6.3.5}$$

$$+ \int_{\Omega(t_{n+1})} W(t_{n+1}^-) \cdot F_0(V(t_{n+1}^-)) \mathrm{d}\Omega - \int_{\Omega(t_n)} W(t_n^+) \cdot F_0(V(t_n^-)) \mathrm{d}\Omega \tag{6.3.6}$$

$$+ \sum_{e=1}^{(n_{el})_n} \int_{\mathcal{E}_n^e} (\mathcal{L}_V W) \cdot \tilde{\tau} (\mathcal{L}_V V) \mathrm{d}\mathcal{E} - \int_{\mathcal{Q}_n} \bar{n} \cdot v (W \cdot F_0(V)) \mathrm{d}\mathcal{Q} \tag{6.3.7}$$

$$+ \int_{\mathcal{Q}_n} W \cdot \Big(F_i(V) - \tilde{K}_{ij} V_{,j} \Big) \bar{n}_i \; \mathrm{d}\mathcal{Q} \Big\} = 0. \tag{6.3.8}$$

6.4 Finite element basis functions

In this section we describe the finite element basis functions for vector-valued problems. In this thesis the shape functions are chosen such that only one component in the vector is nonzero. This is the case, if we choose the shape functions to be

$$\Psi_i = (0, \ldots, 0, \psi_i(x), 0, \ldots, 0)^T \in \mathbb{R}^{d+2}$$

where Ψ_i is a vector-valued shape function with the scalar shape function ψ_i the only non-zero component. Let us denote by $c(i)$ the index of this non-zero component, then the lth component of Ψ_i can also be written as

$$(\Psi_i(x))_l = \psi_i(x) \delta_{c(i)\, l}, \tag{6.4.1}$$

with the Kronecker delta function δ_{jk}. For notational simplicity, we use the scalar function notation in the remainder and it is straightforward that for vector-valued basis functions we mean (6.4.1).

In the finite element discretization we use finite element spaces that consist of basis functions that are piecewise *linear in time and higher order polynomials in space.* First, in the master element $\hat{\mathcal{E}}$ the basis functions $\hat{\phi}_m : \hat{\mathcal{E}} \to \mathbb{R}$, for $m = 1, \ldots, 2\,n_{dof}$ are defined which are linear in time and have the following form

$$\hat{\phi}_m(\xi) = \lambda_i(\xi_0)\hat{\psi}_k(\bar{\xi}) \quad \text{for } i = 1, 2,\ k = 1, \ldots, n_{dof}$$

where $\lambda_1(\tau) = 1 - \tau,\ \lambda_2(\tau) = \tau$ and n_{dof} denotes the number of degrees of freedom on the reference element $\hat{\Omega}$. The functions $\hat{\psi}_k : \hat{\Omega} \to \mathbb{R}$ are the general Lagrangian basis functions defined on the d-dimensional cube ($d = 2$ or 3) as

$$\hat{\psi}_k(\bar{a}_j) = \delta_{kj}, \quad j = 1, \ldots, n_{dof},$$

with $\hat{\psi}_k$ polynomials on $\hat{\Omega}$, δ_{kj} the Kronecker delta, and $\bar{a}_j,\ j = 1, \ldots, n_{dof}$ the nodes of the finite element mesh in $\hat{\Omega}$.

Next, the basis functions $\phi^e_m : \mathcal{E}^e_n \to \mathbb{R}$ are constructed through the parametrization G^e_n as

$$\phi^e_m = \hat{\phi}_m \circ {G^e_n}^{-1}, \quad m = 1, \ldots, 2\,n_{dof},$$

or $\phi^e_m(x) = \hat{\phi}_m({G^e_n}^{-1}(x))$.

Hence, within the nth space-time slab, the finite element trial solution can be defined as

$$\begin{aligned} V^h(t, \bar{x}) &= \sum_{A=1}^{(n_{np})_{(n)}} \phi_A(t, \bar{x}) V_A = \sum_{A=1}^{(n_{np})_{(n)}} \left(\phi_{A;n+1}(t, \bar{x}) v_{A;n+1} + \phi_{A;n}(t, \bar{x})\tilde{v}_{A;n}\right) \\ &= \sum_{A=1}^{(n_{np})_{(n)}} \hat{\phi}_A({G^e_n}^{-1}(x)) V_A \\ &= \sum_{A=1}^{(n_{np})_{(n)}} \left(\lambda_2(\xi_0)\hat{\psi}_A(\bar{\xi}) v_{A;n+1} + \lambda_1(\xi_0)\hat{\psi}_A(\bar{\xi})\tilde{v}_{A;n}\right), \end{aligned} \tag{6.4.2}$$

where $v_{A;n+1}$ and $\tilde{v}_{A;n}$ are the nodal values of $V^h(t, \bar{x})$ at node A and times t^-_{n+1} and t^+_n, respectively. The vector of unknowns at time level $(n + 1)$ is denoted by

$$v_{n+1} = \left(v^T_{1;n+1}, v^T_{2;n+1}, \ldots, v^T_{n_{np};n+1}\right)^T \tag{6.4.3}$$

and we will call them *primary variables.* The vector of unknowns at the nth time level is

$$\tilde{v}_n = \left(\tilde{v}^T_{1;n}, \tilde{v}^T_{2;n}, \ldots, \tilde{v}^T_{n_{np};n}\right)^T \tag{6.4.4}$$

and we call them *secondary variables.* We denote by $\phi_{A;n+1}(x)$ and $\phi_{A;n}(x)$ the finite element shape functions related to the nodes A at the time levels $(n+1)$ and n, respectively, and $(n_{np})_n$ is the number of nodal points for the nth space-time slab. Then,

$$\phi^e_{A;n+1}(x) = \lambda_2(\xi_0)\hat{\psi}_A(\bar{\xi}), \quad \phi^e_{A;n}(x) = \lambda_1(\xi_0)\hat{\psi}_A(\bar{\xi}). \tag{6.4.5}$$

Due to the definition of the basis functions in physical space, we can write

$$\phi_{A;n+1}(t^-_{n+1}, \bar{x}(\xi)) = \hat{\psi}_A(\bar{\xi}), \quad \phi_{A;n}(t^+_n, \bar{x}(\xi)) = \hat{\psi}_A(\bar{\xi}),$$

and the trial solution at the time levels t^-_{n+1} and t^+_n are therefore equal to

$$V^h(t^-_{n+1}, \bar{x}) = \sum_{A=1}^{(n_{np})_{(n)}} \hat{\psi}_A(\bar{\xi}) v_{A;n+1}, \quad V^h(t^+_n, \bar{x}) = \sum_{A=1}^{(n_{np})_{(n)}} \hat{\psi}_A(\bar{\xi}) \tilde{v}_{A;n}.$$

The test functions on the nth space-time slab are defined as:

$$W^h(t, \bar{x}) = \sum_{A=1}^{(n_{np})_{(n)}} \phi_A(x) w_A = \sum_{A=1}^{(n_{np})_{(n)}} \hat{\phi}_A(G_n^{-1}(x)) w_A$$

and we write them in the form

$$W^h(t, \bar{x}) = \sum_{A=1}^{(n_{np})_{(n)}} \left(\hat{\psi}_A(\bar{\xi}) w_{A;n+1} + \hat{\psi}_A(\bar{\xi})(\lambda_1(\xi_0) - \lambda_2(\xi_0)) \tilde{w}_{A;n} \right), \text{ for } (t, \bar{x}) \in \mathcal{E}_n \tag{6.4.6}$$

where $w_{A;n+1}$ and $\tilde{w}_{A;n}$ are the nodal values of the weighting function corresponding to $v_{A;n+1}$ and $\tilde{v}_{A;n}$, respectively. The values of the test functions at the time levels t^-_{n+1} and t^+_n are

$$W^h(t^-_{n+1}, \bar{x}) = \sum_{A=1}^{(n_{np})_{(n)}} \left\{ \hat{\psi}_A(\bar{\xi}) w_{A;n+1} - \hat{\psi}_A(\bar{\xi}) \tilde{w}_{A;n} \right\},$$

and

$$W^h(t^+_n, \bar{x}) = \sum_{A=1}^{(n_{np})_{(n)}} \left\{ \hat{\psi}_A(\bar{\xi}) w_{A;n+1} + \hat{\psi}_A(\bar{\xi}) \tilde{w}_{A;n} \right\},$$

respectively.

Remark 6.4.1 *The motivation for the choice of the test functions given in (6.4.6) goes back to the extensive work done by Shakib et al. [51]. If we choose the test function in the same way as the trial functions, we end up with a nonlinear system of equations where the temporal coupling between the primary and secondary variables will be non-symmetric. In [50], Shakib showed that an alternative to avoid this problem is to use a left preconditioning matrix of the form*

$$\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

We can incorporate this preconditioning directly into the weighting function, which results in (6.4.6).

If we introduce the relations for $V^h(x)$ and $W^h(x)$, given by (6.4.2) and (6.4.6), into the weak formulation (6.3.5-6.3.8), and assume that there are no essential boundary conditions, then the time-discontinuous Galerkin least-squares finite element discretization results in a system of nonlinear algebraic equations for the coefficients v and $\tilde{v}$ in every space-time slab:

$$G(v, \tilde{v}) = 0 \tag{6.4.7}$$

$$\tilde{G}(v, \tilde{v}) = 0. \tag{6.4.8}$$

Note, we omitted the subscript n for the vectors v and $\tilde{v}$ since the same applies to all space-time slabs $\mathcal{E}_n$. In [50], implicit/explicit predictor multi-corrector algorithms were developed to reduce the nonlinear system to a sequence of linear systems. The solution strategy employed will be presented later in Section 6.6. Denote by

$$R = G(v, \tilde{v}),$$
$$\tilde{R} = \tilde{G}(v, \tilde{v}),$$

the primary and secondary residual vectors and by

$$M_{11} = \frac{\partial G(v, \tilde{v})}{\partial v}, \quad M_{12} = \frac{\partial G(v, \tilde{v})}{\partial \tilde{v}},$$
$$M_{21} = \frac{\partial \tilde{G}(v, \tilde{v})}{\partial v}, \quad M_{22} = \frac{\partial \tilde{G}(v, \tilde{v})}{\partial \tilde{v}}$$

the consistent tangent matrices, respectively. The residual vectors and tangent matrices can be constructed in each space-time slab from their element contribution as

$$R = \bigwedge_{e=1}^{n_{el}} R^e, \quad R^e = (R^e_a), \quad a = 1, \ldots, n_{en}$$
$$M_{11} = \bigwedge_{e=1}^{n_{el}} M^e_{11}, \quad M^e_{11} = (M^{ab}_{11}), \quad a, b = 1, \ldots, n_{en}$$

where n_{en} is the number of element nodes, R^e_a is the residual at node a of element $\mathcal{E}^e_n$, M^{ab}_{11} is the entry of the element matrix M^e_{11} corresponding to the element nodes a and b and $\bigwedge$ is the assembly operator. Note, the secondary residual $\tilde{R}$ and the matrices M_{12}, M_{21} and M_{22} are constructed analogously. In the next section we will compute the entries of the element residual and matrices, respectively.

6.5 Space-time finite element discretization

In this section we discuss the Galerkin least-squares finite element discretization of the symmetrized incompressible Navier-Stokes equations. The trial and test functions are chosen as in (6.4.2) and (6.4.6) respectively.

Let $V^e(\bar{x})$ and $\tilde{V}^e(\bar{x})$ respectively, be the restrictions of $V^h(t_{n+1}^-, \bar{x})$ and $V^h(t_n^+, \bar{x})$ to the eth element

$$V^e(\bar{x}) := V^h(t_{n+1}^-, F_{n+1}^e(\bar{\xi})),$$
$$\tilde{V}^e(\bar{x}) := V^h(t_n^+, F_n^e(\bar{\xi})),$$

and let

$$\bar{V}^e(\bar{x}) = \frac{1}{2}(V^e(\bar{x}) + \tilde{V}^e(\bar{x})),$$
$$\overset{\triangle}{V}{}^e(\bar{x}) = \frac{1}{2}(V^e(\bar{x}) - \tilde{V}^e(\bar{x}))$$

denote the average- and difference-in-time values of the entropy solution in the space-time slab $\mathcal{E}_n$.

The basis functions at node a of element $\mathcal{E}_n^e$, $\phi_a^e(t, \bar{x})$ in the physical coordinates are related to the basis functions in the computational coordinates $\hat{\phi}_a(\xi)$ as $\phi_a^e(x) = \hat{\phi}_a(\xi)$, for $a = 1, \ldots, n_{en}$, where n_{en} is the number of element nodes (e.g. $n_{en} = 4$ for bilinear and $n_{en} = 9$ for biquadratic quadrilaterals). Note here that we used A to denote the nodes in a space-time slab and a for the nodes of a space-time element. Thus, we obtain

$$\frac{\partial \phi_a^e(t, \bar{x})}{\partial x_\ell} = \frac{\partial \hat{\phi}_a(G_n^{e\,-1}(x))}{\partial x_\ell} = \sum_{j=0}^{d} \frac{\partial \xi_j}{\partial x_\ell} \frac{\partial \hat{\phi}_a(\xi)}{\partial \xi_j} \quad \text{for } \ell = 0, \ldots, d.$$

The above relation can be written in the following closed form

$$\nabla_x \phi_a^e(t, \bar{x}) = \mathcal{I}^T \nabla_\xi \hat{\phi}_a(\xi)$$

where $\nabla_\xi = (\frac{\partial}{\partial \xi_0}, \frac{\partial}{\partial \xi_1}, \ldots, \frac{\partial}{\partial \xi_d})^T$ and $\nabla_x = (\frac{\partial}{\partial x_0}, \ldots, \frac{\partial}{\partial x_d})^T$. Since our solution is vector valued, we need to generalize the above gradient operators. Denote by ∇_ξ and $\nabla_{\bar{\xi}}$ the local gradient in the element space-time and spatial coordinate systems, respectively. That is

$$\nabla_\xi = \begin{pmatrix} \frac{\partial}{\partial \xi_0} I_m \\ \frac{\partial}{\partial \xi_1} I_m \\ \vdots \\ \frac{\partial}{\partial \xi_d} I_m \end{pmatrix}, \quad \nabla_{\bar{\xi}} = \begin{pmatrix} \frac{\partial}{\partial \xi_1} I_m \\ \frac{\partial}{\partial \xi_2} I_m \\ \vdots \\ \frac{\partial}{\partial \xi_d} I_m \end{pmatrix},$$

where I_m is the $m \times m$ identity matrix. The gradients with respect to the physical space coordinates, ∇_x and $\nabla_{\bar{x}}$, are defined analogously. Using the definition of the transformation from the reference to the physical coordinates, discussed in Section 6.1, the spatial gradient operator can be decomposed as

$$\nabla_x = \mathcal{I}^T \nabla_\xi = \begin{pmatrix} \frac{1}{\Delta t} & \bar{\mathcal{I}}_t^T \\ 0 & \mathcal{I}_{\bar{x}}^T \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial \xi_0} \\ \nabla_{\bar{\xi}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\Delta t}\frac{\partial}{\partial \xi_0} + \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \\ \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \end{pmatrix}. \tag{6.5.1}$$

The spatial gradient and time derivative of $\phi_a^e(t, \bar{x})$ are then related to the gradient of $\hat{\phi}_a(\xi)$ as

$$\nabla_{\bar{x}} \phi_a^e(t, \bar{x}) = \bar{\mathcal{I}}^T \nabla_\xi \hat{\phi}_a(\xi) = \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\phi}_a(\xi), \tag{6.5.2}$$

$$\frac{\partial \phi_a^e(t, \bar{x})}{\partial t} = \mathcal{I}_t^T \nabla_\xi \hat{\phi}_a(\xi) = \frac{1}{\Delta t}\frac{\partial}{\partial \xi_0}\hat{\phi}_a(\xi) + \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\phi}_a(\xi). \tag{6.5.3}$$

The second equality in (6.5.2) follows from the property of the transformation from reference to physical coordinates, see (6.1.6). Using (6.4.5), it follows that the time derivative of the basis functions of node a of an element $\mathcal{E}_n^e$ at time levels $n+1$ and n are related to the basis functions on the master element as

$$\begin{aligned} \frac{\partial \phi_{a;n+1}^e(t, \bar{x})}{\partial t} &= \mathcal{I}_t^T \nabla_\xi (\lambda_2(\xi_0)\hat{\psi}_a(\bar{\xi})) = \mathcal{I}_t^T \left(\hat{\psi}_a(\bar{\xi}), \xi_0 \nabla_{\bar{\xi}}^T \hat{\psi}_a(\bar{\xi}) \right)^T \\ &= \frac{1}{\Delta t}\hat{\psi}_a(\bar{\xi}) + \xi_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}), \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \phi_{a;n}^e(t, \bar{x})}{\partial t} &= \mathcal{I}_t^T \nabla_\xi (\lambda_1(\xi_0)\hat{\psi}_a(\bar{\xi})) = \mathcal{I}_t^T \left(-\hat{\psi}_a(\bar{\xi}), (1-\xi_0) \nabla_{\bar{\xi}}^T \hat{\psi}_a(\bar{\xi}) \right)^T \\ &= -\frac{1}{\Delta t}\hat{\psi}_a(\bar{\xi}) + (1-\xi_0) \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}), \end{aligned}$$

respectively.

The time derivative of the trial solution in the element $\mathcal{E}_n^e$ can then be given as:

$$V_{,t}^h(t, \bar{x}) = \frac{2}{\Delta t} \overset{\triangle}{V^e}(\bar{x}) + \sum_{a=1}^{n_{en}} \left\{ \xi_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) v_{a;n+1} + (1-\xi_0) \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \tilde{v}_{a;n} \right\},$$

and the time derivative of the test function can be written as:

$$W_{,t}^h(t, \bar{x}) = \sum_{a=1}^{n_{en}} \left\{ \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) w_{a;n+1} + \left(-\frac{2}{\Delta t}\hat{\psi}_a(\bar{\xi}) + (1-2\xi_0) \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \right) \tilde{w}_{a;n} \right\}.$$

Note here that when the mesh is not deforming in time, $\bar{\mathcal{I}}_t^T = (0, \ldots, 0) \in \mathbb{R}^d$, therefore, on a fixed mesh we have

$$V_{,t}^h(t, \bar{x}) = \frac{2}{\Delta t} \overset{\triangle}{V^e}(\bar{x}) \quad \text{and} \quad W_{,t}^h(t, \bar{x}) = -\frac{2}{\Delta t} \sum_{a=1}^{n_{en}} \hat{\psi}_a(\bar{\xi}) \tilde{w}_{a;n}.$$

Using (6.5.2), the space derivatives of the basis functions are given as

$$\begin{aligned}\nabla_{\bar{x}}\psi^e_{a;n+1}(t,\bar{x}) &= \bar{\mathcal{I}}^T\nabla_\xi\left(\lambda_2(\xi_0)\hat{\psi}_a(\bar{\xi})\right) = \bar{\mathcal{I}}^T\left(\hat{\psi}_a(\bar{\xi}),\xi_0\nabla^T_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\right)^T\\ &= \xi_0\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\\ \nabla_{\bar{x}}\psi^e_{a;n}(t,\bar{x}) &= \bar{\mathcal{I}}^T\nabla_\xi\left(\lambda_1(\xi_0)\hat{\psi}_a(\bar{\xi})\right) = \bar{\mathcal{I}}^T\left(-\hat{\psi}_a(\bar{\xi}),(1-\xi_0)\nabla^T_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\right)^T\\ &= (1-\xi_0)\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\end{aligned}$$

The spatial gradient of the trial and test functions is now equal to

$$\nabla_{\bar{x}}V^h(t,\bar{x}) = \sum_{a=1}^{n_{en}}\left\{\xi_0\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})v_{a;n+1} + (1-\xi_0)\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\tilde{v}_{a;n}\right\},$$

and

$$\begin{aligned}\nabla_{\bar{x}}W^h(t,\bar{x}) &= \sum_{a=1}^{n_{en}}\Big\{\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})w_{a;n+1}+\\ &\qquad\left(-2\hat{\psi}_a(\bar{\xi})\nabla_{\bar{x}}\xi_0 + (\lambda_1(\xi_0)-\lambda_2(\xi_0))\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\right)\tilde{w}_{a;n}\Big\}\\ &= \sum_{a=1}^{n_{en}}\left\{\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})w_{a;n+1} + (\lambda_1(\xi_0)-\lambda_2(\xi_0))\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\tilde{w}_{a;n}\right\}.\end{aligned}$$

Note that the last equality follows from the $\nabla_{\bar{x}}\xi_0 = 0$ property of the transformation from the reference to the physical coordinates.

Next, the integrals in the weak formulation (6.3.5-6.3.8) are transformed to the reference element to facilitate their numerical evaluation.

The **convective part of the Galerkin contribution** in the weak formulation can be transformed into an expression on the reference element as:

$$\begin{aligned}\boxed{B^e_{Gal_{conv}}(V^h,W^h)} :&= -\int_{\mathcal{E}^e_n} W^h_{,\ell}\cdot F_\ell(V^h)\mathrm{d}\mathcal{E} = -\int_{\mathcal{E}^e_n}\nabla_x W^h\cdot\mathcal{F}(V^h)\mathrm{d}\mathcal{E}\\ &= -\int_{\hat{\mathcal{E}}}\left(\frac{1}{\Delta t}\frac{\partial}{\partial\xi_0}W^h + \bar{\mathcal{I}}^T_t\nabla_{\bar{\xi}}W^h\right)\cdot F_0(V^h)|J_{\mathcal{E}}|\mathrm{d}\hat{\mathcal{E}}\\ &\quad -\int_{\hat{\mathcal{E}}}\mathcal{I}^T_{\bar{x}}\nabla_{\bar{\xi}}W^h\cdot\mathcal{F}^a(V^h)|J_{\mathcal{E}}|\mathrm{d}\hat{\mathcal{E}}\end{aligned}$$

where the convective flux tensor is defined as $\mathcal{F}^{aT} = (F_1^T, F_2^T, F_3^T)$ with F_i the flux in the ith coordinate direction. The test and trial functions in the integral over the

reference element are equal to $W^h = W^h(t(\xi), \bar{x}(\xi))$ and $V^h = V^h(t(\xi), \bar{x}(\xi))$ and the integrals can be evaluated straightforwardly using Gaussion quadrature.

The **diffusive term in the Galerkin contribution** can be transformed to an integral over the reference element as:

$$\boxed{B^e_{Gal_{diff}}(V^h, W^h)} := \int_{\mathcal{E}^e_n} W^h_{,i} \cdot \tilde{K}_{ij} V^h_{,j} \mathrm{d}\mathcal{E} = \int_{\mathcal{E}^e_n} \nabla_{\bar{x}} W^h \cdot \mathcal{F}^d(V^h) \mathrm{d}\mathcal{E}$$
$$= \int_{\hat{\mathcal{E}}} \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} W^h \cdot \mathcal{F}^d(V^h) \, |J_{\mathcal{E}}| \, \mathrm{d}\hat{\mathcal{E}}$$

where the diffusive flux tensor $\mathcal{F}^{d^T} = (F^{d^T}_1, F^{d^T}_2, F^{d^T}_3)$ and $F^d_i = \tilde{K}_{ij} V^h_{,j}$ for $i, j = 1, 2, 3$. We can write the diffusive flux tensor as

$$\mathcal{F}^d(V^h) = \mathcal{K} \nabla_{\bar{x}} V^h = \mathcal{K} \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} V^h$$

where $\mathcal{K}$ is the tensor defined as $\mathcal{K} = [\tilde{K}_{ij}]$ for $i, j = 1, 2, 3$.

The **least-squares term** in the weak formulation can be written in a closed form as

$$\boxed{B^e_{LS}(V^h, W^h)} := \int_{\mathcal{E}^e_n} (\mathcal{L}_V W^h) \cdot \tilde{\tau} (\mathcal{L}_V V^h) \mathrm{d}\mathcal{E}$$
$$= \int_{\mathcal{E}^e_n} (\tilde{A}^T \nabla_x W^h - \nabla_{\bar{x}} \cdot \mathcal{K} \nabla_{\bar{x}} W^h) \cdot \tilde{\tau} (\tilde{A}^T \nabla_x V^h - \nabla_{\bar{x}} \cdot \mathcal{K} \nabla_{\bar{x}} V^h) \mathrm{d}\mathcal{E}$$

where $\tilde{A}^T = (\tilde{A}_0, \ldots, \tilde{A}_d)$. Let us transform the Navier-Stokes differential operator into a differential operator on the master element:

$$\tilde{A}^T \nabla_x - \nabla_{\bar{x}} \cdot \mathcal{K} \nabla_{\bar{x}} = \tilde{A}^T \mathcal{I}^T \nabla_\xi - \nabla_{\bar{x}} \cdot \mathcal{K} \nabla_{\bar{x}}$$
$$= \tilde{A}_0 \left(\frac{1}{\Delta t} \frac{\partial}{\partial \xi_0} + \bar{\mathcal{I}}^T_t \nabla_{\bar{\xi}} \right) + \mathcal{A}^T \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} - \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}}$$

with $\mathcal{A}^T = (\tilde{A}_1, \ldots, \tilde{A}_d)$. Then, the least-squares term can be written as

$$B^e_{LS}(V^h, W^h) = \int_{\hat{\mathcal{E}}} \left(\frac{1}{\Delta t} \tilde{A}_0 \frac{\partial W^h}{\partial \xi_0} + \tilde{A}_0 \bar{\mathcal{I}}^T_t \nabla_{\bar{\xi}} W^h + \mathcal{A}^T \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} W^h - \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} W^h \right)$$
$$\cdot \tilde{\tau} \left(\frac{2}{\Delta t} \tilde{A}_0 \overset{\triangle}{V^e} + \tilde{A}_0 \bar{\mathcal{I}}^T_t \nabla_{\bar{\xi}} V^h + \mathcal{A}^T \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} V^h - \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}^T_{\bar{x}} \nabla_{\bar{\xi}} V^h \right) |J_{\mathcal{E}}| \mathrm{d}\hat{\mathcal{E}},$$

where we used

$$\frac{\partial V^h}{\partial \xi_0} = \sum_{a=1}^{n_{en}} \left\{ \hat{\psi}(\bar{\xi}_a) v_{a;n+1} - \hat{\psi}(\bar{\xi}_a) \tilde{v}_{a;n} \right\} = 2 \overset{\triangle}{V^e}(\bar{x}).$$

The **boundary integral** in the weak formulation can be transformed to the integral on the reference element by using the following definition given in [55]:

Given a parametrization $F : (0,1)^{d-1} \to S$, where S is a hypersurface in $\mathbb{R}^d$, integration over the surface S is defined as:

$$\int_S f(x)\mathrm{d}x = \int_{(0,1)^{d-1}} f(F(\xi)) \left| \frac{\partial F}{\partial \xi_1} \wedge \cdots \wedge \frac{\partial F}{\partial \xi_{d-1}} \right| \mathrm{d}\xi, \tag{6.5.4}$$

where the outer product $v = w_1 \wedge \cdots \wedge w_{d-1}$, for $d-1$ vectors $w_i \in \mathbb{R}^d$, is defined component-wise by the following rule

$$v^j = \det(w_1, \ldots, w_{d-1}, e_j),$$

with e_j the jth basis vector in $\mathbb{R}^d$.

Using the above definition we can transform the integral over any of the six space-time faces $\mathcal{Q}_n^e$ of the element $\mathcal{E}_n^e$ to an integral over the appropriate face on the boundary $\hat{\mathcal{Q}}$ of the reference element $\hat{\mathcal{E}}$ using the parametrization G_n^e. The Jacobian of the transformation will be denoted by $J_{\mathcal{Q}}$ and it is defined by $\left| \frac{\partial F}{\partial \xi_1} \wedge \cdots \wedge \frac{\partial F}{\partial \xi_{d-1}} \right|$. Then, the boundary integral in the weak formulation can be transformed to the reference element as

$$\boxed{B^e_{Bound}(V^h, W^h)} := \int_{\hat{\mathcal{Q}}} W^h \cdot \Big(F_i(V^h) - \bar{n} \cdot v F_0(V^h) - \mathcal{K}_i \, \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h \Big) \bar{n}_i \; |J_{\mathcal{Q}}| \mathrm{d}\hat{\mathcal{Q}},$$

where $\mathcal{K}_i = (\tilde{K}_{i1}, \tilde{K}_{i2}, \tilde{K}_{i3})$ for $i = 1, 2, 3$.

Finally, the **jump term** in the weak formulation can be written in terms of the reference coordinates as

$$\begin{aligned} \boxed{B^e_{jump}(V^h, W^h)} := & \int_{\hat{\Omega}} W^h(G_n^e(1,\bar{\xi})) F_0(V^e(G_n^e(1,\bar{\xi}))) \; |J_{\bar{x}}(1,\bar{\xi})| \; \mathrm{d}\hat{\Omega} \\ & - \int_{\hat{\Omega}} W^h(G_n^e(0,\bar{\xi})) F_0(V^e(G_{n-1}^e(1,\bar{\xi}))) \; |J_{\bar{x}}(0,\bar{\xi})| \; \mathrm{d}\hat{\Omega}. \end{aligned}$$

Since the jump term is added to the weak formulation to ensure the weak continuity between the space-time slabs, let us denote the trial solution related to an element in the previous time slab by

$$V^e_{(n)}(\bar{x}) := V^e(t_n^-, \bar{x}) = V^e(G_{n-1}^e(1,\bar{\xi})).$$

In the remainder of this section we compute the weak formulation when the test function defined in (6.4.6) is used.

Define the vectors v and $\tilde{v}$ on each element $\mathcal{E}_n^e$ of the nth space-time slab similarly as in (6.4.3) and (6.4.4), respectively. When we collect all contributions to the weak formulation and set the test function equal to

$$\boxed{W^h(t, \bar{x}) = \hat{\psi}_a(\bar{\xi}),}$$

then we obtain the following nodal entry of the element residual in the nonlinear system of algebraic equations for v and $\tilde{v}$ in the space-time element $\mathcal{E}_n^e$:

$$
\begin{aligned}
R_a^e(v,\tilde{v}) = & \int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot F_0(V^e(\bar{x}))\ |J_{\bar{x}}(1,\bar{\xi})|\ \mathrm{d}\hat{\Omega} - \int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot F_0(V^e_{(n)}(\bar{x}))\ |J_{\bar{x}}(0,\bar{\xi})|\ \mathrm{d}\hat{\Omega} \\
& - \int_{\hat{\mathcal{E}}} \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot F_0(V^h)\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}} - \int_{\hat{\mathcal{E}}} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot \mathcal{F}^a(V^h)\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}} \\
& + \int_{\hat{\mathcal{E}}} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot \mathcal{K}\ \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}} \\
& + \int_{\hat{\mathcal{E}}} \Big(\tilde{A}_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) + \mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{x}) \Big) \cdot \tilde{\tau} \\
& \qquad \Big(\frac{2}{\Delta t} \tilde{A}_0 \overset{\triangle}{V^e} + \tilde{A}_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} V^h + \mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h \Big)\ |J_{\mathcal{E}}| \mathrm{d}\hat{\mathcal{E}} \\
& + \int_{\hat{\mathcal{Q}}} \hat{\psi}_a(\bar{\xi}) \cdot \Big(F_i(V^h) - \bar{n} \cdot v F_0(V^h) - \mathcal{K}_i\ \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h \Big) \bar{n}_i\ |J_{\mathcal{Q}}| \mathrm{d}\hat{\mathcal{Q}} \qquad (6.5.5)
\end{aligned}
$$

for $a = 1, \ldots, n_{en}$.

Similarly, for the test function

$$
\boxed{W^h(t,\bar{x}) = (\lambda_1(\xi_0) - \lambda_2(\xi_0)) \hat{\psi}_a(\bar{\xi}),}
$$

we obtain the following nodal element entry in the nonlinear system of algebraic equations for v and $\tilde{v}$

$$
\begin{aligned}
\tilde{R}_a^e(v,\tilde{v}) = & - \int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot F_0(V^e(\bar{x}))\ |J_{\bar{x}}(1,\bar{\xi})|\ \mathrm{d}\hat{\Omega} - \int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot F_0(V^e_{(n)}(\bar{x}))\ |J_{\bar{x}}(0,\bar{\xi})|\ \mathrm{d}\hat{\Omega} \\
& + \int_{\hat{\mathcal{E}}} \Big(\frac{2}{\Delta t} \hat{\psi}_a(\bar{\xi}) - (1-2\xi_0) \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \Big) \cdot F_0(V^h)\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}} \\
& - \int_{\hat{\mathcal{E}}} (1-2\xi_0) \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot \mathcal{F}^a(V^h)\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}} \\
& + \int_{\hat{\mathcal{E}}} (1-2\xi_0) \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot \mathcal{K}\ \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}} \\
& + \int_{\hat{\mathcal{E}}} \Big(-\frac{2}{\Delta t} \tilde{A}_0 \hat{\psi}_a(\bar{\xi}) \Big) \cdot \tilde{\tau} \\
& \qquad \Big(\frac{2}{\Delta t} \tilde{A}_0 \overset{\triangle}{V^e} + \tilde{A}_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} V^h + \mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} V^h \Big)\ |J_{\mathcal{E}}| \mathrm{d}\hat{\mathcal{E}} \\
& + \int_{\hat{\mathcal{E}}} (1-2\xi_0) \Big(\tilde{A}_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) + \mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \Big) \cdot \tilde{\tau}
\end{aligned}
$$

$$\left(\frac{2}{\Delta t}\tilde{A}_0\overset{\triangle}{V^e} + \tilde{A}_0\bar{\mathcal{I}}_t^T\nabla_{\bar{\xi}}V^h + \mathcal{A}^T\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}V^h - \mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\cdot\mathcal{K}\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}V^h\right)|J_{\mathcal{E}}|\mathrm{d}\hat{\mathcal{E}}$$

$$+\int_{\hat{\mathcal{Q}}}(1-2\xi_0)\hat{\psi}_a(\bar{\xi})\cdot\left(F_i(V^h)-\bar{n}\cdot vF_0(V^h)-\mathcal{K}_i\,\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}V^h\right)\bar{n}_i\;|J_{\mathcal{Q}}|\mathrm{d}\hat{\mathcal{Q}}, \tag{6.5.6}$$

for $a = 1, \ldots, n_{en}$.

6.6 The solution of the nonlinear system

The time-discontinuous Galerkin finite element method requires the solution of a large set of coupled nonlinear algebraic equations. These equations are solved with a Newton method, but for efficiency and robustness the choice of variables in the linearization process is important. In [50], a predictor multi-corrector method is developed in order to reduce the nonlinear system to a sequence of linear systems. In particular two finite element discretizations are considered, constant-in-time and linear-in-time. In this section, we describe a different method of solving the nonlinear equations by considering the structure of the matrices that result from the finite element discretization. In Section 6.6.1 we compare this method with the original algorithm of Shakib [51] using the advection-diffusion equation as a model problem. For convenience, we summarize the main steps of the predictor multi-corrector algorithm defined in [50] in Appendix C.1.

In this section we present an iterative technique, based on a Newton method, to solve the nonlinear algebraic system (6.4.7)-(6.4.8). Define

$$\begin{aligned} R^{(i)} &= G(v^{(i)}, \tilde{v}^{(i)}), \\ \tilde{R}^{(i)} &= \tilde{G}(v^{(i)}, \tilde{v}^{(i)}), \end{aligned} \tag{6.6.1}$$

where $v^{(i)}$ and $\tilde{v}^{(i)}$ are the ith iterative approximation of v and $\tilde{v}$, respectively. Then, the ith approximation of the Jacobian matrix of (6.6.1) can be defined as:

$$M^{(i)} = \begin{pmatrix} M_{11}^{(i)} & M_{12}^{(i)} \\ M_{21}^{(i)} & M_{22}^{(i)} \end{pmatrix},$$

where

$$\begin{aligned} M_{11}^{(i)} &= \frac{\partial G(v^{(i)}, \tilde{v}^{(i)})}{\partial v}, \quad M_{12}^{(i)} = \frac{\partial G(v^{(i)}, \tilde{v}^{(i)})}{\partial \tilde{v}}, \\ M_{21}^{(i)} &= \frac{\partial \tilde{G}(v^{(i)}, \tilde{v}^{(i)})}{\partial v}, \quad M_{22}^{(i)} = \frac{\partial \tilde{G}(v^{(i)}, \tilde{v}^{(i)})}{\partial \tilde{v}}. \end{aligned}$$

In a Newton algorithm we need to solve the following system:

$$\begin{aligned} M_{11}^{(i)}\Delta v^{(i)} + M_{12}^{(i)}\Delta\tilde{v}^{(i)} &= -R^{(i)}, \\ M_{21}^{(i)}\Delta v^{(i)} + M_{22}^{(i)}\Delta\tilde{v}^{(i)} &= -\tilde{R}^{(i)}, \end{aligned} \tag{6.6.2}$$

where $\Delta v^{(i)} = v^{(i+1)} - v^{(i)}$ and $\Delta\tilde{v}^{(i)} = \tilde{v}^{(i+1)} - \tilde{v}^{(i)}$. Re-writing (6.6.1) to this form will help us recognize or investigate the properties associated with the individual blocks of the equation system. The solution of this large linear system is non-trivial and in order to simplify the solution procedure, we first investigate the properties of the individual terms. We can recognize that the matrices M_{12} and M_{21} represent the coupling between the old and new time levels of a space-time slab.

Note that the constant-in-time approximation of the space-time Galerkin least-squares variational equation leads to an algebraic system, having half as many equations and unknowns as the linear-in-time approximation. The constant-in-time approximation has low order of time accuracy, but has good stability properties and is computationally efficient. We use this algorithm in Chapter 7 for solving steady problems.

Assumption 6.6.1 *Assume that the flux Jacobian matrices $\tilde{A}_\ell$, $\ell = 0, \ldots, 3$ and $\tilde{K}_{ij}$, $i, j = 1, 2, 3$ for the symmetrized incompressible Navier-Stokes equations are constant in time in each space-time slab during each Newton iteration step.*

Note that the flux Jacobian matrices are evaluated at the mid-time of the space-time slab, that is using $\bar{V}^e$, and that we update them after each Newton iteration step.

First, observe that the Jacobian matrices M_{11} and M_{12} can be written in a close relation as

$$\begin{aligned} M_{1i}^{ab} = \delta_{1i} &\int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot \tilde{A}_0 \hat{\psi}_b(\bar{\xi}) \; |J_{\bar{x}}(1,\bar{\xi})| \; \mathrm{d}\hat{\Omega} \\ &- \int_{\hat{\mathcal{E}}} \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot \tilde{A}_0 f_i(\xi_0) \; \hat{\psi}_b(\bar{\xi}) \; |J_{\mathcal{E}}| \; \mathrm{d}\hat{\mathcal{E}} \\ &- \int_{\hat{\mathcal{E}}} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot \mathcal{A} f_i(\xi_0) \; \hat{\psi}_b(\bar{\xi}) \; |J_{\mathcal{E}}| \; \mathrm{d}\hat{\mathcal{E}} \\ &+ \int_{\hat{\mathcal{E}}} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \cdot \mathcal{K} \; \mathcal{I}_{\bar{x}}^T f_i(\xi_0) \nabla_{\bar{\xi}} \hat{\psi}_b(\bar{\xi}) \; |J_{\mathcal{E}}| \; \mathrm{d}\hat{\mathcal{E}} \\ &+ \int_{\hat{\mathcal{E}}} \Big(\tilde{A}_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) + \mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \Big) \\ &\quad \cdot \tilde{\tau} \left(\underline{\pm \frac{1}{\Delta t} \tilde{A}_0 \; \hat{\psi}_b(\bar{\xi})} + f_i(\xi_0) \Big[\tilde{A}_0 \bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}} \hat{\psi}_b(\bar{\xi}) + \right. \\ &\qquad \left. + \mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_b(\bar{\xi}) - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_b(\bar{\xi}) \Big] \right) |J_{\mathcal{E}}| \mathrm{d}\hat{\mathcal{E}} \end{aligned}$$

$$+\int_{\hat{\mathcal{Q}}} \hat{\psi}_a(\bar{\xi}) \cdot f_i(\xi_0)\Big(\tilde{A}_i\ \hat{\psi}_b(\bar{\xi}) - \mathcal{K}_i\ \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_b(\bar{\xi})\Big)\bar{n}_i\ |J_{\mathcal{Q}}| \mathrm{d}\hat{\mathcal{Q}} \tag{6.6.3}$$

for $i = 1, 2$, where δ_{ij} is the Kronecker delta, $f_i(\xi_0) = 1 - \lambda_i(\xi_0)$ (i.e. for M_{11}^{ab} the coefficient f_i is ξ_0 and for M_{12}^{ab} it is $(1 - \xi_0)$). The $\pm$ sign in front of the underlined term means that it's sign is $+$ for M_{11}^{ab} and $-$ in case of M_{12}^{ab}.

In the above integrals we used the relation

$$\frac{\partial V^h(t(\xi), \bar{x}(\xi))}{\partial \bar{\xi}_i} = \sum_{a=1}^{n_{en}} \left\{ \xi_0 \frac{\partial \hat{\psi}_a(\bar{\xi})}{\partial \bar{\xi}_i} v_{a;n+1} + (1 - \xi_0) \frac{\partial \hat{\psi}_a(\bar{\xi})}{\partial \bar{\xi}_i} \tilde{v}_{a;n} \right\}, \quad i = 1, 2, 3,$$

and combined with

$$\frac{\partial V^h(t(\xi), \bar{x}(\xi))}{\partial v_a} = \lambda_2(\xi_0)\hat{\psi}_a(\bar{\xi}) = \xi_0 \hat{\psi}_a(\bar{\xi}),$$
$$\frac{\partial V^h(t(\xi), \bar{x}(\xi))}{\partial \tilde{v}_a} = \lambda_1(\xi_0)\hat{\psi}_a(\bar{\xi}) = (1 - \xi_0)\hat{\psi}_a(\bar{\xi}),$$

and

$$\frac{\partial}{\partial v_a}\left(\nabla_{\bar{x}} V^h(t(\xi), \bar{x}(\xi))\right) = \xi_0 \mathcal{I}_{\bar{x}} \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}),$$
$$\frac{\partial}{\partial \tilde{v}_a}\left(\nabla_{\bar{x}} V^h(t(\xi), \bar{x}(\xi))\right) = (1 - \xi_0) \mathcal{I}_{\bar{x}} \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}),$$

for $a = 1, \ldots, n_{en}$ explains the function $f(\xi_0)$. The sign $\pm$ of the underlined term arises from

$$\frac{\partial \overset{\triangle}{V^e}(\bar{x})}{\partial v_a} = \frac{1}{2}\hat{\psi}_a(\bar{\xi}), \quad \frac{\partial \overset{\triangle}{V^e}(\bar{x})}{\partial \tilde{v}_a} = -\frac{1}{2}\hat{\psi}_a(\bar{\xi}).$$

Furthermore, we used

$$\frac{\partial F_\ell(V^h)}{\partial v_a} = \frac{\partial F_\ell(V^h)}{\partial V^h}\frac{\partial V^h}{\partial v_a} = \tilde{A}_\ell(V^h)\xi_0 \hat{\psi}_a(\bar{\xi}),$$
$$\frac{\partial F_\ell(V^h)}{\partial \tilde{v}_a} = \frac{\partial F_\ell(V^h)}{\partial V^h}\frac{\partial V^h}{\partial \tilde{v}_a} = \tilde{A}_\ell(V^h)(1 - \xi_0)\hat{\psi}_a(\bar{\xi}).$$

Similarly, we obtain the matrices M_{21} and M_{22} as follows

$$M_{2i}^{ab} = (\delta_{2i} - 1)\int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot \tilde{A}_0(V^e)\hat{\psi}_b(\bar{\xi})\ |J_{\bar{x}}(1, \bar{\xi})|\ \mathrm{d}\hat{\Omega}$$

$$+\int_{\hat{\mathcal{E}}} \left(\frac{2}{\Delta t}\hat{\psi}_a(\bar{\xi}) - (1 - 2\xi_0)\bar{\mathcal{I}}_t^T \nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\right) \cdot \tilde{A}_0 f_i(\xi_0)\hat{\psi}_b(\bar{\xi})\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}}$$

$$-\int_{\hat{\mathcal{E}}} (1 - 2\xi_0)\mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi}) \cdot \mathcal{A}(V^h) f_i(\xi_0)\hat{\psi}_b(\bar{\xi})\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}}$$

$$+\int_{\hat{\mathcal{E}}} (1 - 2\xi_0)\mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi}) \cdot \mathcal{K}\ \mathcal{I}_{\bar{x}}^T f_i(\xi_0)\nabla_{\bar{\xi}}\hat{\psi}_b(\bar{\xi})\ |J_{\mathcal{E}}|\ \mathrm{d}\hat{\mathcal{E}}$$

$$+\int_{\hat{\mathcal{E}}}\left(-\frac{2}{\Delta t}\tilde{A}_0\hat{\psi}_a(\bar{\xi})+(1-2\xi_0)\left[\tilde{A}_0\bar{\mathcal{I}}_t^T\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})+\mathcal{A}^T\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})-\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\cdot\mathcal{K}\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\hat{\psi}_a(\bar{\xi})\right]\right)$$

$$\tilde{\tau}\left(\underline{\mp\frac{1}{\Delta t}\tilde{A}_0\ \hat{\psi}_b(\bar{\xi})}+f_i(\xi_0)\left[\tilde{A}_0\bar{\mathcal{I}}_t^T\nabla_{\bar{\xi}}\hat{\psi}_b(\bar{\xi})+\mathcal{A}^T\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\hat{\psi}_b(\bar{\xi})-\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\cdot\mathcal{K}\mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\hat{\psi}_b(\bar{\xi})\right]\right)|J_{\mathcal{E}}|\mathrm{d}\hat{\mathcal{E}}$$

$$+\int_{\hat{\mathcal{Q}}}(1-2\xi_0)\hat{\psi}_a(\bar{\xi})\cdot f_i(\xi_0)\Big(\tilde{A}_i(V^h)\ \hat{\psi}_b(\bar{\xi})-\mathcal{K}_i\ \mathcal{I}_{\bar{x}}^T\nabla_{\bar{\xi}}\hat{\psi}_b(\bar{\xi})\Big)\bar{n}_i\ |J_{\mathcal{Q}}|\mathrm{d}\hat{\mathcal{Q}},\tag{6.6.4}$$

where the sign of the underlined term is $-$ for M_{22}^{ab} and $+$ for M_{21}^{ab}.

Let us review some of the main points of the predictor multi-corrector method, summarized in Appendix C.1. This method neglects the use of the matrices M_{12} and M_{21}, that means that the coupling matrices between the old and new time levels are not taken into account. The predictor multi-corrector method in [50], proposed to solve the linear system (6.6.2) iteratively. First, iterations for the $\triangle v^{(i)}$ equation are performed, followed by iterations for the $\triangle\tilde{v}^{(i)}$ equation. Using the notations of Appendix C.1, the matrices M^* and $\tilde{M}^*$ are approximations of M_{11} and M_{22}, respectively. In [50] two different algorithms are proposed, and implicit and an explicit one, but we only consider the first one. In the implicit case two set of matrices are used. The first one is to set $\tilde{M}^* = M^* = (M_{11}-(\mathit{contribution}_1))$, where $\mathit{contribution}_1$ is the underlined term in (6.6.3). The second alternative is: $M^* = (M_{11}-(\mathit{contribution}_1))$, as before, and $\tilde{M}^* = (M_{22}-(\mathit{contribution}_2))$, where $\mathit{contribution}_2$ is the underlined term in (6.6.4). In our computations using the predictor multi-corrector method we only consider the first choice. The benefit of this choice is that we can employ the same left-hand-side matrix for both systems.

Since we are interested in time accuracy using a linear-in-time approximation, a space-time definition of the stabilization matrix is needed. When the least-squares term is added to the linear system and Δt is small, the predictor multi-corrector scheme will not converge with the current definition of the stabilization matrix. Therefore, this method cannot be used for small time steps, which is crucial to obtain time accurate solutions. It is necessary to extend the definition of the stabilization parameter τ_m in Definition 4.4.1 given in the following way:

Definition 6.6.1 *The stabilization parameter τ_m on the space-time elements $\mathcal{E}_n^e$ is defined as*

$$\tau_m(x)=\min\left\{\frac{\Delta t}{\rho},\frac{h_e}{2\rho|u(x)|}\xi(Re_e(x))\right\}\tag{6.6.5}$$

where we used the same notations as in Definition 4.4.1.

Note that the definition of the stabilization parameter τ_e is $\tau_e(x)=\frac{\tau_m}{c_v}$, where τ_m is defined in (6.6.5). For steady state computations the Δt term can be neglected from the definition of the stabilization parameters.

Instead of using the predictor multi-corrector method discussed in Appendix C.1, we also investigate an alternative method which uses the structure of the Jacobian matrix $M^{(i)}$. Let us introduce a different approach to solve the system (6.6.2). Consider $R^{(i)}$ and $\tilde{R}^{(i)}$ in terms of the unknown vectors

$$\bar{v} = (\bar{v}_1^T, \ldots, \bar{v}_{n_{np}}^T)^T,$$
$$\overset{\triangle}{v} = (\overset{\triangle}{v_1}^T, \ldots, \overset{\triangle}{v}_{n_{np}}^T)^T,$$

then, the Jacobian matrix corresponding to the system

$$R^{(i)} = R(\bar{v}^{(i)}, \overset{\triangle}{v}^{(i)}),$$
$$\tilde{R}^{(i)} = \tilde{R}(\bar{v}^{(i)}, \overset{\triangle}{v}^{(i)})$$

becomes

$$\hat{M}^{(i)} = \begin{pmatrix} M_{11}^{(i)} + M_{12}^{(i)} & M_{11}^{(i)} - M_{12}^{(i)} \\ M_{21}^{(i)} + M_{22}^{(i)} & M_{21}^{(i)} - M_{22}^{(i)} \end{pmatrix}.$$

It is appealing to reduce the matrix $\hat{M}^{(i)}$ to a block diagonal form. Under some specific flow conditions, given more precisely later, it can be shown that $M_{12}^{(i)} \approx M_{11}^{(i)}$ and $M_{21}^{(i)} \approx -M_{22}^{(i)}$. The Newton algorithm, then can be simplified by solving the following two linear systems

$$2M_{11}^{(i)} \Delta\bar{v}^{(i)} = -R^{(i)},$$
$$-2M_{22}^{(i)} \Delta\overset{\triangle}{v}^{(i)} = -\tilde{R}^{(i)}, \tag{6.6.6}$$

where $\Delta\bar{v}^{(i)} = \bar{v}^{(i+1)} - \bar{v}^{(i)}$ and $\Delta\overset{\triangle}{v}^{(i)} = \overset{\triangle}{v}^{(i+1)} - \overset{\triangle}{v}^{(i)}$. The solution of the linear systems (6.6.2) is then computed as:

$$\Delta v^{(i)} = \Delta\bar{v}^{(i)} + \Delta\overset{\triangle}{v}^{(i)},$$
$$\Delta\tilde{v}^{(i)} = \Delta\bar{v}^{(i)} - \Delta\overset{\triangle}{v}^{(i)},$$

and this results in a new solution technique, described more precisely in Appendix C.2. Our experience shows that the Newton method described in this section results in a robust discretization technique. In the remaining part of this section we give conditions under which the above approximation is valid, and in the next section we compare the different solution techniques.

The use of this new approach requires, however, that the matrix has a particular structure. Numerical experiments show that when the Reynolds number is small (Re $\approx O(1)$), the "diag" method is not valid, therefore, this solution method cannot be used. For Re $>>$ 1, this method is applicable and useful for real applications. The failure of the method at small Reynolds numbers is due to the fact that the

approximations $M_{12}^{(i)} \approx M_{11}^{(i)}$ and $M_{21}^{(i)} \approx -M_{22}^{(i)}$ are not valid in this regime. This becomes more clear when considering the matrices for the Navier-Stokes equations:

$$M_{11}^{ab} - M_{12}^{ab} = \int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot \tilde{A}_0 \hat{\psi}_b(\bar{\xi}) \; |J_{\bar{x}}(1,\bar{\xi})| \; \mathrm{d}\hat{\Omega}$$
$$+ 2 \int_{\hat{\mathcal{E}}} \left(\mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_a(\bar{\xi}) \right) \cdot \tilde{\tau} \left(\frac{1}{\Delta t} \tilde{A}_0 \; \hat{\psi}_b(\bar{\xi}) \right)$$

$$M_{22}^{ab} + M_{21}^{ab} = \int_{\hat{\Omega}} \hat{\psi}_a(\bar{\xi}) \cdot \tilde{A}_0(V^e) \hat{\psi}_b(\bar{\xi}) \; |J_{\bar{x}}(1,\bar{\xi})| \; \mathrm{d}\hat{\Omega}$$
$$+ 2 \int_{\hat{\mathcal{E}}} \left(-\frac{2}{\Delta t} \tilde{A}_0 \hat{\psi}_a(\bar{\xi}) \right) \cdot \tilde{\tau} \left(\mathcal{A}^T \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_b(\bar{\xi}) - \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \cdot \mathcal{K} \mathcal{I}_{\bar{x}}^T \nabla_{\bar{\xi}} \hat{\psi}_b(\bar{\xi}) \right) |J_{\mathcal{E}}| \mathrm{d}\hat{\mathcal{E}}.$$

From these, it can be concluded that indeed if the Reynolds number is small, then the diffusion matrices dominate in the stabilization operator and $M_{21}^{(i)} \approx -M_{22}^{(i)}$ is not a good approximation.

6.6.1 Evaluation of approximate Newton algorithms

In this section we present some numerical results for the scalar advection-diffusion equation

$$\phi_{,t} + a \cdot \nabla \phi = \mu \phi_{,ii}, \quad i = 1,2,3,$$

on the domain $\Omega = [0,1]^3$, with the velocity $a = (1,0,0)$, and a positive constant μ, using different methods to solve the resulting algebraic system. Dirichlet boundary conditions $\phi = 2$ are applied at the inlet ($x_1 = 0$) and along the side walls and Neumann conditions $\frac{\partial \phi}{\partial x_1} = 0$ at the outlet ($x_1 = 1$). The initial condition is $\phi = 2 - 10x_1(1-x_1)x_2(1-x_2)x_3(1-x_3)$.

We solved the advection-diffusion equation with a Galerkin least-squares discretization. The computational mesh consists of $4 \times 4 \times 4$ uniform hexahedral elements, quadratic finite element basis functions in space and linear-in-time. We compare three methods:

(1) solving the linear system (6.6.2) with a direct method, which we call from now on the "full system" method,

(2) the third-order predictor multi-corrector algorithm (denoted by "pred-mult"), discussed in Appendix C.1,

(3) the approximation method (6.6.6), described in the previous section, that is using the diagonal block matrices M_{11} and M_{22}, which we call "diag" method.

First we compare the convergence to the steady state solution of the three methods. We choose $\mu = 1/100$ and a time step $\Delta t = 0.5$. In Figure 6.2, the normalized residual is plotted for all three methods, as a function of time. The normalized residual at time t is defined as

$$\|\phi(t) - \phi(t - \Delta t)\|_{l_\infty(\Omega)}.$$

For this case all three methods are identical. Both for the "pred-mult" and "diag" methods 10 predictor steps are performed and the linear system is solved with a GMRES algorithm with a tolerance of 10^{-12}.

We observe that if the time step becomes too small (in this particular example when $\Delta t < 0.2$), the predictor multi-corrector algorithm, using Definition 4.4.1 for the stabilization parameters, cannot solve the linear system. The residual of the solution is increasing and the method does not converge. The computed stabilization parameter for this example is $\tau = 0.21$ for each element in the computational mesh. For a time step of $\Delta t = 0.001$, we use the stabilization parameter given in Definition 6.6.1 and recomputed all three methods, see Figure 6.3. We can, however, still observe the instability in the predictor multi-corrector method. This instability is due to the presence of the underlined term in (6.6.3) in the M_{11} matrix. If we omit this term from the M_{11} matrix, then the "pred-mult" method becomes stable and all three methods have the same convergence behavior, see Figure 6.4. Table 6.1 indicates that all three methods, with the proper choice of the stabilization parameter, are

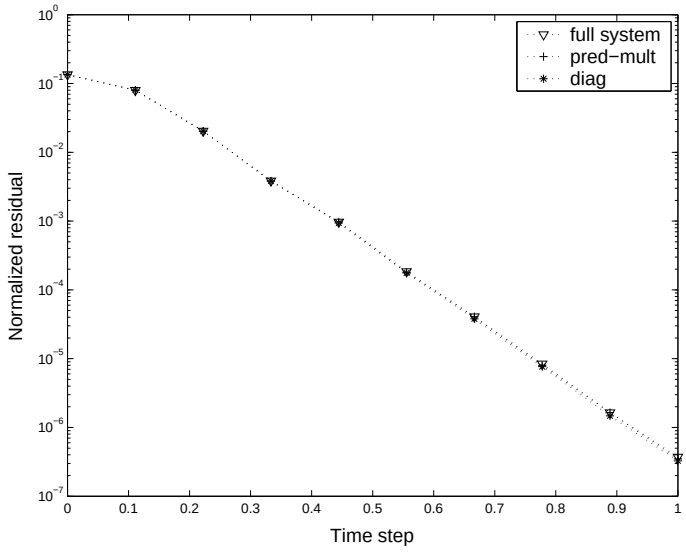

Figure 6.2: Convergence to steady state for three different solution techniques used to solve the linear system for the advection-diffusion equation. $\Delta t = 0.5$ and $\mu = 1/100$.

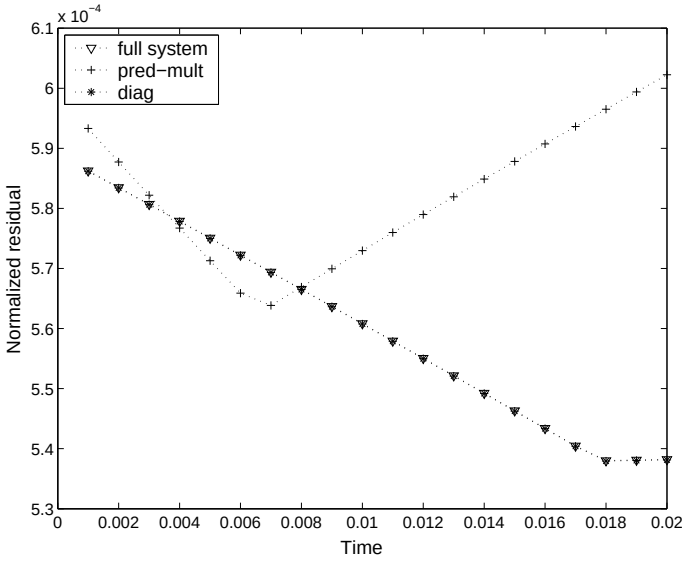

Figure 6.3: Convergence to steady state for three different solution techniques used to solve the linear system for the advection-diffusion equation. $\Delta t = 0.001$ and $\mu = 1/100$.

second-order accurate in time, but that the time dependent part must be removed from the stabilization operator when the predictor multi-corrector scheme is used, which results in an inconsistent finite element discretization.

We also investigate the use of a first order predictor multi-corrector method, denoted by "pred-mult-c", that is, a constant-in-time approximation is used. Figure 6.5 compares the "full system", the "diag" and the "pred-mult-c" methods for a large time step (here the Definition 4.4.1 of τ_m will be used) and in Figure 6.6 a small time step is used for which the stabilization parameter of Definition 6.6.1 is valid.

"full system"	*"pred-mult"*	*"diag"*
1.987963	1.9888002	1.9878829

Table 6.1: Time accuracy, advection-diffusion equation, $\mu = 1/100$.

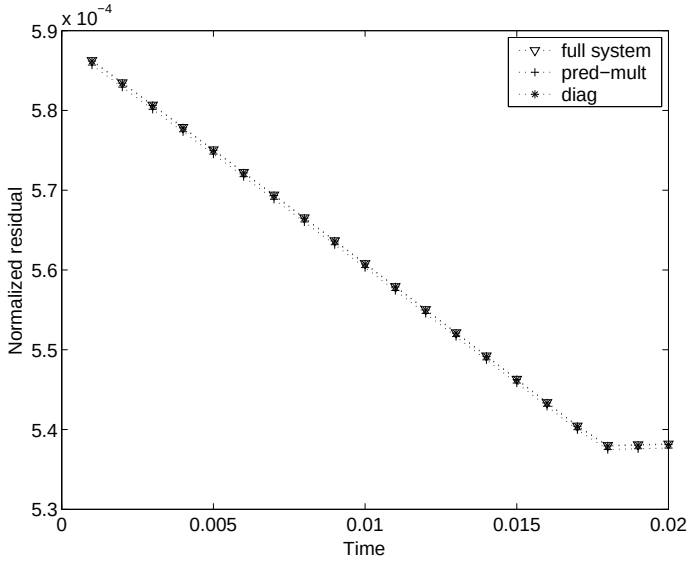

Figure 6.4: Convergence to steady state for three different solution techniques used to solve the linear system for the advection-diffusion equation. $\Delta t = 0.001$ and $\mu = 1/100$.

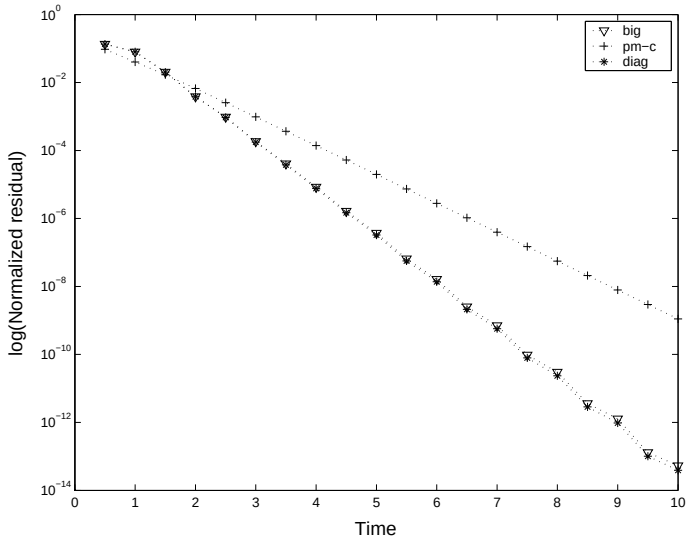

Figure 6.5: Convergence to steady state for three different solution techniques used to solve the linear system for the advection-diffusion equation. $\Delta t = 0.5$ and $\mu = 1/100$.

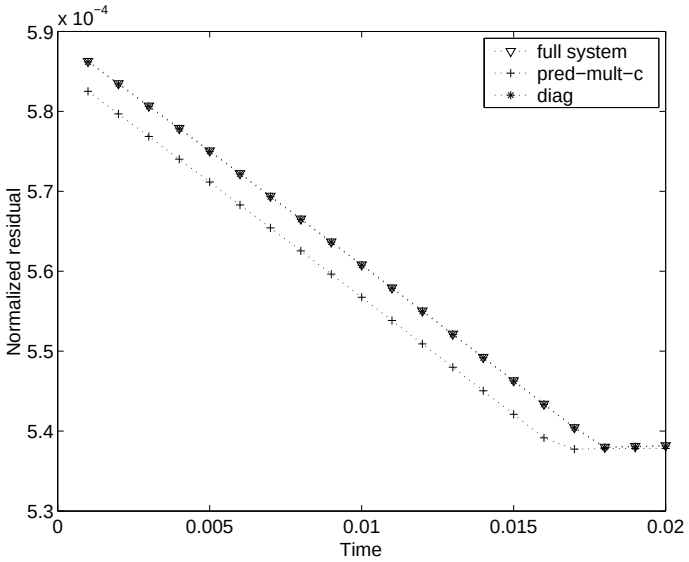

Figure 6.6: Convergence to steady state for three different solution techniques used to solve the linear system for the advection-diffusion equation. $\Delta t = 0.001$ and $\mu = 1/100$.

6.7 Concluding remarks

In this chapter we introduced the time-discontinuous Galerkin least-squares finite element discretization for the incompressible Navier-Stokes equations. The discretization method is described for the symmetrized form of the Navier-Stokes equations, the same formulation can, however, be used for the set of primitive variables (p, u, T), which we use in the next chapter of this thesis. Having established the method for entropy variables, we can transform the variational equation to the pressure primitive variables, which are commonly used for incompressible flows.

The time-discontinuous Galerkin method results in a large system of nonlinear algebraic equations. For the solution of this system we propose a new solution strategy. This new technique is then compared with other solution methods using the advection-diffusion equation as a model problem. More complicated cases will be considered in Chapter 7.

Chapter 7

Numerical examples

This chapter discusses several test cases and applications to verify and demonstrate the Galerkin least-squares finite element method for the incompressible Navier-Stokes equations. The main emphasis is to employ the newly designed stabilization matrices and to compare the results with existing and well studied results.

The computations presented in this chapter were performed on a personal computer. In the construction of the finite element code, we used the deal.II library. Deal.II is a C++ program library [2] which provides an excellent interface to handle complex data structures and algorithms. It is well suited for mesh adaptation, enables the use of a variety of finite elements in several space dimensions (continuous or discontinuous), and also includes higher order elements.

7.1 Channel flow

The main difficulty in solving the incompressible Navier-Stokes equations is their nonlinearity arising from the convective acceleration terms, i.e. $u_i \frac{\partial u_j}{\partial x_i}$. There are, however, a few special cases for which the convective acceleration vanishes due to the nature of the geometry of the flow. In these cases it is possible to find the exact solution of the Navier-Stokes equations. One of these examples is the so-called Poiseuille flow, which we discuss in this section. Since our computer program is a fully three dimensional code, we verify it on two examples, which are similar to the Poiseuille flow.

Let us introduce some notations. A point $x \in \mathbb{R}^3$ has coordinates $x = (x_1, x_2, x_3)$ and $u_i,\ i = 1, 2, 3$ denote the components of the velocity vector u in the ith Cartesian coordinate direction. Consider a two dimensional steady flow between two infinite parallel flat walls. The fluid particles are moving in the x_1-direction, parallel to the plates, and there is no velocity in the x_2 or x_3-direction, that is $u_2 \equiv 0$ and $u_3 \equiv 0$. In

this case, it follows from the continuity equation that $\frac{\partial u_1}{\partial x_1} = 0$. Furthermore, there is no variation of u in the $z-$direction for infinite plates. Therefore, $u_3(x)$ is a function of x_2 only. If these conditions are used in the Navier-Stokes equations, we obtain for the equations of motion the following system:

$$0 = -\frac{1}{\rho}\frac{\partial p}{\partial x_1} + \frac{\nu}{\mathrm{Re}}\frac{\partial^2 u_1}{\partial x_2^2}. \tag{7.1.1}$$

$$0 = -\frac{1}{\rho}\frac{\partial p}{\partial x_2} \tag{7.1.2}$$

$$0 = -\frac{1}{\rho}\frac{\partial p}{\partial x_3}. \tag{7.1.3}$$

It follows that $p(x)$ is a linear function of x_1 only, and combined with (7.1.1), we obtain

$$u_1(x) = \frac{\mathrm{Re}}{2\mu}\left(\frac{\partial p}{\partial x_1}\right) x_2^2 + c_1 x_2 + c_2, \quad c_1, c_2 \in \mathbb{R}.$$

Note that the pressure gradient is constant. The two constants are determined from the boundary conditions.

- at $x_2 = 0$ and $x_2 = b$

$$u_1 = 0 \Longrightarrow c_2 = 0, \quad c_1 = -\frac{\mathrm{Re}}{2\mu}\left(\frac{\partial p}{\partial x_1}\right) b.$$

Therefore, the fluid motion is given by

$$\boxed{u_1(x) = \frac{\mathrm{Re}}{2\mu}\left(\frac{\partial p}{\partial x_1}\right) x_2(x_2 - b).} \tag{7.1.4}$$

The pressure variation throughout the fluid can be obtained from

$$p(x) = \left(\frac{\partial p}{\partial x_1}\right) x_1 + p_0, \tag{7.1.5}$$

where p_0 is the reference pressure.

Let us compute the solution of the temperature equation, which for the Poiseuille flow is given by

$$\frac{\kappa}{\mathrm{Pr}\,\mathrm{Ec}}\frac{\partial^2 T}{\partial x_2^2} = -\mu\left(\frac{\partial u_1}{\partial x_2}\right)^2. \tag{7.1.6}$$

Solving the above ODE, we obtain

$$T(x) = -\frac{C}{48}(2x_2 - b)^4 + c_1 x_2 + c_2, \quad \text{with} \quad C = \frac{\mathrm{Re}^2\,\mathrm{Pr}\,\mathrm{Ec}}{4\mu\kappa}\left(\frac{\partial p}{\partial x_1}\right)^2. \tag{7.1.7}$$

Assuming equal temperatures of the walls, that is

- at $x_2 = 0$ and $x_2 = b$: $\quad T = T_0$,

the coefficients c_1 and c_2 can be specified and we obtain that the temperature distribution is represented by a parabola of degree four:

$$\boxed{T(x) - T_0 = \frac{Cb^4}{48}\left[1 - \left(2\frac{x_2}{b} - 1\right)^4\right]}, \quad \text{with} \quad C = \frac{\mathrm{Re}^2 \,\mathrm{Pr}\,\mathrm{Ec}}{4\mu\kappa}\left(\frac{\partial p}{\partial x_1}\right)^2. \tag{7.1.8}$$

For the Poiseuille flow the boundary conditions consist of:

Periodic boundary: We assume that the flow is periodic in the x_3-direction.

No-slip flow boundary conditions: The no-slip walls are:

- $x_2 = 0$ and $x_2 = b$,

along which the velocity and the temperature are prescribed, that is $u = (0, 0, 0)^T$ at the walls and both walls have the same temperature, $T = T_0$.

Inflow boundary conditions: The inflow boundary is

- $x_1 = 0$,

where the velocity field is prescribed, $u = (u_1(x), 0, 0)^T$, with $u_1(x)$ given in (7.1.4), and the temperature field is set equal to the initial temperature $T = T_0$. The variation in temperature in the flow domain is due to friction only.

Outflow boundary conditions: The outflow boundary is

- $x_1 = L$,

where L is the length of the channel. We set the temperature at the outflow boundary equal to its exact value given in (7.1.8). The flow is characterized by a decreasing linear pressure field. By prescribing the pressure gradient and initializing the pressure to p_0, the pressure at the outflow boundary can be obtained as

$$p_{\text{out}} = \left(\frac{\partial p}{\partial x_1}\right) L + p_0. \tag{7.1.9}$$

This pressure is prescribed at the outflow boundary and enters in the advective fluxes over the outflow boundary. Furthermore, in the diffusive fluxes we insert the following constraints:

$$n^T S r = 0 \tag{7.1.10}$$

$$n^T S s = 0 \tag{7.1.11}$$

where r and s are the two tangential vectors at the outflow boundary and the matrix $S = (s_{ij})$, for $i, j = 1, 2, 3$, contains the stresses $s_{ij} = u_{i,j} + u_{j,i}$. For our particular test case, the constraints (7.1.10-7.1.11) reduce to

$$n^T S r = s_{12} = u_{1,2} + u_{2,1} = u_{1,2} = 0$$
$$n^T S s = s_{13} = u_{1,3} + u_{3,1} = 0$$

where the last equality is straightforwardly satisfied due to the periodicity in the x_3-direction.

7.1.1 Verification of the numerical method

Poiseuille flow. In this example we verify the numerical solution of the Poiseuille flow obtained by solving the incompressible Navier-Stokes equations using the newly defined stabilization matrix τ_Y for primitive variables (4.3.13). The Reynolds number, which is based on the maximum velocity and channel width, was set to $\mathrm{Re} = 1$. We study the influence of the parameter ω in the stabilization matrix τ_Y on the convergence of the solution to steady state and on the accuracy of the numerical method. In these computations we have used *linear* basis functions in space and constant-in-time. Note that the numerical scheme is exact for the velocity and pressure if quadratic basis functions are used, which was also confirmed by the numerical simulations.

The Poiseuille flow is solved on a unit cube $[0, 1]^3$. Dirichlet conditions are applied at the inlet and along the solid walls and periodic boundary conditions at the side wall, see the introduction to Section 7.1. The computational domain is subdivided into equidistant elements and the convergence of the solution to steady state is observed when different values of ω are employed in the stabilization operator. Let us recall that for the linearized incompressible Navier-Stokes equations, we obtained that the Galerkin least-squares method is stable when the ω parameter satisfies:

$$\omega \in \left(\frac{-1 - \sqrt{5}}{2}, \frac{-1 + \sqrt{5}}{2} \right).$$

Therefore, we choose some specific values of ω in this interval, for example, one close to the minimum and the other to the maximum of admissible values. When $\omega = 0$, the stabilization matrix is identical to the one discussed in [23]. We also introduce the diagonal stabilization matrix, which is the stabilization used in [13]. In Figure 7.1, we observe that for the different stabilization matrices the solution converges to steady state at approximately the same speed when the element size is equal to $h = 1/16$. The algebraic system was solved with the first order predictor multi-corrector method, described in Section 6.6. The resulting linear system was solved in each iteration of the multi-corrector method with the GMRES iterative solver [2].

The first important remark is that for $\omega = 0.6$, GMRES is more robust and needs much less iterations to converge with the same tolerance than for other values of ω.

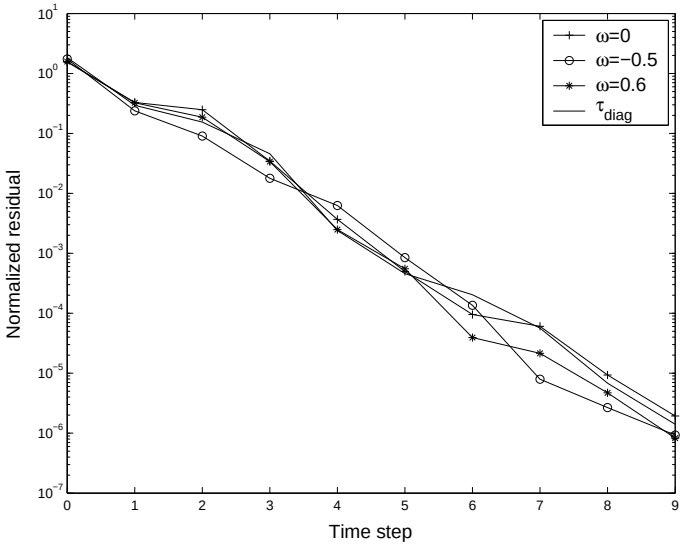

Figure 7.1: Poiseuille flow, residual convergence for different values of ω, $h = 1/16$, $\mathrm{Re} = 1$.

The same robustness of the iterative solver is observed when the mesh is refined, but for $\omega = 0.6$ the solution now converges much faster to steady state, see Figure 7.2. Next, we studied what is the influence of ω on the spatial accuracy. In Table 7.1, some values of the L^2-norm of the error are given, which are plotted in Figure 7.3. We conclude that the ω parameter does not influence the accuracy of the method and that the algorithm converges with second order accuracy to the exact solution.

A square channel. In this example we validate the performance of our computer program in three space dimensions. Consider the following divergence free initial velocity field for the incompressible Navier-Stokes equations in the domain $\Omega = [0, 1]^3$,

Element size	$\omega = -0.5$	$\omega = 0$	$\omega = 0.6$
1/10	0.0956049	0.0958273	0.0956631
1/16	0.0436049	0.0413225	0.043563
1/20	0.0290783	0.0290625	0.0290509
1/25	0.0201648	0.0190827	0.0191477

Table 7.1: L^2 error for Poiseuille flow computations as a function of mesh size, $\mathrm{Re} = 1$.

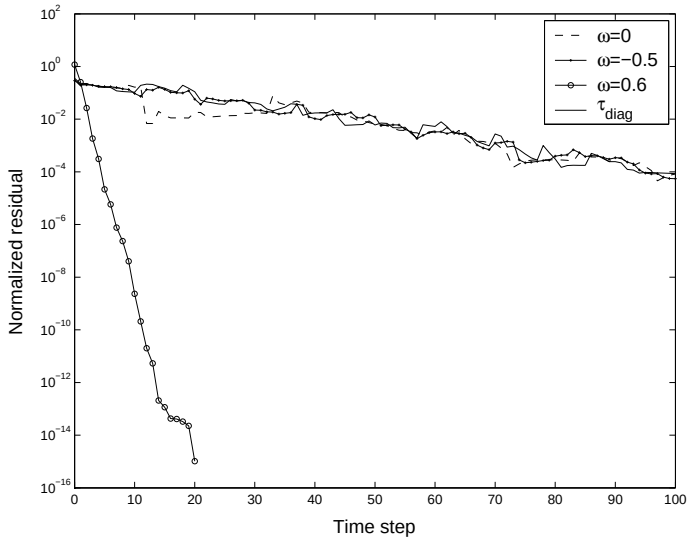

Figure 7.2: Poiseuille flow, residual convergence for different values of ω, $h = 1/25$, $\mathrm{Re} = 1$.

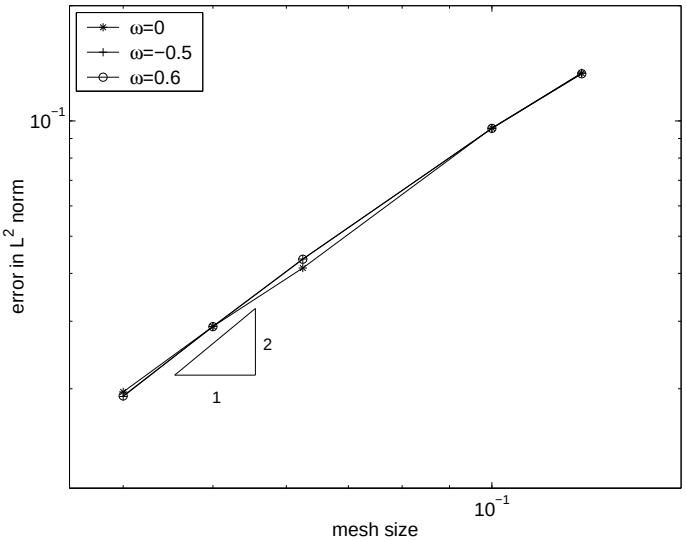

Figure 7.3: Convergence of the Poiseuille flow for different values of ω under mesh refinement, $\mathrm{Re} = 1$.

which is also specified at the inflow ($x_1 = 0$) boundary

$$u_1 = 16x_2(1 - x_2)x_3(1 - x_3), \tag{7.1.12}$$

$$u_2 = u_3 = 0. \tag{7.1.13}$$

At the outflow boundary ($x_1 = 1$) the pressure p is prescribed as well as the natural boundary conditions $\frac{\partial u_1}{\partial n} = 0, \frac{\partial u_2}{\partial n} = 0, \frac{\partial u_3}{\partial n} = 0$, with n the unit outward normal vector. At the four side walls we impose the no-slip boundary condition $u = 0$. The time step is initially set equal to $\Delta t = 0.1$ and since we are interested in the convergence to steady state, we gradually increase the time step till we reach the steady state. The test is performed on a $4 \times 4 \times 4$ mesh. In this approximation, the interpolation functions are *quadratic* in space and constant-in-time. We give examples for two values of the channel length. The Reynolds number is based on the maximum inflow velocity and the width of the channel.

(a) Consider Re = 1. In Figure 7.4 the convergence to steady state of the solution is shown on a logarithmic scale when the non-diagonal stabilization matrix (4.3.13) with $\omega = 0.6$, and the diagonal stabilization matrix $\tau_{diag} = \text{diag}\ (\tau_c, \tau_m, \tau_m, \tau_m)$ are used. Both stabilization operators give essentially the same convergence to steady state. Note that the choice for $\omega = 0.6$ is due to its nice properties we experienced in the previous example.

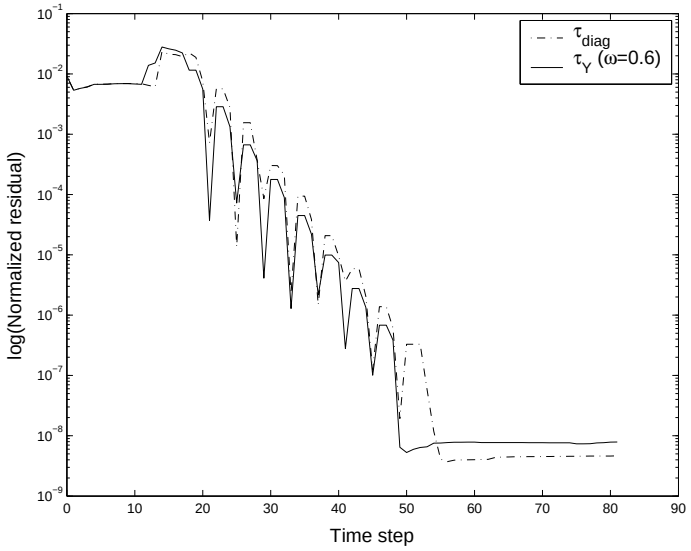

Figure 7.4: Residual convergence for τ_{diag} and τ_Y for $\omega = 0.6$, with Re = 1.

Note that $u_{\max} = 1$ at the inflow and from the momentum equations, it is straightforward to see that the pressure field is linear in x_1. Figure 7.5 shows the velocity and pressure contours at $x_3 = 1/2$. It can be seen that both the velocity field becomes two dimensional further downstream, with a vanishing u_2 component, and the pressure field is linear.

(b) Consider the same test case as in (a) where we double the length of the channel in the streamwise direction. The convergence to steady state is nearly identical to case (a), velocity and pressure contours are shown in Figure 7.6 and 7.7, respectively.
In Figure 7.8 we plot the vertical velocity close to the outflow boundary in a cross section at $x_3 = 0.5$, for both case (a) and (b). This plot shows that a significant channel length is required to reduce the effect of inflow boundary conditions.

"Stokes flow". In this example we verify the accuracy of the method when higher order polynomial basis functions are used and study how the accuracy is influenced by the use of different stabilization matrices.

Consider the three dimensional domain $[0,1]^3$ with boundary $\partial\Omega$, and the following divergence free velocity field

$$u_1 = \phi(x_2, x_3) \tag{7.1.14}$$

$$u_2 = 0 \tag{7.1.15}$$

$$u_3 = 0, \tag{7.1.16}$$

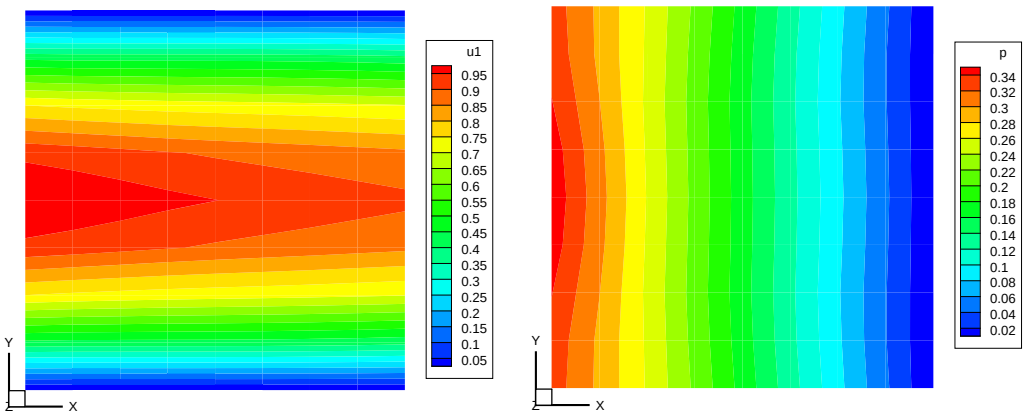

Figure 7.5: Velocity and pressure contours in the channel $[0,1]^3$ at $\mathrm{Re} = 1$.

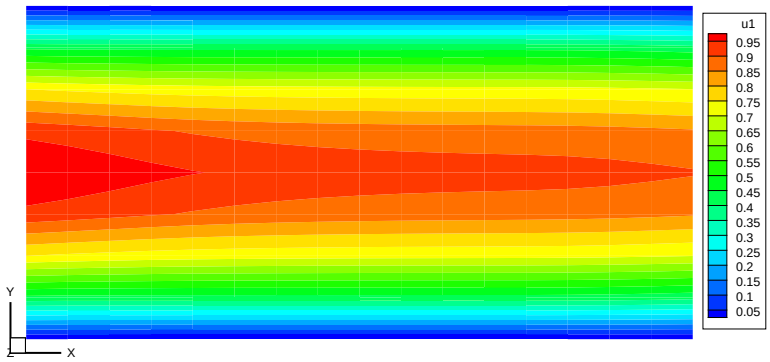

Figure 7.6: Velocity contours for the channel $[0, L] \times [0, 1]^2$ for $\mathrm{Re} = 1$, $L = 2$.

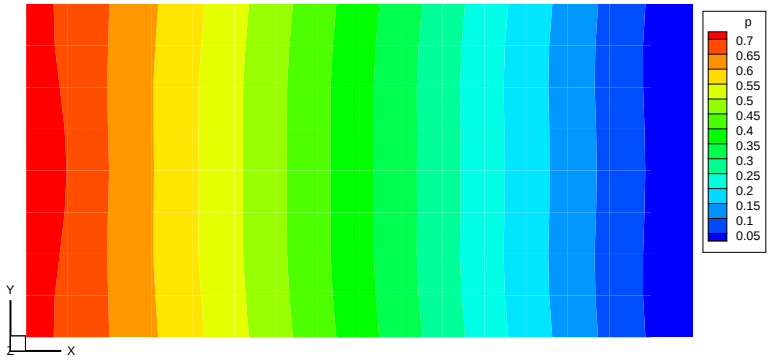

Figure 7.7: Pressure contours for the channel $[0, L] \times [0, 1]^2$ for $\mathrm{Re} = 1$, $L = 2$.

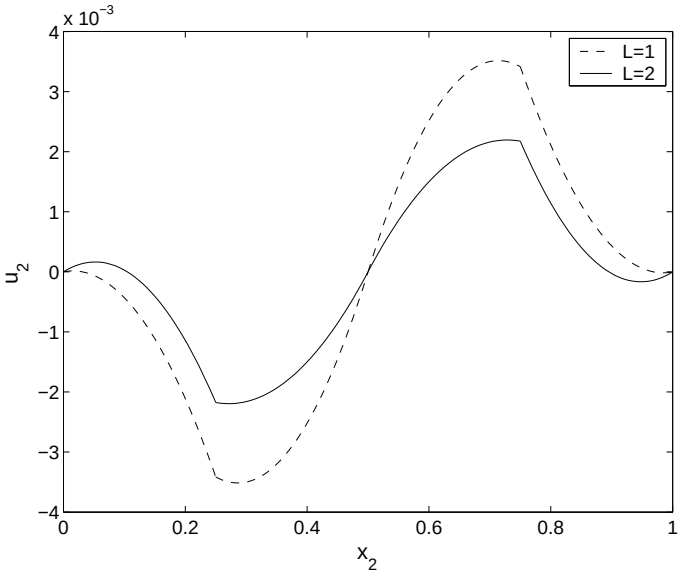

Figure 7.8: Vertical velocity along the outflow boundary at $x_3 = 0.5$ for the square channel of length $L = 1$ and $L = 2$, respectively, at $\mathrm{Re} = 1$.

with

$$\phi(x_2, x_3) = C\left[\frac{x_3(1-x_3)}{2} - \frac{1}{2}\sum_{n=0}^{\infty}\frac{(-1)^n\cos(\alpha_n(2x_3-1))\cosh(\alpha_n(2x_2-1))}{\alpha_n^3\cosh(\alpha_n)}\right] \tag{7.1.17}$$

where $C > 0$ is a constant and

$$\alpha_n = \left(n + \frac{1}{2}\right)\pi.$$

This field is the exact solution of the boundary value problem

$$C\Delta u_1 = -1 \text{ in } \Omega = [0,1]^3$$
$$u_1 = 0 \text{ on } \Gamma_1,$$

where $\partial\Omega = \Gamma_1 \cup \Gamma_2$, with $\Gamma_1 = \{x_2 = 0\} \cup \{x_2 = 1\} \cup \{x_3 = 0\} \cup \{x_3 = 1\}$.

It is straightforward to verify that a constant pressure gradient in the x_1 direction is obtained

$$p(x) = -\frac{1}{C}x_1 + p_0,$$

which together with the velocity field $u = (\phi(x_2, x_3), 0, 0)^T$ and ϕ given in (7.1.17), is an exact solution of the incompressible Navier-Stokes equations.

The boundary conditions are chosen such that the velocity field (7.1.14-7.1.16) is prescribed along the inflow boundary ($x_1 = 0$), $u = (0, 0, 0)^T$ on Γ_1, a constant pressure p is given at the outflow boundary and all normal derivatives of the velocities vanish at the outflow boundary.

For all cases $\mathrm{Re} = 1$ and $C = 28$ were used. In this approximation, the finite element basis functions are linear and quadratic in space and constant-in-time.

Consider the stabilization matrix τ_Y in (4.3.13) as a function of ω. In Figure 7.9, the convergence of the solution to steady state is plotted on a logarithmic scale for $\omega = 0, -0.5$ and 0.6 on a $4 \times 4 \times 4$ mesh when quadratic basis functions are used. The plot shows that the parameter ω does not influence the convergence to the steady solution for higher order polynomial basis functions. The performance of the GMRES method is observed to be more robust when $\omega = 0.6$. Therefore, we verified for this value of ω the accuracy of the method for linear and quadratic polynomial basis functions. We conclude that this choice of ω does not degrade accuracy, see Figure 7.10.

7.2 Driven cavity flow

The driven cavity flow is a classical problem to test the performance of numerical methods for incompressible flows. The top boundary of a square domain slides in

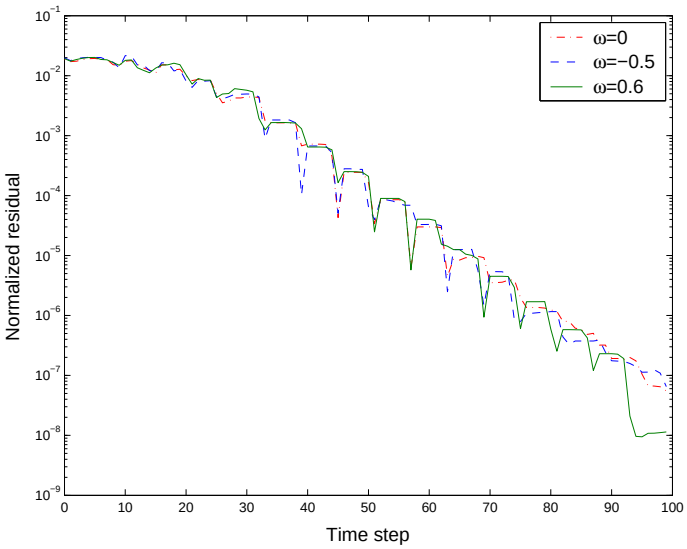

Figure 7.9: Convergence to steady state for the "Stokes flow", $h = 1/4$, $\mathrm{Re} = 1$.

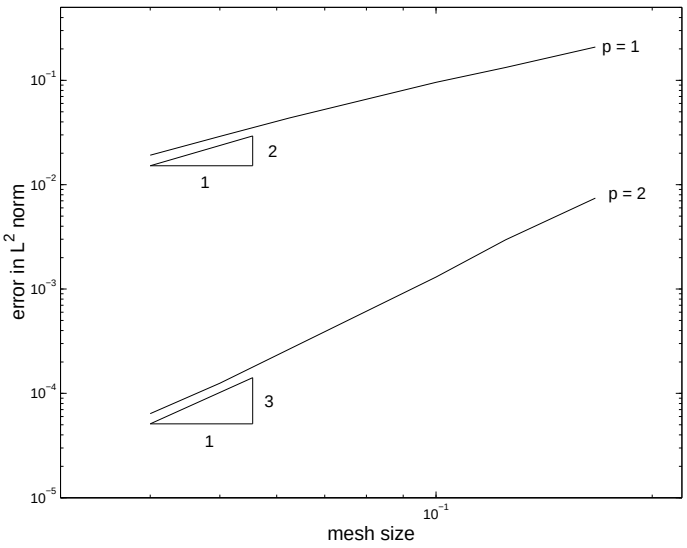

Figure 7.10: Spatial accuracy under p-refinement for the "Stokes flow", with $\omega = 0.6$ in the stabilization matrix τ_Y, Re $= 1$.

the x_1-direction with a velocity of $u = (1, 0, 0)^T$ causing recirculation inside the cavity. The velocity field is discontinuous at the top corners, therefore, these points are singularities. This problem is challenging since the numerical method should be able to control these singularities and simultaneously represent the smooth regions of the flow accurately. Note that, since our computational domain is three dimensional, $[0, 1]^3$, we prescribe periodic boundary conditions in the third direction.

First, consider Re $= 1$, isothermal incompressible flow, computed using a mesh of square elements, linear polynomial basis functions in space and constant in time. For this Reynolds number, a main vortex develops in the center of the cavity. Since primitive variables are used in these computations, the stabilization matrix τ_Y, given in (4.3.13) is employed with $\omega = 0.6$. We started our computations using a uniform 30×30 mesh and compared the results with a computation using the same number of elements but with grid clustering near the walls. The main benefit of the clustered mesh is that the discontinuities at the two top corners are much better represented. In order to better capture the details of the flow, we further refined the mesh, using now a 40×40 grid. Figures 7.11 and 7.12 show for a uniform and a clustered mesh, respectively, the pressure and vertical velocity contours and Figures 7.13 and 7.14 show the corresponding horizontal velocity contours and streamlines. Figure 7.15 shows the convergence to the steady state, where m is defined as

$$m = \max_{i=1}^{n_{dofs}} |V^{n+1}(i) - V^n(i)|,$$

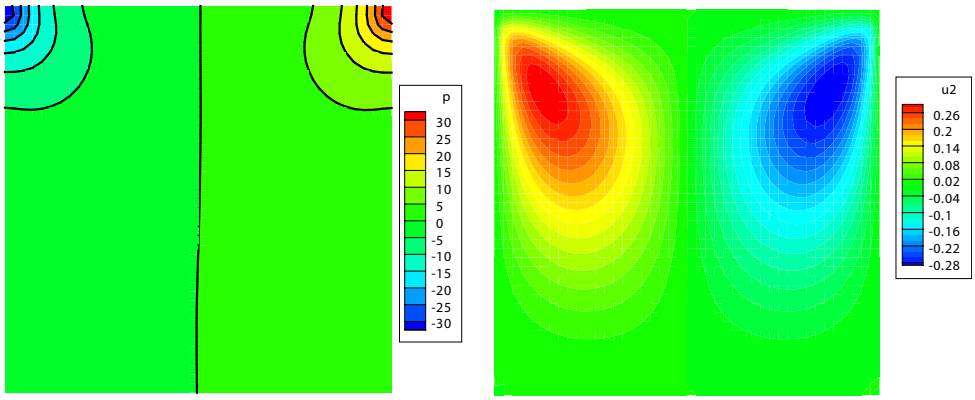

Figure 7.11: Pressure p and vertical velocity contours u_2 for the driven cavity flow on a 40×40 uniform mesh, $\mathrm{Re} = 1$.

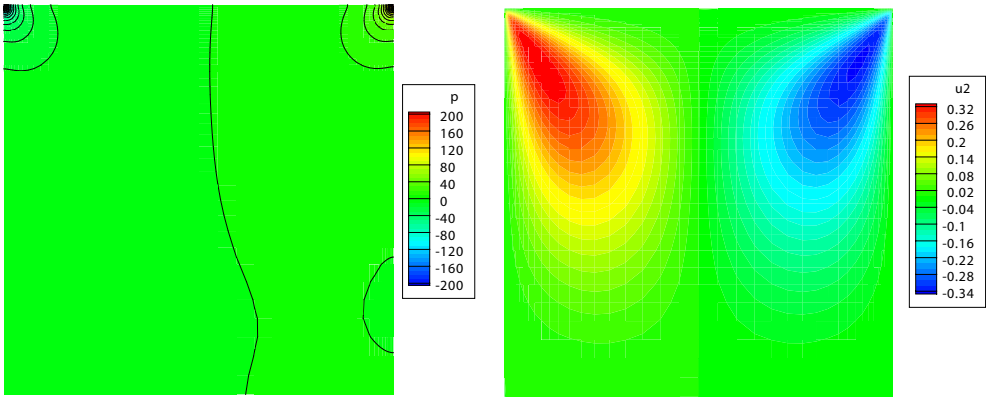

Figure 7.12: Pressure p and vertical velocity contours u_2 for the driven cavity flow on a 40×40 clustered mesh, $\mathrm{Re} = 1$.

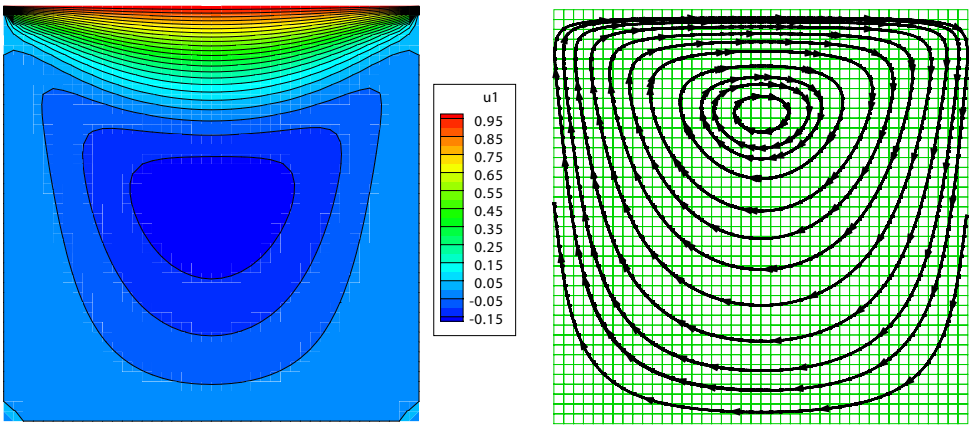

Figure 7.13: Horizontal velocity contours u_1 and streamlines for the driven cavity flow on a 40×40 uniform mesh, $\mathrm{Re} = 1$.

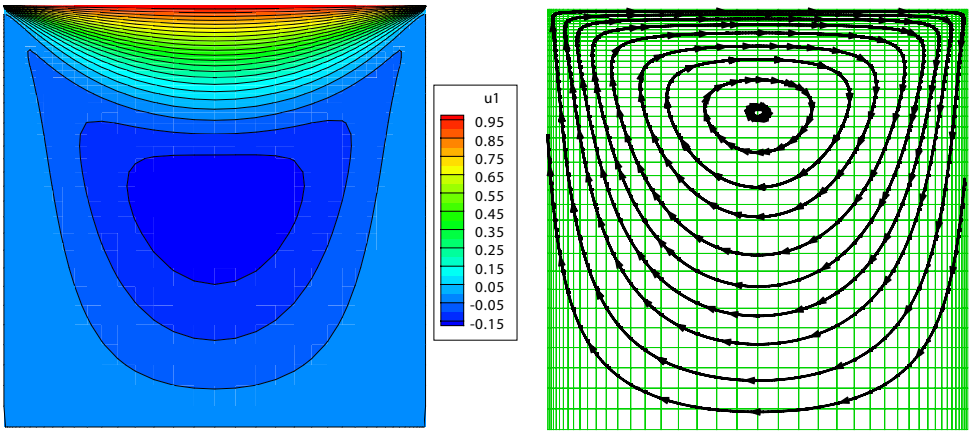

Figure 7.14: Horizontal velocity contours u_1 and streamlines for the driven cavity flow on a 40×40 clustered mesh, $\mathrm{Re} = 1$.

with n_{dofs} the number of degrees of freedom in the finite element discretization and V^{n+1} and V^n the solution at the time levels t_{n+1} and t_n, respectively.

Since the pressure is discontinuous at the upper corners, the difference between the maximal and minimal values of the pressure is a good indication for the performance of the numerical method on the current computational mesh. The larger the pressure difference, the better the discontinuities are represented. In Table 7.2, we listed the pressure difference for the various meshes we considered and we also included the result obtained in [23]. We conclude that mesh clustering has a significant effect on capturing the singularities.

Consider now a Re = 400, isothermal driven cavity flow, using a 50×50 clustered grid. For this Reynolds number, in addition to the main vortex in the center of the cavity, a secondary eddy appears in the right bottom corner of the domain, see Figures 7.16 and 7.17. The results of this computation are compared with the results of Ghia et al. in [16], which were performed on a uniform grid with 129×129 points. In Figures 7.18 and 7.19 we compare the results for the velocity values for lines passing through the geometric center of the cavity. The comparison shows that the present results correlate well with the results obtained in [16] and for example some typical points, such as local minima or maxima are well represented.

When we further increase the Reynolds number to Re = 800, two new eddies appear in addition to the center vortex. In this test case we used the same 50×50 clustered mesh as for the Re = 400 case. We employed a linear-in-time approximation and solved the resulting algebraic system with the predictor multi-corrector method, described in Section 6.6. The convergence of the solution to the steady state is plotted in Figure 7.20. For this Reynolds number case we illustrate in Figures 7.21-7.25 the velocity, pressure, vorticity contours and the streamlines, respectively. These results show that the secondary vortex in the lower left corner is also well captured, including the thin shear layers present in the flow.

Mesh	*Number of elements*	*pressure variation*
uniform	30×30	68
clustered	30×30	270
uniform, [23]	40×40	347
clustered	40×40	412

Table 7.2: Pressure difference for the driven cavity flow using different meshes at Re = 1.

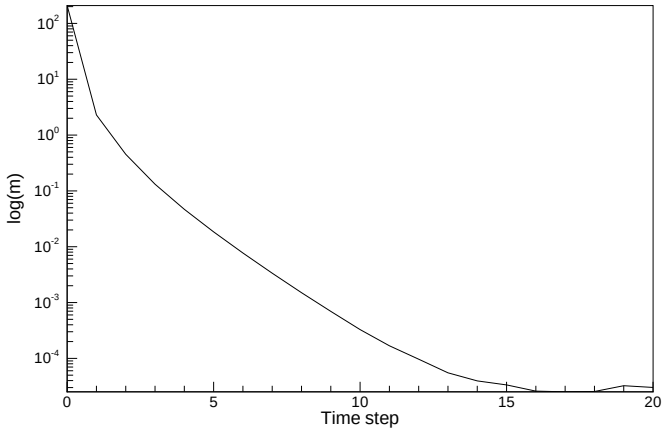

Figure 7.15: Convergence to steady state for the driven cavity flow on a 40×40 clustered mesh, $\mathrm{Re} = 1$.

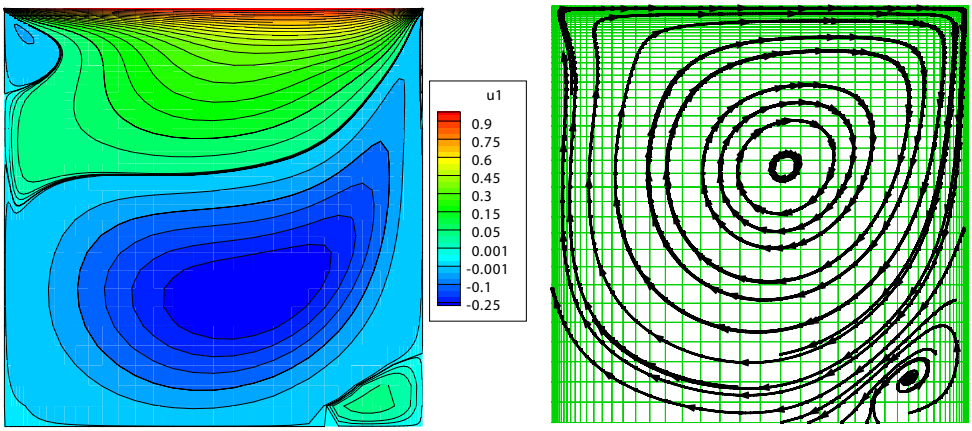

Figure 7.16: Horizontal velocity contours u_1 and streamlines for the driven cavity flow on a 50×50 clustered mesh, $\mathrm{Re} = 400$.

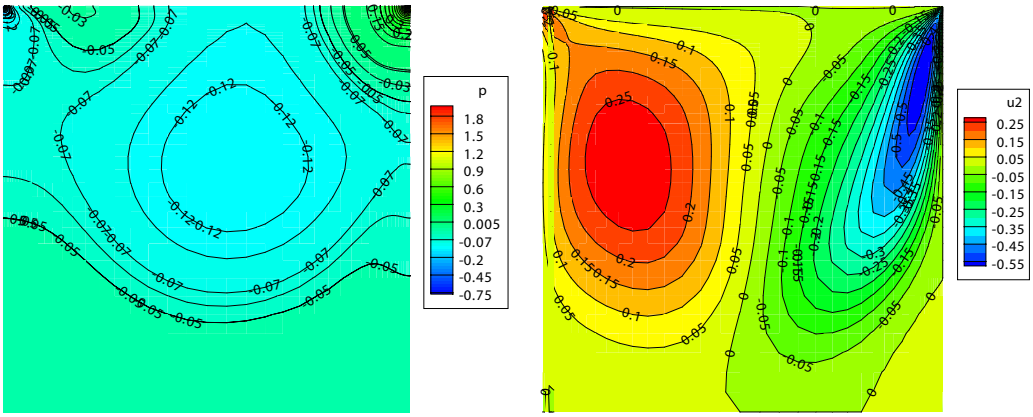

Figure 7.17: Pressure p and vertical velocity contours u_2 for the driven cavity flow on a 50×50 clustered mesh, Re $= 400$.

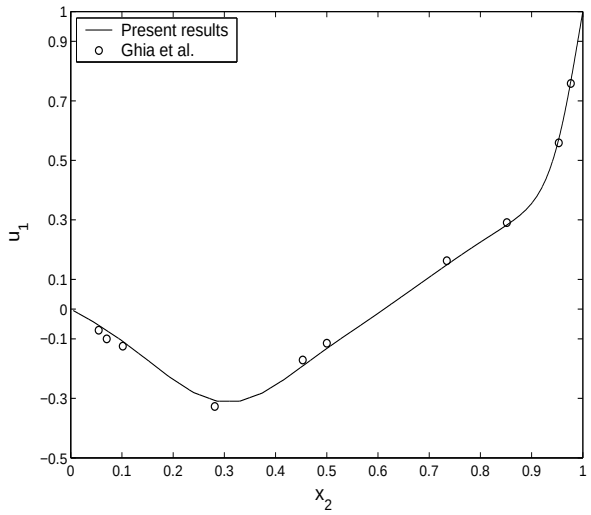

Figure 7.18: Comparison of driven cavity flow results with those obtained in [16] for values of the horizontal velocity along $x = 0.5$ at Re $= 400$.

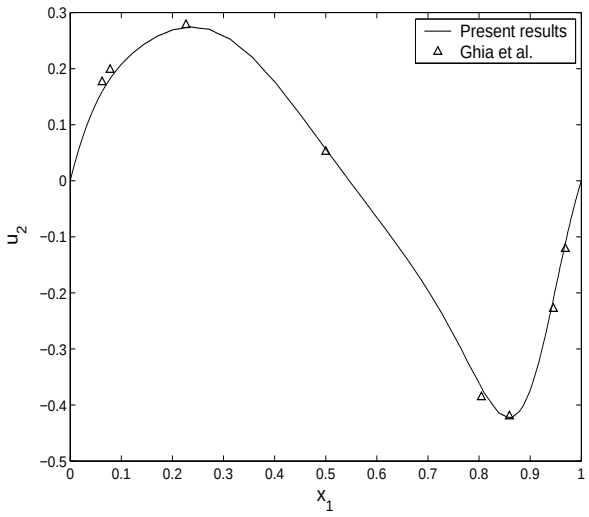

Figure 7.19: Comparison of driven cavity flow results with those obtained in [16] for values of the vertical velocity along $y = 0.5$, $\mathrm{Re} = 400$.

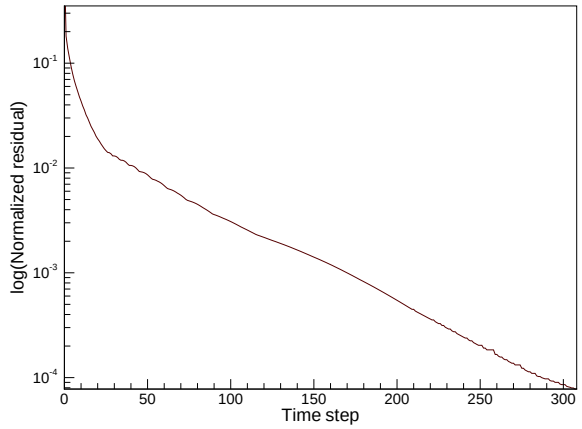

Figure 7.20: Convergence of the driven cavity flow on a 50×50 clustered mesh, $\mathrm{Re} = 800$.

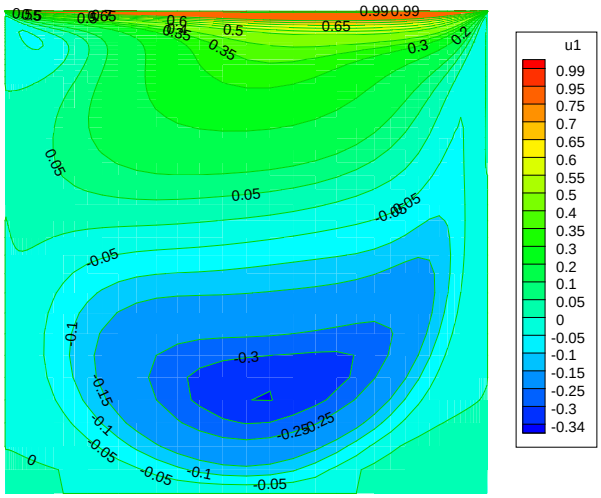

Figure 7.21: Horizontal velocity contours of the driven cavity flow on a 50×50 clustered mesh, $\mathrm{Re} = 800$.

7.3 Flow past a circular cylinder

Simulation of the flow past a circular cylinder is a challenging problem for numerical solution methods. Depending on the Reynolds number, the flow around the cylinder can have different characteristics. At Reynolds numbers less than or equal to about 40, the flow is steady and separates. A pair of symmetrical counter rotating eddies forms downstream of the cylinder. As the Reynolds number increases, the eddies become unstable and periodic vortex shedding occurs. The eddies are then transported downstream and result in the so-called Karman vortex street.

The periodic shedding of vortices is very important from the engineering point of view, since vortex shedding can induce significant structural vibrations and loads. These engineering problems occur at high Reynolds numbers where the flow is turbulent. In this thesis we do not consider high Reynolds number flow, it is, however, important to verify if the numerical method is capable to accurately represent the flow details at moderate Reynolds numbers.

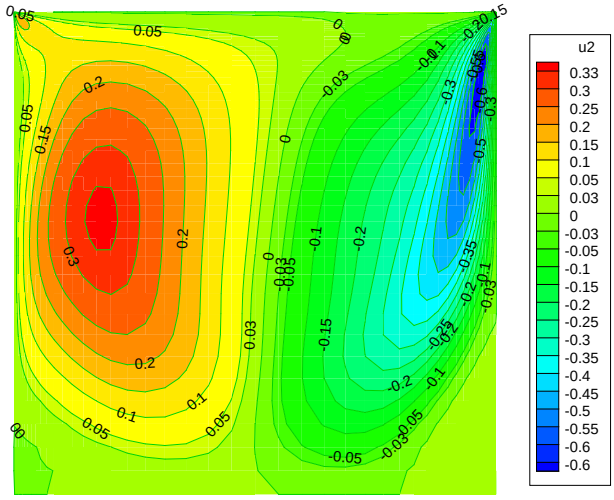

Figure 7.22: Vertical velocity contours of the driven cavity flow on a 50×50 clustered mesh, $\mathrm{Re} = 800$.

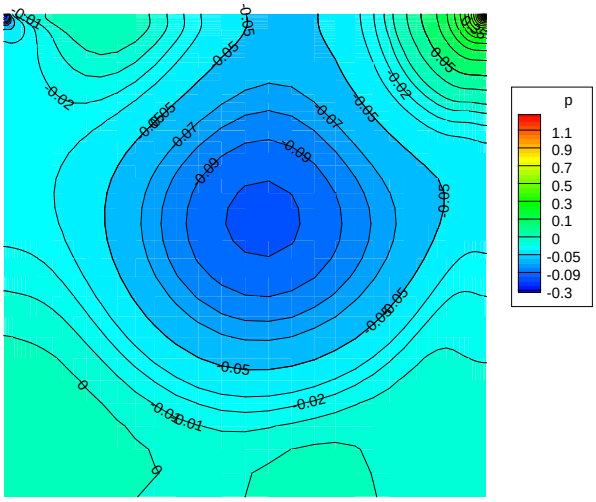

Figure 7.23: Pressure contours of the driven cavity flow on a 50×50 clustered mesh, $\mathrm{Re} = 800$.

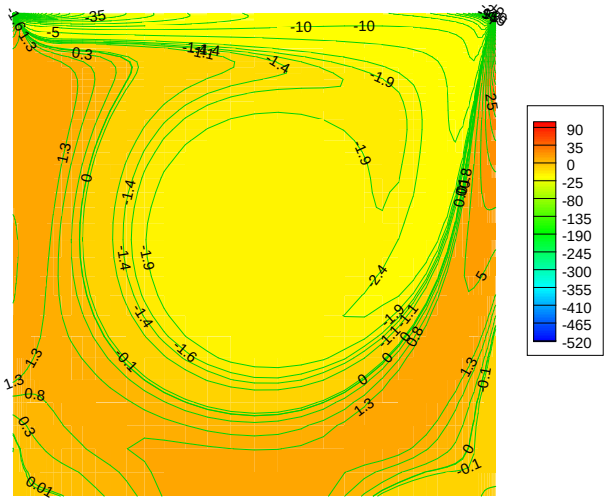

Figure 7.24: Vorticity contours of the driven cavity flow on a 50×50 clustered mesh, $\mathrm{Re} = 800$.

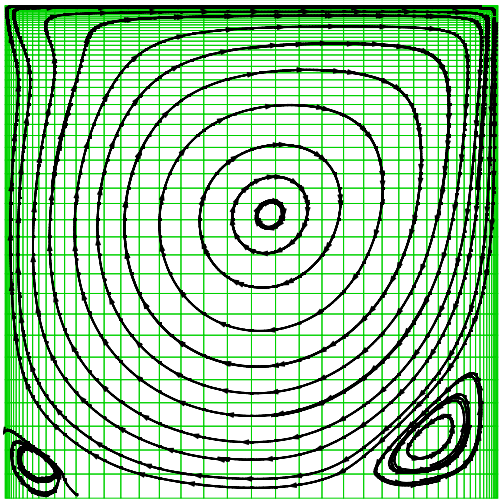

Figure 7.25: Streamlines of the driven cavity flow on a 50×50 clustered mesh, $\mathrm{Re} = 800$.

7.3.1 Problem statement and finite element mesh

In this section we present the computational domain and the boundary conditions that apply for the cylinder test case. The finite element mesh and the size of the domain are shown in Figure 7.26. In the design of the mesh we focus on ensuring adequate resolution to capture the relevant flow details. At the cylinder wall a grid clustering is used in order to have enough elements to efficiently resolve the developing boundary layer. Furthermore, in the downstream region, the number of elements is crucial to capture the vortex street, therefore, more elements are used in this direction.

In Figure 7.26 we show a slice of the finest mesh used in our computations. Since we are interested in developing a three dimensional code, a fully three dimensional mesh is created and we apply periodic boundary conditions in the third direction.

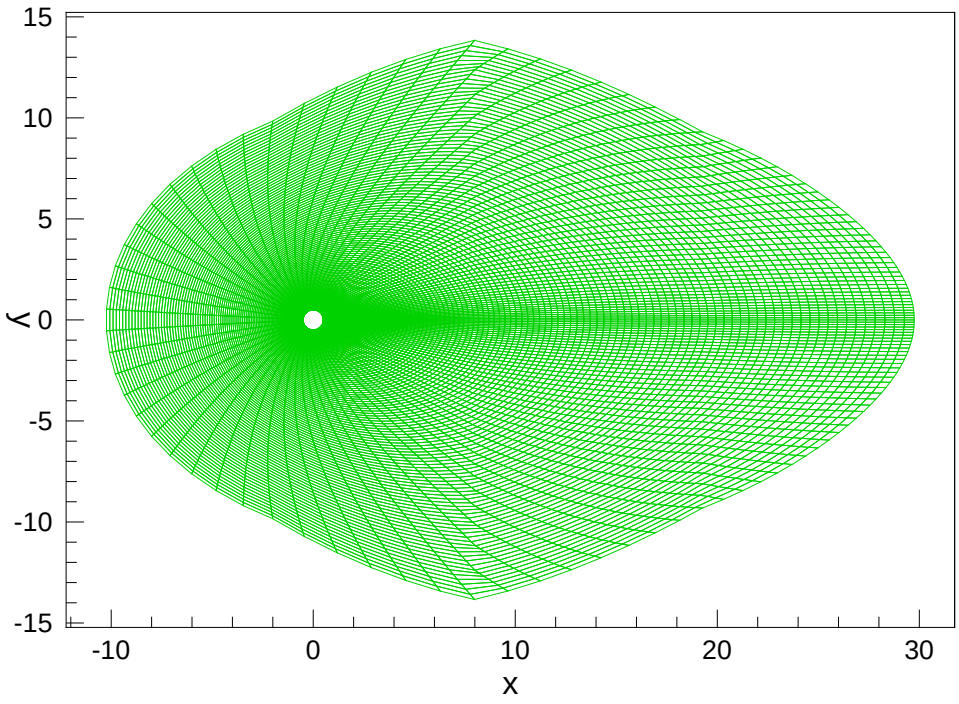

Figure 7.26: Horizontal slice of the finite element mesh for the circular cylinder with 43200 three dimensional elements, a total of 290400 degrees of freedom.

The computations are, however essentially two-dimensional and we use only three elements in the direction along the cylinder. The mesh generated for the circular cylinder is a topological " O" grid. The finest mesh uses $120 \times 120 \times 3$ elements in the radial, angular and x_3 coordinate directions, respectively.

The boundary conditions consists of unit horizontal velocity $u = (1, 0, 0)^T$ along the inflow boundary, zero pressure and zero normal derivatives of the velocities along the outflow boundary, and zero velocity components at the wall of the cylinder.

7.3.2 Results

In this section we describe two Reynolds number cases, Re = 40 and Re = 200. Consider first Re = 40, which is a standard test case to verify the numerical method. For this Reynolds number we study the accuracy of the numerical method under mesh refinement. In order to verify the spatial accuracy of the method, we consider three meshes: $60 \times 60 \times 3$, $85 \times 85 \times 3$ and $120 \times 120 \times 3$. Figures 7.27-7.29 show the streamlines, vorticity and pressure contours at steady state on the finest mesh.

The length of the wake and the separation point can be used to study the accuracy of the numerical method. The length of the wake behind the cylinder is obtained by plotting the horizontal velocity along the centerline behind the cylinder, see Figure 7.30 for the three meshes considered in the computations. When comparing our result on the finest mesh in Table 7.3 with the experimental data described in [10], we observe a difference in the wake length of only 1.8%. In Figure 7.31 we plot the wake length as a function of h^2, where h is the average mesh size in the horizontal direction. This plot shows a quadratic convergence, with a predicted wake length, as $h \to 0$, equal to 2.25. Similarly, we plot in Figure 7.32 the computed drag coefficient

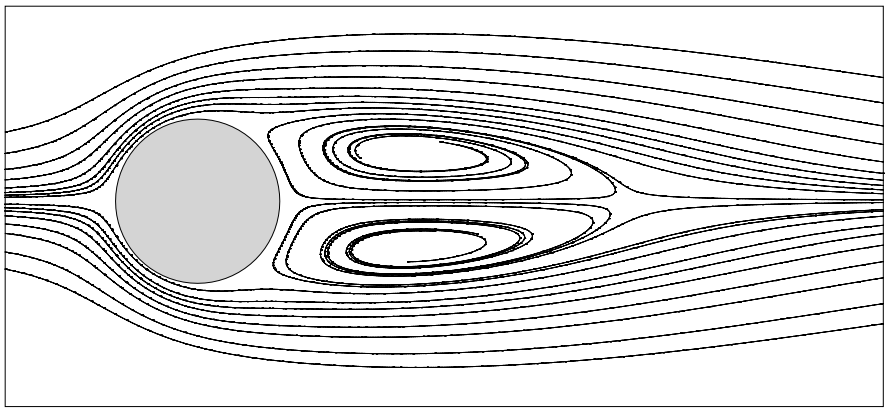

Figure 7.27: Streamlines for the circular cylinder at Re = 40 on a $120 \times 120 \times 3$ mesh.

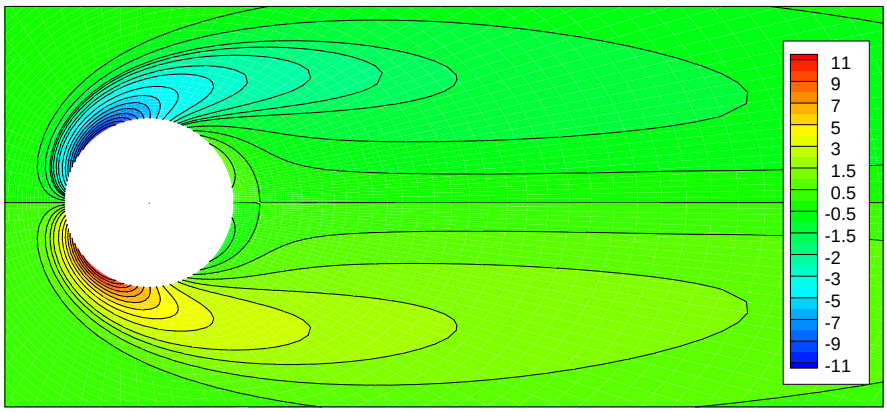

Figure 7.28: Vorticity contours for the circular cylinder at Re = 40 on a $120 \times 120 \times 3$ mesh.

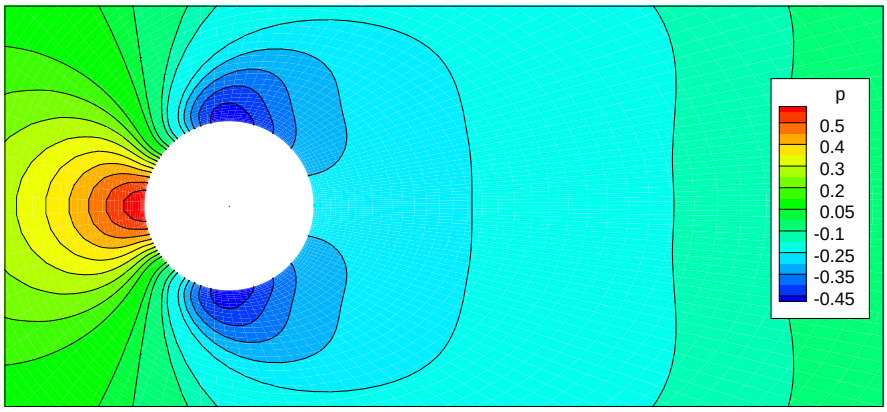

Figure 7.29: Pressure contours for the circular cylinder at Re = 40 on a $120 \times 120 \times 3$ mesh.

as a function of h^2, where also a quadratic convergence is observed. Table 7.3 shows that the separation point is also close to the one obtained from experiments.

The next test case we consider is Re = 200. At this Reynolds number vortex shedding occurs and the Reynolds number is small enough that the boundary layers can be resolved on the present mesh. Initially, a pair of symmetric eddies develops behind the cylinder, reaching a steady state at about $t = 50$. Streamlines of this steady solution are shown in Figure 7.33. In the computations to obtain an initial steady solution, the basis functions are chosen to be linear-in-space and constant-in-time, using a time step of 0.1.

After an initial steady solution was obtained, the time step was reduced to 0.02. For another 200 time steps, a very small un-symmetry in the eddies was observed, therefore, a small perturbation was added to begin the vortex shedding process. To obtain a time accurate solution, the time step was further reduced to 0.01 and a predictor multi-corrector method has been employed with only one iteration in each time step (no corrector passes were performed). The resulting linear system was solved iteratively with the BiCGStab solver. Since for time accuracy, it is important to solve the linear system accurately, the tolerance of the solver was set to $O(10^{-6})$. No preconditioners were employed in the BiCGStab method.

In Figure 7.34 and 7.35, the vorticity contours and streamlines of the developing

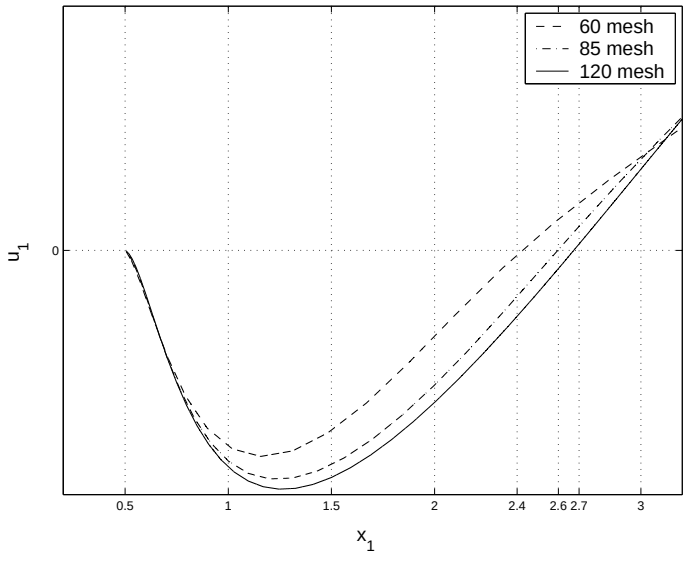

Figure 7.30: Horizontal velocity along the centerline behind the cylinder on the finest mesh, Re = 40.

	Wake length	Separation point, θ
Experiment, [10]	2.13	53.5^0
Numerical simulations		
120×120 mesh	2.174	53.9^0
85×85 mesh	2.096	53.1^0
60×60 mesh	1.927	51.5^0

Table 7.3: Wake length and separation point in degrees, measured from the x_1-axis for the flow around the circular cylinder, Re = 40.

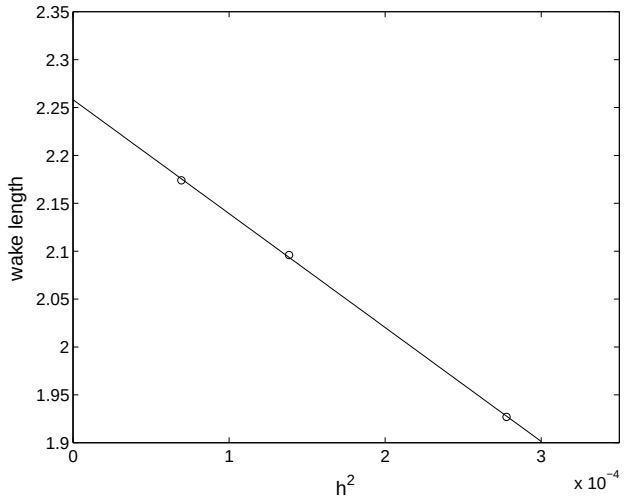

Figure 7.31: Predicted wake length for the circular cylinder at Re = 40.

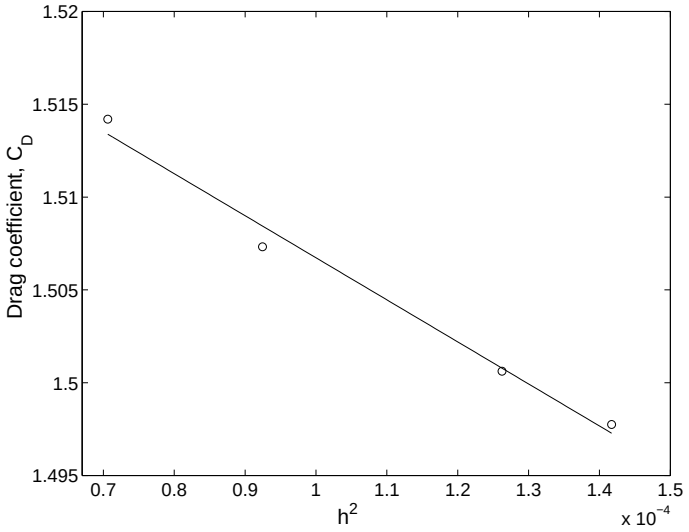

Figure 7.32: Drag coefficient for the circular cylinder at Re = 40.

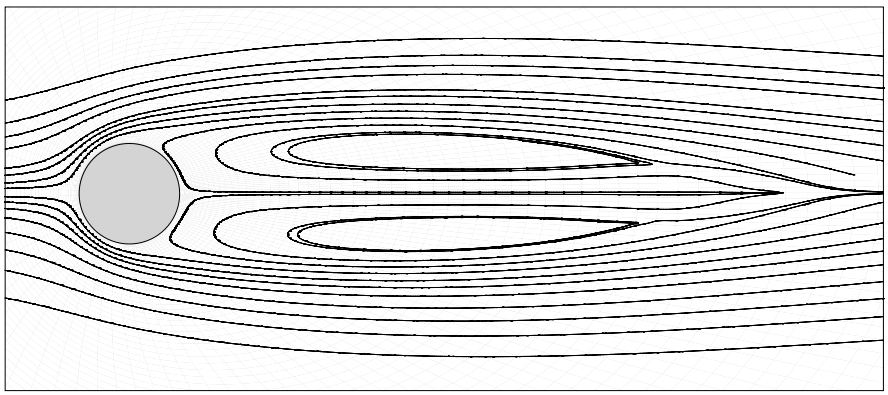

Figure 7.33: Streamlines for the initial steady solution for the flow around the circular cylinder at $t = 50$, Re = 200.

vortex street can be observed at time $t = 70$. We can further follow the streamlines for times $t = 100$ and $t = 110$, in Figures 7.36 and 7.37, respectively. In Figures 7.38-7.40 we plot these streamlines close to the cylinder wall, showing the detailed flow structure near the cylinder, in particular the separation region. These simulations show that the numerical discretization with the stabilization operators developed in this thesis are well capable of capturing the von Karman vortex street behind the cylinder.

7.4 Concluding remarks

In this chapter we verified our three dimensional finite element code for a variety of numerical examples. We investigated the influence of the stabilization matrix on the accuracy of the Galerkin least-squares finite element method for the incompressible Navier-Stokes equations, when primitive variables are used. The main conclusion is that the parameter ω in the newly designed stabilization matrix (4.3.13) does not influence the spatial accuracy of the numerical discretization. We verified this result also for higher order discretizations. The parameter ω does influence, however, the convergence to steady state and improves the conditioning of the linear system, resulting in a faster convergence for GMRES and BiCGStab solvers for $\omega = 0.6$. For the value $\omega = 0.6$ we also obtained that the iterative solver to solve the linear system (GMRES or BiCGStab) is more robust.

The finite element method has been demonstrated to capture the detailed flow structures for both the driven cavity and the flow about a circular cylinder. The stabi-

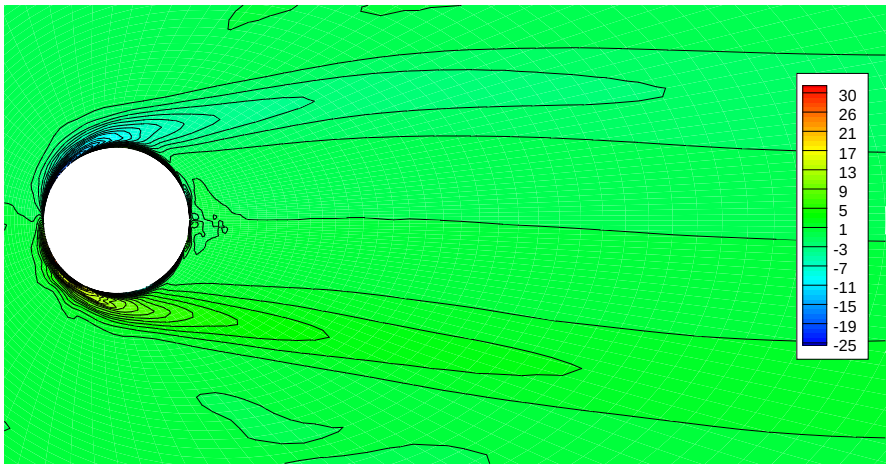

Figure 7.34: Vorticity contours for the circular cylinder at $t = 70$, Re $= 200$.

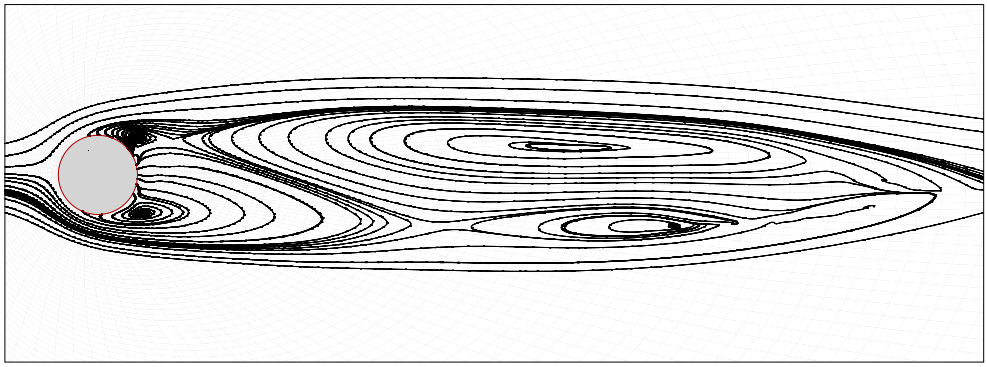

Figure 7.35: Streamlines for the circular cylinder at $t = 70$, $\mathrm{Re} = 200$.

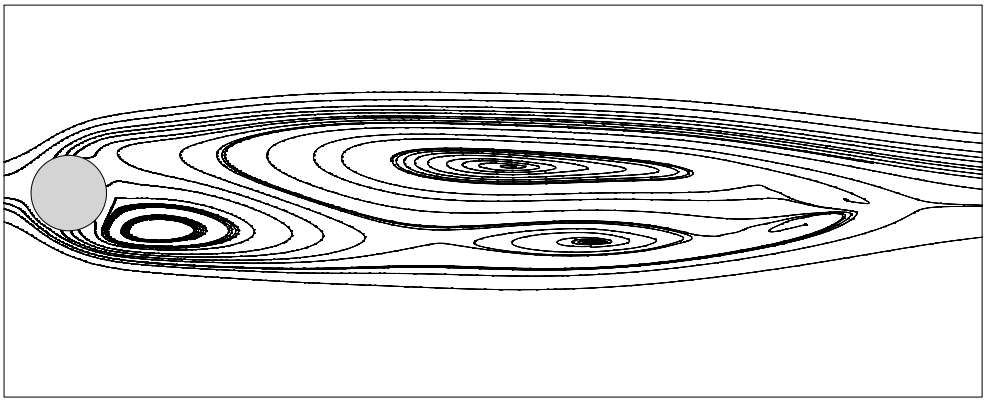

Figure 7.36: Streamlines for the circular cylinder at $t = 100$, $\mathrm{Re} = 200$.

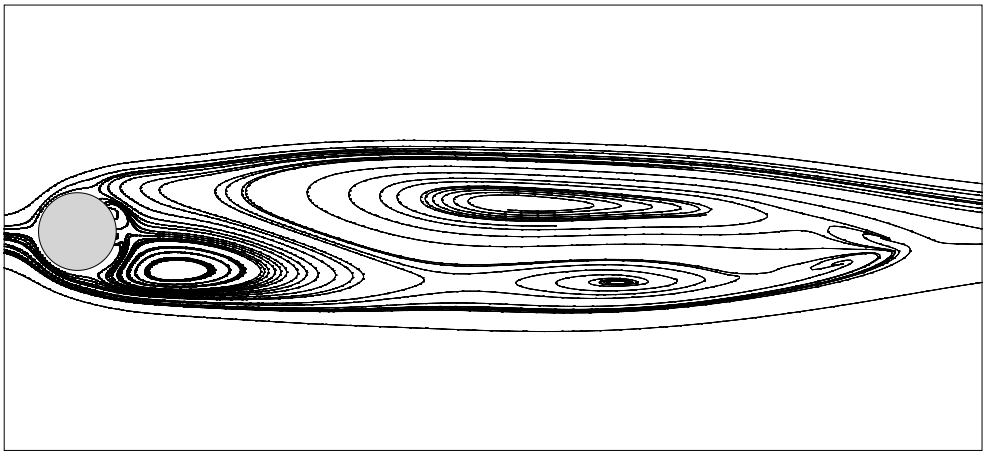

Figure 7.37: Streamlines for the circular cylinder at $t = 110$, $\mathrm{Re} = 200$.

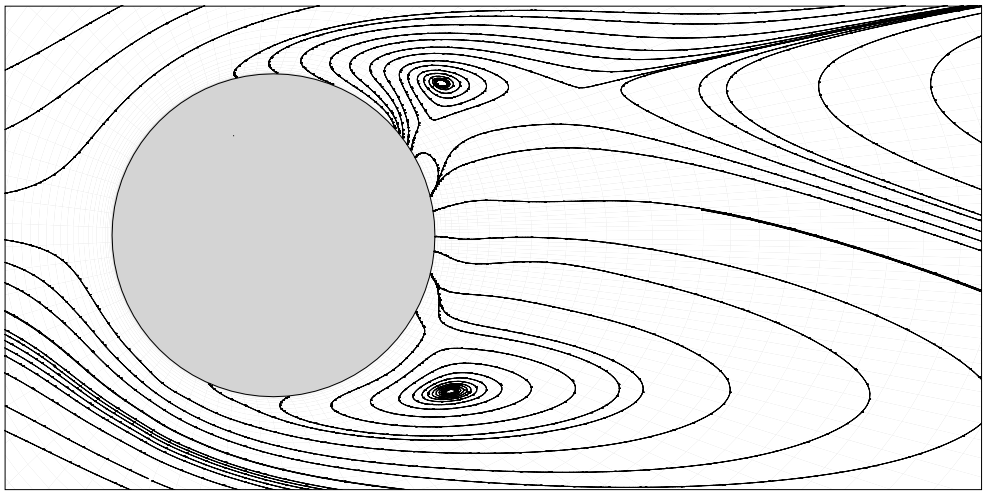

Figure 7.38: Streamlines near the cylinder wall at $t = 70$, $\mathrm{Re} = 200$.

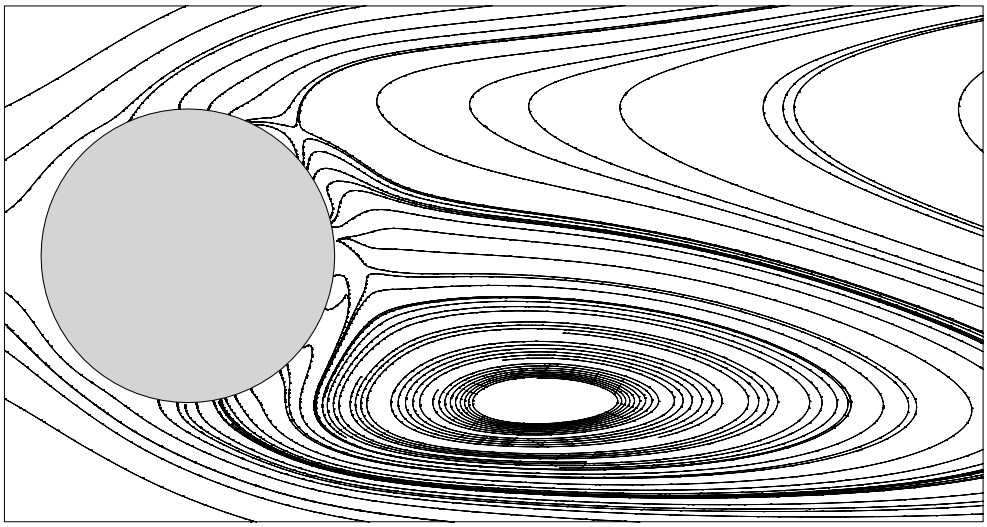

Figure 7.39: Streamlines near cylinder wall at $t = 100$, Re $= 200$.

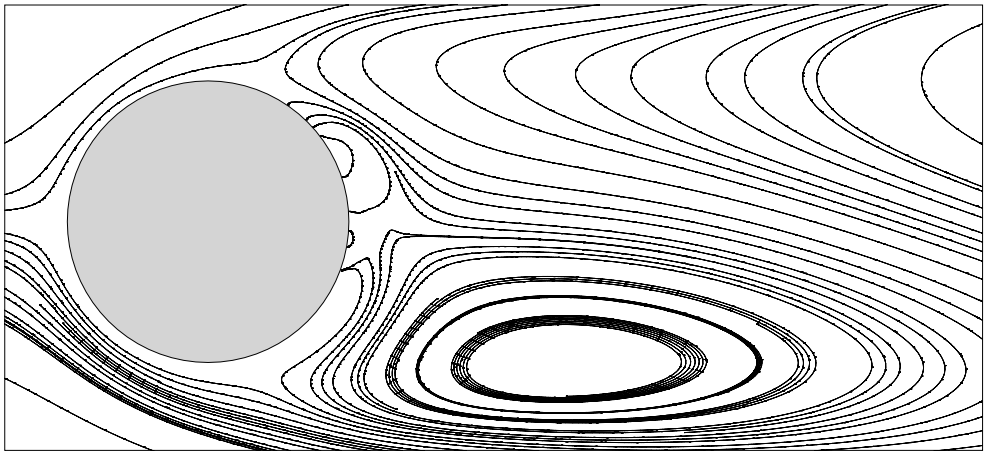

Figure 7.40: Streamlines near the cylinder wall at $t = 110$, Re $= 200$.

lization matrices developed in this thesis performed well in stabilizing the numerical method without seriously degrading accuracy. Further investigations on deforming meshes and for higher Reynolds numbers will be needed to demonstrate also the effectiveness for predicting the dynamic behavior of the cylinder due to periodic vortex shedding.

We believe that a similar investigation is needed to verify the accuracy of the method when entropy variables are used. Furthermore, we incorporated, but did not investigate, the evolution of temperature field in the various test cases.

Chapter 8

Conclusions and further research

In this chapter we give an overview of the most important results and conclusions of this thesis. Furthermore, we indicate some directions for further research.

8.1 Conclusions

The goal of this research was to design and analyze stabilization matrices in a Galerkin least-squares finite element method for the Navier-Stokes equations, suitable for a wide spectrum of flow problems. To achieve this goal, we have employed the symmetric form of the Navier-Stokes equations using entropy variables. This form gives the possibility to study the incompressible limit of the equations.

The space-time Galerkin least-squares method was employed, which permits discrete discontinuities in time and is suitable for problems requiring moving and deforming meshes. This method has been successfully used for both compressible and incompressible flows. Having established the method for entropy variables, we may transform it to any set of variables for which the incompressible limit is well-defined. Using this advantage, we gave a consistent mathematical derivation of a class of stabilization operators suitable for space-time Galerkin least-squares discretizations of the incompressible Navier-Stokes equations for both entropy and primitive variables. This derivation is based on a dimensional analysis of the stabilization matrix to determine its dependence on the flow variables. Next, we analyzed the resulting class of stabilization matrices such that we can ensure that the Galerkin least-squares finite element discretization results in a stable discretization technique, at least for the locally linearized problem.

The Galerkin formulation of the symmetric compressible Navier-Stokes equations automatically satisfies the second law of thermodynamics. The Galerkin method applied

to the compressible Navier-Stokes equations requires a stabilization operator to compensate the lack of stability for advection dominated problems, while maintaining the accuracy of the Galerkin method for smooth solutions. We have proposed a new definition of stabilization operators for compressible flows without shocks, which has certain advantages over the original definitions given in [32] and [50]. The new matrix has a simple structure that is much easier to implement. Additionally, when the incompressible limit is taken, this matrix results in the stabilization matrix obtained for incompressible flows, therefore provides a larger range of applicability. Furthermore, we gave necessary and sufficient conditions on the positive definiteness of the designed stabilization matrix for entropy variables. Under the condition of positive definiteness of the stabilization matrix, the Galerkin least-squares method for the symmetrized compressible Navier-Stokes equations satisfies the entropy condition, which results in a nonlinear stability condition, as we discussed in detail in this thesis.

Since we are interested in the incompressible limit, it is necessary to study the various thermodynamic limits. We proposed a general form of the fundamental equation, that is valid in the incompressible limit. The main statement is that the thermodynamic state of a single species material is determined by three measurable quantities, the volume expansivity α_p, the isothermal compressibility β_T and the specific heat at constant pressure c_p (or specific heat at constant volume c_v).

8.2 Further research

In the numerical examples discussed in this thesis we investigated the properties of the newly designed stabilization operator for the incompressible Navier-Stokes equations in terms of the primitive variables (p, u, T). An important continuation of this research consists of analyzing the influence of the stabilization operator for entropy variables on the accuracy and robustness of the Galerkin least-squares method for incompressible flows. Furthermore, for primitive variable computations, we did not discuss the temperature equation, which would be of great interest for many applications.

The main objective of this thesis is to give a unified formulation of stabilization operators valid for wide range of fluid flow applications. We analyzed the mathematical properties of the proposed matrices for both types of flows but we did not verify their performance for compressible flows. Therefore, this remains for future investigation.

The time-discontinuous Galerkin least-squares finite element discretization results in a large system of nonlinear algebraic equations. The predictor multi-corrector algorithm originating from the constant-in-time approximation of the Galerkin least-squares variational equation has good stability properties but is of low order of accuracy in time. This algorithm, introduced in [50], is well-suited for solving steady problems and we have employed this method in our steady computations. For unsteady problems, a linear-in-time approximation of the space-time Galerkin least-squares variational

equation is needed. In [50], a predictor multi-corrector method is proposed to reduce the resulting large algebraic system to two weakly coupled systems. This algorithm employs the same left hand side matrix for both systems. In this thesis we proposed a different method to solve the nonlinear algebraic system. We compared the algorithm with the predictor multi-corrector method using the advection-diffusion equation as a model problem. We concluded that the two methods have similar properties, further investigation is, however needed. In this method we need to compute two matrices, but our preliminary experience suggests that this can be a better solution technique for incompressible flows. Further improvement of iterative methods to solve the resulting linear system in a Newton method for incompressible flows is needed to obtain more robust solvers.

In the finite element discretization of the incompressible Navier-Stokes equations, we discussed some aspects of the mesh deformation, but did not yet implement them in our computer program. The extension of this work to problems which require deforming meshes, such as risers, would be very useful for many applications.

Further challenge would be to incorporate different equations of state in the formulation. The general equation of state provides a good starting point towards this challenge.

Appendices

A Measurements

In Table A.1 we listed some measured values of α_p and β_T for water [57], at various temperatures. Table A.2 shows some measured values of α_p and β_T for several substances at 293 K, [38].

Temperature (K)	α_p (10^{-3} 1/K)	β_T (10^{-6} 1/bar)
243	-1.400	80.79
253	-0.661	64.25
263	-0.292	55.83
273	-0.068	50.89
283	0.088	47.81
293	0.207	45.89
303	0.303	44.77
313	0.385	44.24
323	0.458	44.17
333	0.523	44.50
343	0.584	45.16
353	0.641	46.14
363	0.696	47.43
373	0.750	49.02

Table A.1: Measured values for β_T and α_p over a range of temperatures for water.

Substance	α_p (10^{-4} 1/K)	β_T (10^{-6} 1/atm)
Liquids		
Benzene	12.4	92.1
Carbon tetrachloride	12.4	90.5
Ethanol	11.2	76.8
Mercury	1.82	38.7
Water	2.1	49.6
Solids		
Copper	0.501	0.735
Diamond	0.030	0.187
Iron	0.354	0.589
Lead	0.861	2.21

Table A.2: Expansion coefficient α_p and isothermal compressibility β_T, measured at 293 K for several substances.

B Flux Jacobian matrices

B.1 Flux Jacobian matrices for entropy variables

The advective flux Jacobian matrices in terms of the entropy variables V have the form:

$$\tilde{A}_0 = \rho^2 \beta_T T \begin{pmatrix} 1 & u_1 & u_2 & u_3 & e_2 \\ u_1 & c_1 & u_1u_2 & u_1u_3 & u_1e_3 \\ u_2 & u_1u_2 & c_2 & u_2u_3 & u_2e_3 \\ u_3 & u_1u_3 & u_2u_3 & c_3 & u_3e_3 \\ e_2 & u_1e_3 & u_2e_3 & u_3e_3 & e_5 \end{pmatrix}$$

$$\tilde{A}_1 = \rho^2 \beta_T T \begin{pmatrix} u_1 & c_1 & u_1u_2 & u_1u_3 & u_1e_3 \\ c_1 & a_1 & u_2c_1 & u_3c_1 & b_1 \\ u_1u_2 & u_2c_1 & u_1c_2 & u_1u_2u_3 & u_1u_2e_4 \\ u_1u_3 & u_3c_1 & u_1u_2u_3 & u_1c_3 & u_1u_3e_4 \\ u_1e_3 & b_1 & u_1u_2e_4 & u_1u_3e_4 & d_1 \end{pmatrix}$$

$$\tilde{A}_2 = \rho^2\beta_T T \begin{pmatrix} u_2 & u_1u_2 & c_2 & u_2u_3 & u_2e_3 \\ u_1u_2 & u_2c_1 & u_1c_2 & u_1u_2u_3 & u_1u_2e_4 \\ c_2 & u_1c_2 & a_2 & u_3c_2 & b_2 \\ u_2u_3 & u_1u_2u_3 & u_3c_2 & u_2c_3 & u_2u_3e_4 \\ u_2e_3 & u_1u_2e_4 & b_2 & u_2u_3e_4 & d_2 \end{pmatrix}$$

$$\tilde{A}_3 = \rho^2\beta_T T \begin{pmatrix} u_3 & u_1u_3 & u_2u_3 & c_3 & u_3e_3 \\ u_1u_3 & u_3c_1 & u_1u_2u_3 & u_1c_3 & u_1u_3e_4 \\ u_2u_3 & u_1u_2u_3 & u_3c_2 & u_2c_3 & u_2u_3e_4 \\ c_3 & u_1c_3 & u_2c_3 & a_3 & b_3 \\ u_3e_3 & u_1u_3e_4 & u_2u_3e_4 & b_3 & d_3 \end{pmatrix}$$

where we used the following notations:

$$\begin{aligned} a_i &= u_i(u_i^2 + \frac{3}{\rho\beta_T}), & c_i &= u_i^2 + \frac{1}{\rho\beta_T}, \\ k &= \frac{|u|^2}{2}, & e_1 &= h + k, \\ \bar{\gamma} &= \frac{\alpha_p}{\rho\beta_T c_v}, & e_2 &= e_1 - T\bar{\gamma}c_v, \\ e_3 &= e_2 + \frac{1}{\rho\beta_T}, & e_4 &= e_2 + \frac{2}{\rho\beta_T}, \\ d &= \frac{\alpha_p T}{\rho\beta_T}, & e_5 &= e_1^2 - 2e_1 d + \frac{2k + c_p T}{\rho\beta_T}, \\ b_i &= u_i^2 e_4 + \frac{e_1}{\rho\beta_T}, & d_i &= u_i(e_5 + \frac{2e_1}{\rho\beta_T}). \end{aligned}$$

The viscous flux Jacobian matrices in terms of the entropy variables, which satisfy the relation $\tilde{K}_{ij} = \tilde{K}_{ji}^T$, are given as:

$$\tilde{K}_{11} = T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 4\mu/3 & 0 & 0 & 4\mu u_1/3 \\ 0 & 0 & \mu & 0 & \mu u_2 \\ 0 & 0 & 0 & \mu & \mu u_3 \\ 0 & 4\mu u_1/3 & \mu u_2 & \mu u_3 & k_1 \end{pmatrix} \quad \tilde{K}_{22} = T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & \mu u_1 \\ 0 & 0 & 4\mu/3 & 0 & 4\mu u_2/3 \\ 0 & 0 & 0 & \mu & \mu u_3 \\ 0 & \mu u_1 & 4\mu u_2/3 & \mu u_3 & k_2 \end{pmatrix}$$

$$\tilde{K}_{33} = T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & \mu u_1 \\ 0 & 0 & \mu & 0 & \mu u_2 \\ 0 & 0 & 0 & 4\mu/3 & 4\mu u_3/3 \\ 0 & \mu u_1 & \mu u_2 & 4\mu u_3/3 & k_3 \end{pmatrix} \quad \tilde{K}_{12} = T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2\mu/3 & 0 & -2\mu u_2/3 \\ 0 & \mu & 0 & 0 & \mu u_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \mu u_2 & -2\mu u_1/3 & 0 & \mu u_1 u_2/3 \end{pmatrix}$$

$$\tilde{K}_{13} = T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2\mu/3 & -2\mu u_3/3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & \mu u_1 \\ 0 & \mu u_3 & 0 & -2\mu u_1/3 & \mu u_1 u_3/3 \end{pmatrix} \quad \tilde{K}_{23} = T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2\mu/3 & -2\mu u_3/3 \\ 0 & 0 & \mu & 0 & \mu u_2 \\ 0 & 0 & \mu u_3 & -2\mu u_2/3 & \mu u_2 u_3/3 \end{pmatrix}$$

where κ is the coefficient of thermal conductivity and $k_i = \frac{1}{3}\mu u_i^2 + \mu|u|^2 + \kappa T$ for $i = 1, 2, 3$.

The incompressible limit of the Navier-Stokes equations are obtained by setting α_p and β_T equal to zero. The Jacobian matrices in the incompressible limit are given by:

$$\tilde{A}_0^{inc} = \rho T \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & u_1 \\ 0 & 0 & 1 & 0 & u_2 \\ 0 & 0 & 0 & 1 & u_3 \\ 0 & u_1 & u_2 & u_3 & r \end{pmatrix}$$

$$\tilde{A}_1^{inc} = \rho T \begin{pmatrix} 0 & 1 & 0 & 0 & u_1 \\ 1 & 3u_1 & u_2 & u_3 & 2u_1^2 + e_1 \\ 0 & u_2 & u_1 & 0 & 2u_1 u_2 \\ 0 & u_3 & 0 & u_1 & 2u_1 u_3 \\ u_1 & 2u_1^2 + e_1 & 2u_1 u_2 & 2u_1 u_3 & u_1(r + 2e_1) \end{pmatrix}$$

$$\tilde{A}_2^{inc} = \rho T \begin{pmatrix} 0 & 0 & 1 & 0 & u_2 \\ 0 & u_2 & u_1 & 0 & 2u_1 u_2 \\ 1 & u_1 & 3u_2 & u_3 & 2u_2^2 + e_1 \\ 0 & 0 & u_3 & u_2 & 2u_2 u_3 \\ u_2 & 2u_1 u_2 & 2u_2^2 + e_1 & 2u_2 u_3 & u_2(r + 2e_1) \end{pmatrix}$$

$$\tilde{A}_3^{inc} = \rho T \begin{pmatrix} 0 & 0 & 0 & 1 & u_3 \\ 0 & u_3 & 0 & u_1 & 2u_1 u_3 \\ 0 & 0 & u_3 & u_2 & 2u_2 u_3 \\ 1 & u_1 & u_2 & 3u_3 & 2u_3^2 + e_1 \\ u_3 & 2u_1 u_3 & 2u_2 u_3 & 2u_3^2 + e_1 & u_3(r + 2e_1) \end{pmatrix}$$

with $k = |u|^2/2$, $r = 2k + c_p T$ and $e_1 = h + k$. Note that the viscous flux Jacobian matrices $\tilde{K}_{ij}$ are independent of α_p and β_T and do not change in the incompressible limit.

B.2 Flux Jacobian matrices for primitive variables

The Euler Jacobian matrices with respect to the primitive variables $Y = (p, u_1, u_2, u_3, T)^T$ are:

$$A_0(Y) = \begin{pmatrix} \rho\beta_T & 0 & 0 & 0 & -\rho\alpha_p \\ \rho\beta_T u_1 & \rho & 0 & 0 & -\rho\alpha_p u_1 \\ \rho\beta_T u_2 & 0 & \rho & 0 & -\rho\alpha_p u_2 \\ \rho\beta_T u_3 & 0 & 0 & \rho & -\rho\alpha_p u_3 \\ e_1^p & \rho u_1 & \rho u_2 & \rho u_3 & e_4^p \end{pmatrix}$$

$$A_1(Y) = \begin{pmatrix} \rho\beta_T u_1 & \rho & 0 & 0 & -\rho\alpha_p u_1 \\ \rho\beta_T u_1^2 + 1 & 2\rho u_1 & 0 & 0 & -\rho\alpha_p u_1^2 \\ \rho\beta_T u_1 u_2 & \rho u_2 & \rho u_1 & 0 & -\rho\alpha_p u_1 u_2 \\ \rho\beta_T u_1 u_3 & \rho u_3 & 0 & \rho u_1 & -\rho\alpha_p u_1 u_3 \\ u_1 e_2^p & e_3^p + \rho u_1^2 & \rho u_1 u_2 & \rho u_1 u_3 & u_1 e_4^p \end{pmatrix}$$

$$A_2(Y) = \begin{pmatrix} \rho\beta_T u_2 & 0 & \rho & 0 & -\rho\alpha_p u_2 \\ \rho\beta_T u_1 u_2 & \rho u_2 & \rho u_1 & 0 & -\rho\alpha_p u_1 u_2 \\ \rho\beta_T u_2^2 + 1 & 0 & 2\rho u_2 & 0 & -\rho\alpha_p u_2^2 \\ \rho\beta_T u_2 u_3 & 0 & \rho u_3 & \rho u_2 & -\rho\alpha_p u_2 u_3 \\ u_2 e_2^p & \rho u_1 u_2 & e_3^p + \rho u_2^2 & \rho u_2 u_3 & u_2 e_4^p \end{pmatrix}$$

$$A_3(Y) = \begin{pmatrix} \rho\beta_T u_3 & 0 & 0 & \rho & -\rho\alpha_p u_3 \\ \rho\beta_T u_1 u_3 & \rho u_3 & 0 & \rho u_1 & -\rho\alpha_p u_1 u_3 \\ \rho\beta_T u_2 u_3 & 0 & \rho u_3 & \rho u_2 & -\rho\alpha_p u_2 u_3 \\ \rho\beta_T u_3^2 + 1 & 0 & 0 & 2\rho u_3 & -\rho\alpha_p u_3^2 \\ u_3 e_2^p & \rho u_1 u_3 & \rho u_2 u_3 & e_3^p + \rho u_3^2 & u_3 e_4^p \end{pmatrix}$$

where

$$e_1^p = \rho\beta_T e_1 - \alpha_p T, \quad e_2^p = e_1^p + 1, \quad e_3^p = \rho e_1, \quad e_4^p = -\rho\alpha_p e_1 + \rho c_p.$$

The diffusivity coefficient matrices $K_{ij}(Y)$, for $i, j = 1, 2, 3$ have the form:

$$K_{11}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \chi & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 \\ 0 & \chi u_1 & \mu u_2 & \mu u_3 & \kappa \end{pmatrix}, \quad K_{12}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 \\ 0 & \mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \mu u_2 & \lambda u_1 & 0 & 0 \end{pmatrix}$$

$$K_{13}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & 0 \\ 0 & \mu u_3 & 0 & \lambda u_1 & 0 \end{pmatrix}, \quad K_{21}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & 0 \\ 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda u_2 & \mu u_1 & 0 & 0 \end{pmatrix}$$

$$K_{22}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & 0 \\ 0 & 0 & \chi & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 \\ 0 & \mu u_1 & \chi u_2 & \mu u_3 & \kappa \end{pmatrix}, \quad K_{23}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & \mu u_3 & \lambda u_2 & 0 \end{pmatrix}$$

$$K_{31}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 & 0 \\ 0 & \lambda u_3 & 0 & \mu u_1 & 0 \end{pmatrix}, \quad K_{32}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda u_3 & \mu u_2 & 0 \end{pmatrix}$$

$$K_{33}(Y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & \chi & 0 \\ 0 & \mu u_1 & \mu u_2 & \chi u_3 & \kappa \end{pmatrix},$$

where $\chi = \lambda + 2\mu$.

B.3 Variable transformation matrices

The transformation matrix $V_{,Y}$ has the form

$$V_{,Y} = \begin{pmatrix} \frac{1}{\rho T} & -\frac{u_1}{T} & -\frac{u_2}{T} & -\frac{u_3}{T} & -\frac{h-k}{T^2} \\ 0 & \frac{1}{T} & 0 & 0 & -\frac{u_1}{T^2} \\ 0 & 0 & \frac{1}{T} & 0 & -\frac{u_2}{T^2} \\ 0 & 0 & 0 & \frac{1}{T} & -\frac{u_3}{T^2} \\ 0 & 0 & 0 & 0 & \frac{1}{T^2} \end{pmatrix}$$

where $Y = (p, u_1, u_2, u_3, T)^T$. The inverse matrix transformation is given by

$$Y_{,V} = \begin{pmatrix} \rho T & \rho T u_1 & \rho T u_2 & \rho T u_3 & \rho T(h+k) \\ 0 & T & 0 & 0 & u_1 T \\ 0 & 0 & T & 0 & u_2 T \\ 0 & 0 & 0 & T & u_3 T \\ 0 & 0 & 0 & 0 & T^2 \end{pmatrix}.$$

C The solution of the nonlinear system

C.1 Third-order predictor multi-corrector algorithm

The third-order predictor multi-corrector algorithm can be summarized as follows:

(Initialization)
Set $v_{(0)}$
(Time steps)
for $n_{time} = 0, \ldots, n_{step}$
    (Predictor)
    $v^{(0)} = \tilde{v}^{(0)} = v_{(n)}$
    Set Δt
    (Multi-corrector loop)
    for $i = 0, \ldots, i_{max}$
        form $R^{(i)}(v^{(i)}, \tilde{v}^{(i)}, v_{(n)})$
        form $M^*(v^{(i)}, \tilde{v}^{(i)})$
        solve for $\Delta v^{(i)}$:
            $M^* \Delta v^{(i)} = -R^{(i)}$
        update: $v^{(i+1)} = v^{(i)} + \Delta v^{(i)}$
        if $i = i_{max}$ exit the loop
        form $\tilde{R}^{(i)}(\tilde{v}^{(i)}, v^{(i+1)}, v_{(n)})$
        form $\tilde{M}^*(v^{(i+1)}, \tilde{v}^{(i)})$
        solve for $\Delta \tilde{v}^{(i)}$:
            $\tilde{M}^* \Delta \tilde{v}^{(i)} = -\tilde{R}^{(i)}$
        update: $\tilde{v}^{(i+1)} = \tilde{v}^{(i)} + \Delta \tilde{v}^{(i)}$
    (end multi-corrector loop)
    $v_{(n+1)} = v^{(i+1)}$
(end time steps)

C.2 Modified predictor multi-corrector algorithm

The modification of the third-order predictor multi-corrector algorithm, proposed in this thesis, can be summarized as follows:

(Initialization)
Set $v_{(0)}$
(Time steps)
for $n_{time} = 0, \ldots, n_{step}$
$v^{(0)} = \tilde{v}^{(0)} = v_{(n)}$
Set Δt
for $i = 0, \ldots, i_{max}$
form $R^{(i)}(\bar{v}^{(i)}, \overset{\triangle}{v}{}^{(i)}, v_{(n)})$
form $\tilde{R}^{(i)}(\bar{v}^{(i)}, \overset{\triangle}{v}{}^{(i)}, v_{(n)})$
form $M_{11}(\bar{v}^{(i)}, \overset{\triangle}{v}{}^{(i)})$
form $M_{22}(\bar{v}^{(i)}, \overset{\triangle}{v}{}^{(i)})$
solve for $\Delta\bar{v}^{(i)}$ and $\Delta\overset{\triangle}{v}{}^{(i)}$:

$$2M_{11}\Delta\bar{v}^{(i)} = -R^{(i)}$$

$$-2\tilde{M}_{22}\Delta\overset{\triangle}{v}{}^{(i)} = -\tilde{R}^{(i)}$$

update: $v^{(i+1)} = v^{(i)} + \Delta\bar{v}^{(i)} + \Delta\overset{\triangle}{v}{}^{(i)})$
update: $\tilde{v}^{(i+1)} = \tilde{v}^{(i)} + \Delta\bar{v}^{(i)} - \Delta\overset{\triangle}{v}{}^{(i)}$
(end of the i loop)
$v_{(n+1)} = v^{(i+1)}$
(end time steps)

Bibliography

[1] I. Babuška. Error bounds for finite element method. *Numer. Math*, 16:322–333, 1971.

[2] Wolfgang Bangerth, Ralf Hartmann, and Guido Kanschat. `deal.II` *Differential Equations Analysis Library, Technical Reference*. `http://www.dealii.org`.

[3] T.J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. *Notes in Computational Sci. and Eng.*, 5, 1998.

[4] S.C. Brenner and L.R. Scott. *The mathematical theory of finite element methods. Second edition.* Springer-Verlag, New York, 2002.

[5] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising form Lagrange multipliers. *Rev. Francaise d'Automatique Inform. Rech. Oper., Ser. Rouge Anal. Numer.*, 8:129–151, 1974.

[6] F. Brezzi and M. Fortin. *The mathematical theory of finite element methods. Second edition.* Springer-Verlag, New York, 1991.

[7] F. Brezzi and A. Russo. Stabilization techniques for the finite element method. *Applied and Industrial Mathematics, Venice-2*, pages 47–58, 1998.

[8] A.N. Brooks and T.J.R. Hughes. Streamline upwind/Petrov-Galerkin methods for convective dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Meth. Appl. Mech. Engrg.*, 32:199–259, 1982.

[9] F. Chalot, T.J.R. Hughes, and F. Shakib. Symmetrization of conservation laws with entropy for high-temperature hypersonic computations. *Comput. Syst. Eng.*, 1:495–521, 1990.

[10] M. Coutanceau and R. Bouard. Experimental determination of the viscous flow in a wake of a circular cylinder in a uniform translation. part I. steady flow. *J. Fluid Mech.*, 79:231–256, 1977.

[11] J. Douglas and J. Wang. An absolutely stabilized finite element method for the Stokes problem. *Math. Comp.*, 52:495–508, 1989.

[12] P.K. Dutt. Stable boundary conditions and difference schemes for Navier-Stokes type equations. *SIAM J. Numer. Anal.*, 25(2):245–267, 1988.

[13] L.P. Franca and S.L. Frey. Stabilized finite element methods: II. The incompressible Navier-Stokes equations. *Comput. Meth. Appl. Mech. Engrg*, 99:209–233, 1992.

[14] L.P. Franca, S.L. Frey, and T.J.R. Hughes. Stabilized finite element methods: I. application to the advective-diffusive model. *Comput. Meth. Appl. Mech. Engrg.*, 95:253–276, 1992.

[15] L.P. Franca, T.E. Tezduyar, and A. Masud. *Finite element methods: 1970's and beyond*. CIMNE, 2004.

[16] U. Ghia, K.N. Ghia, and C.T. Shin. High-re solutions for incompressible flow using the Navier-Stokes equations and a multigrid method. *J. Comput. Phys.*, 48:387–441, 1982.

[17] V. Girault and P.A. Raviart. *Finite element methods for Navier-Stokes equations*. Springer-Verlag, Berlin Heidelberg, 1986.

[18] S.K. Godunov. The problem of generalized solution in the theory of quasilinear equations and in gas dynamics. *Russ. Math. Surveys*, 17:145–156, 1962.

[19] M.D. Gunzburger and R.A. Nicolaides. *Incompressible computational fluid dynamics*. Cambridge University Press, 1993.

[20] P. Hansbo and A. Szepessy. A velocity-pressure streamline diffusion finite element method for the incompressible Navier-Stokes equations. *Comput. Meth. Appl. Mech. Engrg.*, 84:175–192, 1990.

[21] A. Harten. On the symmetric form of systems of conservation laws with entropy. *J. Comp. Phys.*, 49:151–164, 1983.

[22] G. Hauke and T.J.R. Hughes. A unified approach for compressible and incompressible flows. *Comput. Meth. Appl. Mech. Engrg*, 113:383–395, 1994.

[23] G. Hauke and T.J.R. Hughes. A comparative study of different sets of variables for solving compressible and incompressible flows. *Comput. Meth. Appl. Mech. Engrg*, 153:1–44, 1998.

[24] T.J.R. Hughes. A simple scheme for developing 'upwind' finite elements. *International Journal for Numerical Methods and Engineering*, 12:1359–1365, 1978.

[25] T.J.R. Hughes and A. Brooks. A multi-dimensional upwind scheme with no crosswind diffusion. *T.J.R. Hughes, ed., Finite element methods for convection dominated flows, AMD*, 34:19–35, 1979.

[26] T.J.R. Hughes and A. Brooks. Galerkin/upwind finite element mesh partitions in fluid mechanics. *J.J.H. Miller, ed., Boundary and interior layers-Computational and asymptotic methods*, pages 103–112, 1980.

[27] T.J.R. Hughes and L.P. Franca. A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. *Comput. Meth. Appl. Mech. Engrg*, 65:85–96, 1987.

[28] T.J.R. Hughes, L.P. Franca, and M. Balestra. A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuška-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations. *Comput. Meth. Appl. Mech. Engrg*, 59:85–99, 1986.

[29] T.J.R. Hughes, L.P. Franca, and G.M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations. *Comput. Meth. Appl. Mech. Engrg*, 73:173–189, 1989.

[30] T.J.R. Hughes, L.P. Franca, and M. Mallet. A new finite element formulation for computational fluid dynamics: I. Symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. *Comput. Meth. Appl. Mech. Engrg*, 54:223–234, 1986.

[31] T.J.R. Hughes, L.P. Franca, and M. Mallet. A new finite element formulation for computational fluid dynamics: VI. Convergence analysis of the generalized SUPG formulation for linear time-dependent multidimensional advective-diffusive systems. *Comput. Meth. Appl. Mech. Engrg*, 63:97–112, 1987.

[32] T.J.R. Hughes and M. Mallet. A new finite element formulation for computational fluid dynamics: III. The generalized streamline operator for multidimensional advective-diffusive systems. *Comput. Meth. Appl. Mech. Engrg*, 58:305–328, 1986.

[33] T.J.R. Hughes and M. Mallet. A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems. *Comput. Meth. Appl. Mech. Engrg*, 58:329–336, 1986.

[34] T.J.R. Hughes, M. Mallet, and A. Mizukami. A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Comput. Meth. Appl. Mech. Engrg*, 54:341–355, 1986.

[35] T.J.R. Hughes and T.E. Tezduyar. Finite element methods for first-order hyperbolic systems with particular emphasis on the compressible Euler equations. *Comput. Meth. Appl. Mech. and Eng.*, 45:217–284, 1984.

[36] G.M. Hulbert and T.J.R. Hughes. Space-time finite element methods for second order hyperbolic equations. *Comput. Meth. Appl. Mech. Engrg*, 84:327–348, 1990.

[37] C. Johnson, A. Szepessy, and P. Hansbo. On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws. *Math. Comp.*, 54(189):107–129, 1990.

[38] E.W. Lemmon et al. Thermodynamic properties of air and mixtures of nitrogen, argon, and oxygen from 60 to 2000 K at pressures to 2000 MPa. *J. Phys. Chem. Ref. Data*, 29(3):331–385, 2000.

[39] M. Marion and R. Temam. *Navier-Stokes Equations: Theory and Approximation.* Elsevier Science, 1998.

[40] A. Masud and T.J.R. Hughes. A space-time Galerkin/least-squares finite element formulation of the Navier-Stokes equations for moving domain problems. *Comput. Meth. Appl. Mech. and Eng.*, 146(1–2):91–126, 1997.

[41] R. Menikoff and J.P. Bradley. The Riemann problem for fluid flow of real materials. *Reviews of modern physics*, 61(1):75–130, 1989.

[42] S. Mittal and T.E. Tezduyar. A unified finite element formulation for compressible and incompressible flows using augmented conservation variables. *Comput. Meth. Appl. Mech. and Eng.*, 161:229–243, 1998.

[43] M.S. Mock. System of conservation laws of mixed type. *J. Diff. Eq.*, 37:70–88, 1980.

[44] R.L. Panton. *Incompressible flow.* Wiley-Interscience, 1984.

[45] M. Polner, L. Pesch, J.J.W. van der Vegt, and R.M.J. van Damme. A unified formulation of stabilization operators for Galerkin least-squares discretizations. *Comput. Meth. Appl. Mech. Engrg*, to be submitted.

[46] M. Polner, J.J.W. van der Vegt, and R.M.J. van Damme. Analysis of stabilization operators for Galerkin least-squares discretizations of the incompressible Navier-Stokes equations. *Comput. Meth. Appl. Mech. Engrg*, 2004.

[47] H. Schlichting. *Boundary layer theory.* Oxford Science Publications, 1998.

[48] Ch. Schwab. *p- and hp- Finite element methods. Theory and applications in solid and fluid mechanics.* Oxford Science Publications, 1998.

[49] J.V. Sengers, R.F. Kayser, et al. *Equations of state for fluids and fluid mixtures.* Elsevier, 2000.

[50] F. Shakib. *Finite element analysis of the compressible Euler and Navier-Stokes equations.* PhD thesis, Stanford University, 1988.

[51] F. Shakib, T.J.R. Hughes, and Z. Johan. A new finite element formulation for computational fluid dynamic: X. The compressible Euler and Navier-Stokes equations. *Comput. Meth. Appl. Mech. and Eng.*, 89:141–219, 1991.

[52] T.E. Tezduyar. Stabilized finite element formulations for incompressible flow computations. *Advances in applied mechanics*, 28:1–44, 1992.

[53] T.E. Tezduyar and T.J.R. Hughes. Development of time-accurate finite element techniques for first-order hyperbolic systems with particular emphasis on the compressible Euler equations. *NASA-Ames University Consortium Interchange, report No. NCA2-OR745-104.*, 1982.

[54] E.F. Toro. *Rieman solvers and numerical methods for fluid dynamics.* Springer-Verlag, Berlin Heidelberg, 1999.

[55] J.J.W. van der Vegt and H. van der Ven. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows. I. General formulation. *J. Comput. Phys.*, 182(2):546–585, 2002.

[56] H. van der Ven and J.J.W. van der Vegt. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows. II. Efficient flux quadrature. *Comput. Meth. Appl. Mech. Engrg*, 191:4747–4780, 2002.

[57] W. Wagner and A. Prus. The IAPWS formulation 1995 for the thermodynamic properties of ordinary water substance for general and scientific use. *J. Phys. Chem. Ref. Data*, 31(2):387–535, 2002.

[58] J.S. Wong, D.L. Darmofal, and J. Peraire. The solution of the compressible Euler equations at low Mach numbers using a stabilized finite element algorithm. *Comput. Meth. Appl. Mech. Engrg*, 190:5719–5737, 2001.

Samenvatting

In dit proefschrift behandelen wij problemen uit de stromingsmechanica vanuit een algemeen uitgangspunt en combineren wij technieken die oorspronkelijk zijn ontwikkeld voor compressibele en incompressibele stromingen in een algemener kader. Om de toepasbaarheid van een generieke aanpak te bestuderen is het nodig om een goede initiële formulering te kiezen. En daarmee is ook de keuze van de variabelen in de beschrijvende vergelijkingen van cruciaal belang. Zo zijn bijvoorbeeld conservatieve variabelen niet geschikt voor een generieke formulering omdat ze een singuliere limiet voor incompressibele stromingen als resultaat hebben. Wanneer entropievariabelen of de zogenaamde primitieve variabelen worden gebruikt dan is de incompressibele limiet van de Navier-Stokes vergelijkingen wel correct gedefinieerd, waardoor deze variabelen geschikt zijn om als uitgangspunt van een generieke formulering te fungeren. De formulering van de Navier-Stokes vergelijkingen in termen van deze variabelen wordt daarom in dit proefschrift in detail bestudeerd.

Aangezien elke groep van variabelen unieke karaktereigenschappen bezit zijn de nauwkeurigheid, stabiliteit, robuustheid en efficiëntie van berekeningen met de numerieke methode sterk afhankelijk van deze keuze.

De numerieke discretisatie, die we onderzoeken, is een tijdsdiscontinue Galerkin kleinste-kwadraten eindige elementen methode. Een essentieel onderdeel van deze en gerelateerde methoden is de stabilisatieoperator. In het algemeen is voor compressibele stromingen een stabilisatieoperator nodig om numerieke oscillaties in gebieden met discontinuïteiten of scherpe gradiënten, die niet nauwkeurig kunnen worden gerepresenteerd op het rekenrooster, te voorkomen. Voor incompressibile stromingen is de stabilisatieoperator ook cruciaal, omdat het in dat geval niet nodig is om elementen te ontwerpen, die aan de inf-sup stabiliteitsconditie voldoen. Ondanks dat er zeer verschillende redenen zijn om een stabilisatieoperator te gebruiken bij het oplossen van compressibele en incompressibele stromingen, toont dit proefschrift aan dat veel ideeën die zijn ontwikkeld in het ene veld ook toegepast kunnen worden in het andere veld.

Het belangrijkste ingrediënt om een generieke formulering te verkrijgen is de bepaling van een stabilisatiematrix die voor beide typen stromingen gebruikt kan worden. De

keuze van deze matrix is cruciaal om de stabiliteit van de numerieke discretisatie te verzekeren zonder de nauwkeurigheid in essentie aan te tasten. Bovendien zou de stabilisatiematrix die voor de incompressibele stromingen is ontwikkeld niet effectief kunnen zijn voor compressibele stromingen, maar ook andersom: de compressibele stabilisatiematrix zou niet correct gedefinieerd kunnen zijn in de incompressibele limiet. Dit proefschrift beschrijft een nieuwe techniek om stabilisatiematrices te ontwikkelen die gebruikt kunnen worden voor beide typen stromingen. De voorgestelde klasse van stabilisatiematrices is verkregen door middel van dimensieanalyse van de stromingsvariabelen. Bij de ontwikkeling van de stabilisatiematrices hebben we gebruik gemaakt van de voordelen van de entropievariabelen en de primitieve variabelen. De verkregen stabilisatiematrices zijn correct gedefinieerd in de incompressibele limiet van zowel de entropievariabelen als de primitieve variabelen. Dit beschouwen we als het belangrijkste resultaat van dit onderzoek. De voorgestelde klasse van stabilisatiematrices met de correcte dimensies wordt verder onderzocht om de stabiliteit van de Galerkin kleinste-kwadraten eindige elementen discretisatie van de gelinearizeerde incompressibele Navier-Stokes vergelijkingen en de niet-lineaire stabiliteit in het compressibele geval te verbeteren. Daarvoor geven we noodzakelijke en voldoende condities voor het positief-definiet zijn van de ontwikkelde stabilisatiematrix voor de entropievariabelen.

De tijdsdiscontinue Galerkin kleinste-kwadraten eindige elementen discretisatie resulteert in een groot systeem van niet-lineaire algebraïsche vergelijkingen. Een benadering van de ruimte-tijd Galerkin kleinste-kwadraten variationele vergelijking, die lineair is in de tijd, is nodig voor het oplossen van niet-stationaire problemen. In dit proefschrift stellen wij een nieuwe methode voor om het niet-lineaire algebraïsche systeem op te lossen en vergelijken wij het algoritme met een predictor multi-corrector-methode, waarbij gebruik gemaakt wordt van de advectie-diffusie vergelijking als modelprobleem.

De ontwikkelde stabilisatiematrix wordt met een aantal numerieke voorbeelden gedemonstreerd. De nadruk ligt hierbij op de invloed van de matrix op de nauwkeurigheid van de numerieke discretisatie. De numerieke voorbeelden tonen dat de nieuwe stabilisatiematrix goede resultaten geeft in het stabiliseren van de numerieke methode zonder de nauwkeurigheid aan te tasten als er primitieve variabelen worden gebruikt.