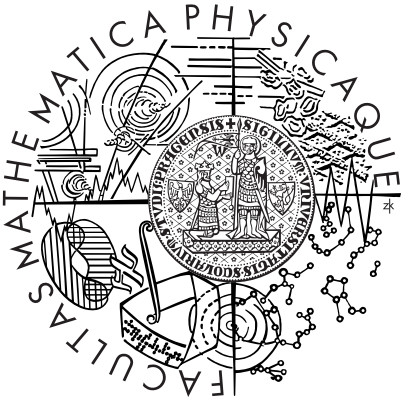
Charles University
Faculty of Mathematics and Physics

DOCTORAL THESIS



Václav Kučera

Higher order methods for the solution of compressible flows

Department of Numerical Mathematics
Supervisor: *Prof. RNDr. Miloslav Feistauer, DrSc., Dr. h. c.*

Title: Higher order methods for the solution of compressible flows
Author: Václav Kučera
Department: Department of Numerical Mathematics
Supervisor: Prof. RNDr. Miloslav Feistauer, DrSc., Dr. h c.
Author's e-mail address: vaclav.kucera@email.cz
Supervisor's e-mail address: feist@karlin.mff.cuni.cz

Abstract: This work is concerned with the theoretical analysis and practical applications of the discontinuous Galerkin finite element method. We derive a discontinuous Galerkin formulation for a model scalar convection-diffusion equation with nonlinear convection and diffusion. The resulting symmetric, nonsymmetric and incomplete variants are theoretically analyzed and error estimates in the $L^2(H^1)$- and $L^\infty(L^2)$-norms are derived. Since these error estimates are suboptimal in the latter norm, we use the Aubin-Nitsche technique to obtain $L^\infty(L^2)$-optimal estimates. The proof admits nonconforming meshes, however this result is limited to the symmetric variant and linear diffusion. Numerical experiments are performed to verify these theoretical results by the experimental order of convergence. Further, we apply the discontinuous Galerkin method to the compressible Euler equations using a semi-implicit discretization with respect to time. We discuss the choice of boundary conditions and shock capturing. Several numerical examples show the resulting scheme is capable of computing transonic, supersonic and low-Mach flows. Finally, we treat the compressible Navier-Stokes equations, incorporating viscous terms into the semi-implicit numerical scheme for the Euler equations. The extension of the discontinuous Galerkin formulation of second order terms to systems of equations is discussed and a new approach is derived. Several numerical examples for the Navier-Stokes equations are presented.

Keywords: Discontinuous Galerkin method, nonlinear convection-diffusion equation, error analysis, optimal error estimates, compressible Euler equations, numerical flux, compressible Navier-Stokes equations, semi-implicit linearized numerical scheme, shock capturing.

Acknowledgments

I would like to thank all those who supported me in my doctoral study and the work on my thesis. I very appreciate the help and guidance received from my supervisor Miloslav Feistauer and I am grateful for numerous remarks, corrections and advices he gave me throughout my work. I would like to thank Vít Dolejší for the help he provided me during the work on this thesis. I am also much obliged to Josef Málek, who enabled me to participate at various mathematical activities.

I wish to express my gratitude to Professor Phil Roe for providing the practically unknown reference [24] on the exact solution of a rotational incompressible flow problem and the recommendation to test our method on this interesting nonstandard example.

Last but not least, I am in debt to my parents, whose support and patience made this work possible.

My thanks also go to institutions that provided financial support for my research work. Through my doctoral study, my work was partially supported by the Grant GAUK 6/2005/R and by the Nečas Center for Mathematical Modelling, project LC06052, financed by MSMT and at the end partly by the grant 7486/2007 of the Charles University.

Contents

List of Figures

List of Tables

Introduction

The discontinuous Galerkin finite element method (DGFEM) is a promising numerical method for the solution of conservation laws and singularly perturbed problems. Such problems can be generally characterized by the presence of boundary layers, where the solution contains steep gradients or discontinuities. For nonlinear conservation laws with small dissipation, so-called shock waves (also called internal boundary layers) may appear. Such phenomena present a challenge for numerical methods, especially when higher order spatial discretizations are considered. The finite volume method (FVM) is often used in such cases. This method uses piecewise constant approximations and its generalization to higher orders is not straightforward. Another method often used is the finite element method (FEM), which usually uses conforming piecewise polynomial approximations. However when such a discretization is applied to conservation laws, undesired oscillations known as the Gibbs phenomenon arise near discontinuities and corrupt the solution. From this point of view the discontinuous Galerkin finite element method (DGFEM) has advantages of both the FEM and FVM. We obtain a higher order method without any requirements on inter-element continuity. This is compensated as in the FVM, where inter-element behavior is treated using an appropriate numerical flux. The DGFEM was applied to nonlinear conservation laws in 1989 by Cockburn and Shu [8]. Later, Bassi and Rebay used it to solve compressible flow in [3] and [4]. During several recent years DGFE schemes have been extensively developed and become more and more popular. Some aspects of the DGFEM and applications to gas dynamics are discussed in [1], [5], [23]. For a survey, see, for example [9] and [10].

In this thesis, we shall be concerned with the theoretical analysis and practical applications of the discontinuous Galerkin. The structure of the work is as follows: in Chapter 1, we shall derive a discontinuous Galerkin formulation of a model scalar convection-diffusion problem with nonlinear convection and diffusion. The DGFEM uses a piecewise polynomial finite element space without any assumption on inter-element continuity, which is replaced by the use of a suitable numerical flux as in the finite volume method. Three different discretizations of the second order diffusion term are considered, the so-called *symmetric*, *nonsymmetric* and *incomplete interior penalty* formulations. Numerical experiments are presented, for which the experimental order of convergence is calculated and compared with

theoretically obtained error estimates proved in Chapter 2.

The second chapter consists of two parts. After some necessary results and assumptions are stated, error estimates are derived for the discontinuous Galerkin discretization of the scalar nonlinear convection- diffusion equation as defined in Chapter 1. Under sufficient regularity assumptions on the exact solution, we prove $O(h^p)$ error estimates in the $L^2(H^1)$- and $L^\infty(L^2)$-norms when the discretization uses piecewise polynomials of order p. In the second part of this chapter, we refine the error estimate with respect to the $L^\infty(L^2)$-norm and derive an $O(h^{p+1})$ bound, which is optimal. Due to the use of the Aubin-Nitsche technique, this result is proved only for the symmetric variant and linear diffusion. However, the technique presented here admits nonconforming meshes with hanging nodes, which improves the result of [17].

In Chapter 3, we apply the DGFEM to the numerical solution of inviscid compressible flows governed by the Euler equations. We discuss the choice of an appropriate numerical flux and boundary conditions, which must be transparent for acoustic phenomena. The semi-implicit linearization with respect to time defined originally in [16] is applied and combined with appropriate shock capturing. The resulting numerical scheme requires the solution of one linear system on each time level, which is solved either using preconditioned GMRES or a sparse direct solver. Several numerical examples for inviscid transonic, supersonic and low-Mach flows are presented and whenever possible, compared with the known exact solution of incompressible flow.

Finally, in the last chapter, we apply the discontinuous Galerkin method to the full compressible Navier-Stokes equations describing viscous compressible flows. Here the main problem lies in the discretization of viscous terms, which are nonlinear with respect to the unknown state vector. The extension of the symmetric and nonsymmetric variants of the DGFEM from the scalar case to the case of nonlinear systems is not straightforward, We discuss two possibilities derived in [21], [14] and present another possibility based on a unified methodology for the analysis of discontinuous Galerkin discretizations of the Poisson equation presented in [2]. The resulting space semidiscretization is semi-implicitly linearized to yield a linear scheme with good stability properties. Again as in Chapter 3, we apply the DGFEM to several test cases.

Chapter 1

Discontinuous Galerkin method for a scalar model equation

In this chapter we shall be concerned with the discontinuous Galerkin finite element method applied to a scalar nonstationary nonlinear convection-diffusion equation, equipped with mixed Dirichlet-Neumann boundary conditions and an initial condition. We describe the symmetric *(SIPG),* nonsymmetric *(NIPG) and* incomplete interior penalty *(IIPG) discontinuous Galerkin finite element discretization of this problem. Further, we test the accuracy of the method and verify orders of convergence theoretically obtained in Chapter 2.*

1.1 Continuous problem

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a Lipschitz-continuous boundary $\partial\Omega = \Gamma_D \cup \Gamma_N, \Gamma_D \cap \Gamma_N = \emptyset$ and $T > 0$. We shall deal with the following *initial-boundary value problem*: find $u : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{2} \frac{\partial f_s(u)}{\partial x_s} = \mathrm{div}\big(\beta(\mathrm{u})\nabla \mathrm{u}\big) + \mathrm{g} \quad \text{in } \mathrm{Q_T}, \tag{1.1}$$

$$u|_{\Gamma_D \times (0,T)} = u_D, \tag{1.2}$$

$$\beta(u)\frac{\partial u}{\partial t}\Big|_{\Gamma_N \times (0,T)} = g_N, \tag{1.3}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{1.4}$$

The function $\beta(u)$ yields a nonlinear diffusion term. By setting $\beta(u) = \varepsilon$, for a constant $\varepsilon \in I\!R$, we obtain the more standard linear case of equation (1.1). Then the right-hand side diffusion term is $\mathrm{div}\big(\beta(\mathrm{u})\nabla \mathrm{u}\big) = \varepsilon \Delta \mathrm{u}$. Usually, we are concerned with the case $0 < \varepsilon \ll 1$, i.e. *dominant convection*.

Conditions (1.2) and (1.3) are the *Dirichlet* and *Neumann* boundary conditions, respectively, $g : Q_T \to \mathbb{R}$, $u_D : \Gamma_D \times (0,T)$, $u_N : \Gamma_N \times (0,T)$ and $u^0 : \Omega \to \mathbb{R}$

are given functions, $f_1, f_2 \in C^1(\mathbb{R})$ are given inviscid fluxes. Further, we assume that the function u_D is a trace of some $u^* \in C([0,T], H^1(\Omega)) \cap L^\infty(Q_T)$. We say that u is a *classical solution* of the scalar convection-diffusion problem, if it is sufficiently regular and satisfies (1.1) - (1.4) pointwise.

1.2 Discretization

As we are confined to two dimensional problems, we assume $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain (the case when Ω is not polygonal and has to be approximated by a different domain Ω_h requires special treatment and will be dealt with in Section 3.5).

Let $\mathcal{T}_h$ be a partition of the closure $\overline{\Omega}$ into a finite number of closed convex polygons, whose interiors are mutually disjoint. In the implementation of all algorithms in this work, we shall use triangular meshes with the usual conforming properties known from the finite element method. Now we shall introduce some notation convenient in DGFE formulations.

For any $K \in \mathcal{T}_h$, we set $|K| = \text{meas}_2(K)$ (two dimensional Lebesgue measure), $h_K = \text{diam}(K)$ –diameter of K, $h = \max_{K \in \mathcal{T}_h} h_K$. We define an index set $I \subset \mathbb{Z}^+ = \{0, 1, 2, \ldots\}$ such that all elements of $\mathcal{T}_h$ are numbered by indices from I, i.e. $\mathcal{T}_h = \{K_i\}_{i \in I}$. If two elements $K_i, K_j \in \mathcal{T}_h$ share a common face, which by definition must be a linear segment, we call them *neighbours* and set $\Gamma_{ij} = \partial K_i \cap \partial K_j$ and $d(\Gamma_{ij}) = \text{meas}_1 \Gamma_{ij}$ = length of Γ_{ij} . For $i \in I$ we define $s(i) = \{j \in I; K_j \text{ is a neighbour of } K_i\}$. The boundary $\partial\Omega$ is formed by a finite number of faces of elements K_i adjacent to $\partial\Omega$. We denote all these boundary faces by S_j, where $j \in I_b \subset \mathbb{Z}^- = \{-1, -2, \ldots\}$ and set $\gamma(i) = \{j \in I_b; S_j \text{ is a face of } K_i\}, \Gamma_{ij} = S_j$ for $K_i \in \mathcal{T}_h$, such that $S_j \subset \partial K_i, j \in I_b$. If K_i is not adjacent to $\partial\Omega$, we set $\gamma(i) = \emptyset$. Furthermore we define $S(i) = s(i) \cup \gamma(i)$.

We can see that

$$s(i) \cap \gamma(i) = \emptyset, \quad \partial K_i = \bigcup_{j \in S(i)} \Gamma_{ij}, \quad \partial K_i \cap \partial\Omega = \bigcup_{j \in \gamma(i)} \Gamma_{ij}. \tag{1.5}$$

If we are concerned with different types of boundary conditions (in our case Neumann and Dirichlet), for $i \in I$ we denote by $\gamma_D(i)$ and $\gamma_N(i)$ the subsets of $\gamma(i)$ such that $\bigcup_{j \in \gamma_D(i)} \Gamma_{ij}$ and $\bigcup_{j \in \gamma_N(i)} \Gamma_{ij}$ form the parts Γ_D and Γ_N, respectively. It is obvious that we suppose $\gamma(i) = \gamma_D(i) \cup \gamma_N(i), \gamma_D(i) \cap \gamma_N(i) = \emptyset,$ for all $i \in I$. By $\mathbf{n}_{ij}$ we denote the unit outer normal to ∂K_i on the face Γ_{ij}. In our case, $\mathbf{n}_{ij}$ is constant along Γ_{ij}.

Over $\mathcal{T}_h$ we define the *broken Sobolev space*

$$H^k(\Omega, \mathcal{T}_h) = \{v; v|_K \in H^k(K)\ \forall K \in \mathcal{T}_h\} \tag{1.6}$$

and for $v \in H^1(\Omega, \mathcal{T}_h)$ we set

$$\begin{aligned} v|_{\Gamma_{ij}} &= \textit{ trace of } v|_{K_i} \text{ on } \Gamma_{ij}, \\ \langle v \rangle_{\Gamma_{ij}} &= \frac{1}{2}(v|_{\Gamma_{ij}} + v|_{\Gamma_{ji}}), \textit{ average of traces of } v \text{ on } \Gamma_{ij}, \\ [v]_{\Gamma_{ij}} &= v|_{\Gamma_{ij}} - v|_{\Gamma_{ji}}, \textit{ jump of traces of } v \text{ on } \Gamma_{ij}, \end{aligned} \tag{1.7}$$

and also

$$|v|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{i \in I} |v|^2_{H^k(K_i)} \right)^{1/2}. \tag{1.8}$$

Finally, we define the space of discontinuous piecewise polynomial functions

$$S_h = S^{p,-1}(\Omega, \mathcal{T}_h) = \{v; v|_K \in P_p(K) \; \forall K \in \mathcal{T}_h\}, \tag{1.9}$$

where $P_p(K)$ is the space of all polynomials on K of degree $\leq p$.

1.3 DGFE formulation

The discrete problem is based on concepts from the finite element and finite volume methods. Let u be a classical solution of our problem. We multiply (1.1) by an arbitrary $\varphi \in H^2(\Omega, \mathcal{T}_h)$, integrate over element $K_i \in \mathcal{T}_h$, apply Green's theorem and obtain

$$\begin{aligned} &\int_{K_i} \frac{\partial u}{\partial t} \varphi \, dx + \int_{\partial K_i} \sum_{s=1}^{2} f_s(u) n_s \varphi \, dS - \int_{K_i} \sum_{s=1}^{2} f_s(u) \frac{\partial \varphi}{\partial x_s} \, dx \\ &+ \int_{K_i} \beta(u) \nabla u \cdot \nabla \varphi \, dx - \int_{\partial K_i} \beta(u) (\nabla u \cdot \mathbf{n}) \varphi \, dS = \int_{K_i} g(t) \varphi \, dx, \end{aligned} \tag{1.10}$$

where $\mathbf{n} = (n_1, n_2)$ denotes the unit outer normal to ∂K_i. Since $\nabla u = \langle \nabla u \rangle$ on any interior edge, it follows that

$$\sum_{i \in I} \sum_{j \in s(i)} \int_{\Gamma_{ij}} \nabla u \cdot \mathbf{n}_{ij} \varphi \, ds = \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j<i}} \int_{\Gamma_{ij}} \langle \nabla u \rangle \cdot \mathbf{n}_{ij} [\varphi] \, ds. \tag{1.11}$$

Now we sum (1.10) over all $i \in I$ and after some manipulation we get

$$\begin{aligned}
&\int_\Omega \frac{\partial u}{\partial t}\,\varphi\,dx + \sum_{i\in I}\sum_{j\in S(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2} f_s(u)n_s\varphi|_{\Gamma_{ij}}\,dS \\
&\quad - \sum_{i\in I}\int_{K_i}\sum_{s=1}^{2} f_s(u)\frac{\partial\varphi}{\partial x_s}\,dx + \sum_{i\in I}\int_{K_i}\beta(u)\nabla u\cdot\nabla\varphi\,dx \\
&\quad - \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\beta(u)\nabla u\rangle\cdot\mathbf{n}_{ij}[\varphi]\,dS - \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u)\nabla u\cdot\mathbf{n}_{ij}\varphi\,dS \\
&\quad = \int_\Omega g(t)\varphi\,dx + \sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}}\beta(u)\nabla u\cdot\mathbf{n}_{ij}\varphi\,dS.
\end{aligned} \tag{1.12}$$

In the second term on the right-hand side, we can use the Neumann boundary condition and replace this term by

$$\sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}} g_N\varphi\,dS. \tag{1.13}$$

Due to the regularity of u, it is clear that $[u(\cdot,t)]_{\Gamma_{ij}} = 0$ and therefore,

$$\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\beta(u)\nabla\varphi\rangle\cdot\mathbf{n}_{ij}[u]\,dS = 0. \tag{1.14}$$

We can add this term to the left-hand side of (1.12). To incorporate the Dirichlet boundary condition, we add the terms

$$\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u)\nabla\varphi\cdot\mathbf{n}_{ij}u\,dS \text{ and } \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u)\nabla\varphi\cdot\mathbf{n}_{ij}u_D\,dS \tag{1.15}$$

to the left- and right-hand side, respectively. This procedure leads to the *non-symmetric* (**NIPG**) discontinuous Galerkin discretization of diffusion terms. The *symmetric* (**SIPG**) variant is obtained by adding terms (1.15) with the '$-$' sign and in the *incomplete interior penalty* (**IIPG**) variant, we do not add these artificial terms to ou formulation.

Since we want to omit the continuity of the solution between elements of the triangulation, we shall compensate it by stabilizing the scheme with the aid of the *interior penalty* terms

$$\nu\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sigma[u][\varphi]\,dS \tag{1.16}$$

and the *boundary penalty* terms

$$\nu \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma u\varphi \, dS = \nu \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma u_D\varphi \, dS. \tag{1.17}$$

The parameter σ is a weight defined by

$$\sigma|_{\Gamma_{ij}} = C_W/d(\Gamma_{ij}). \tag{1.18}$$

The constant $C_W > 0$ must be chosen large enough to ensure *coercivity* of the resulting diffusion form (a detailed analysis will be carried out in Section 2.2). The constant ν must somehow reflect properties of $\beta(u)$. For instance in Section 2.2, we assume $\beta_0 < \beta(u) < \beta_1$ for some constants $\beta_0, \beta_1 > 0$. In this case we set $\nu = \beta_0$. In the linear case ($\beta(u) = \varepsilon$) we simply set $\nu = \varepsilon$ for a given constant $\varepsilon \geq 0$.

The boundary convective terms will be treated similarly as in the finite volume method, i.e. with the aid of a numerical flux $H(u, v, \mathbf{n})$:

$$\int_{\Gamma_{ij}} \sum_{s=1}^{2} f_s(u)\mathbf{n}_s\varphi|_{\Gamma_{ij}} \, dS \approx \int_{\Gamma_{ij}} H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \mathbf{n}_{ij})\varphi|_{\Gamma_{ij}} \, dS. \tag{1.19}$$

The choice of H and $u|_{\Gamma_{ji}}$ for boundary edges differs slightly from the definition in [14]. Here we use

$$u|_{\Gamma_{ji}} = \begin{cases} u_D & \text{on } \Gamma_D, \\ u|_{\Gamma_{ij}} & \text{otherwise.} \end{cases} \tag{1.20}$$

We shall discuss this choice in Section 1.5.1.

Now we can finally write down the following forms, defined for $u, \varphi \in H^2(\Omega, \mathcal{T}_h)$. *Nonsymmetric* diffusion form:

$$\begin{aligned} a_h^N(u,\varphi) = &\sum_{i\in I} \int_{K_i} \beta(u)\nabla u \cdot \nabla\varphi \, dx \\ &- \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \beta(u)\nabla u\rangle \cdot \mathbf{n}_{ij}[\varphi] \, dS + \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \beta(u)\nabla\varphi\rangle \cdot \mathbf{n}_{ij}[u] \, dS \\ &- \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \beta(u)\nabla u \cdot \mathbf{n}_{ij}\varphi \, dS + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \nabla\beta(u)\varphi \cdot \mathbf{n}_{ij} u \, dS, \end{aligned} \tag{1.21}$$

symmetric diffusion form:

$$\begin{aligned} a_h^S(u,\varphi) = &\sum_{i\in I}\int_{K_i}\beta(u)\nabla u\cdot\nabla\varphi\,dx \\ &-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\beta(u)\nabla u\rangle\cdot\mathbf{n}_{ij}[\varphi]\,dS-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\beta(u)\nabla\varphi\rangle\cdot\mathbf{n}_{ij}[u]\,dS \\ &-\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u)\nabla u\cdot\mathbf{n}_{ij}\varphi\,dS-\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\nabla\beta(u)\varphi\cdot\mathbf{n}_{ij}u\,dS \end{aligned} \tag{1.22}$$

and *incomplete* diffusion form:

$$\begin{aligned} a_h^I(u,\varphi) = &\sum_{i\in I}\int_{K_i}\beta(u)\nabla u\cdot\nabla\varphi\,dx \\ &-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\beta(u)\nabla u\rangle\cdot\mathbf{n}_{ij}[\varphi]\,dS-\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u)\nabla u\cdot\mathbf{n}_{ij}\varphi\,dS \end{aligned} \tag{1.23}$$

Further we define the interior and boundary penalty jump terms:

$$J_h(u,\varphi)=\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sigma[u][\varphi]\,dS+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sigma u\varphi\,dS, \tag{1.24}$$

nonsymmetric right-hand side:

$$\begin{aligned} l_h^N(u,\varphi)(t) = &\int_\Omega g(t)\varphi\,dx+\sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}}g_N(t)\varphi\,dS \\ &+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u)\nabla\varphi\cdot\mathbf{n}_{ij}u_D(t)\,dS+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sigma u_D(t)\varphi\,dS, \end{aligned} \tag{1.25}$$

symmetric right-hand side:

$$\begin{aligned} l_h^S(u,\varphi)(t) = &\int_\Omega g(t)\varphi\,dx+\sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}}g_N(t)\varphi\,dS \\ &-\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u)\nabla\varphi\cdot\mathbf{n}_{ij}u_D(t)\,dS+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sigma u_D(t)\varphi\,dS \end{aligned} \tag{1.26}$$

and the *incomplete interior penalty* right-hand side:

$$\begin{aligned} l_h^I(u,\varphi)(t) = &\int_\Omega g(t)\varphi\,dx \\ &+\sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}}g_N(t)\varphi\,dS+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sigma u_D(t)\varphi\,dS. \end{aligned} \tag{1.27}$$

Finally we define the *convective* terms:

$$\begin{aligned} b_h(u,\varphi) = &- \sum_{i\in I} \int_{K_i} \sum_{s=1}^{2} f_s(u) \frac{\partial \varphi}{\partial x_s}\, dx \\ &+ \sum_{i\in I} \sum_{j\in S(i)} \int_{\Gamma_{ij}} H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \mathbf{n}_{ij}) \varphi|_{\Gamma_{ij}}\, dS. \end{aligned} \tag{1.28}$$

Now we can introduce the discrete problem (space semidiscretization with continuous time, also called the method of lines). For simplicity of notation, we omit the superscripts *N, S* and *I* and use the generic notation for the diffusion and right-hand side forms $a_h(u_h,\varphi)$ and $l_h(u,\varphi_h)$. The *symmetric, nonsymmetric* and *incomplete* variants can be obtained by taking in turn $a_h := a_h^S$, $l_h := l_h^S$ and so on.

Definition 1.3.1 *We say that u_h is a DGFE solution of the convection-diffusion problem (1.1) - (1.4), if*

$$\begin{aligned} &a)\ u_h \in C^1([0,T]; S_h), \\ &b)\ \frac{d}{dt}(u_h(t),\varphi_h) + b_h(u_h(t),\varphi_h) + \nu J_h(u_h(t),\varphi_h) + a_h(u_h(t),\varphi_h) \\ &\qquad = l_h(u,\varphi_h)(t), \quad \forall \varphi_h \in S_h,\ \forall t \in (0,T), \\ &c)\ u_h(0) = u_h^0, \end{aligned} \tag{1.29}$$

where u_h^0 is an S_h approximation of the initial condition u^0.

1.4 Time discretization

We proceed as in the finite element method. Let $\mathcal{B} = \{w_\alpha\}_{\alpha=1}^n$ be a basis in the space S_h, with $n = \dim S_h$. We seek the approximate solution $u_h \in S_h$ in the form

$$u_h(t) = \sum_{\alpha=1}^{n} \xi_\alpha(t) w_\alpha. \tag{1.30}$$

Due to the linearity of the forms (1.21) - (1.28) in the variable φ, we can use, as test functions in (1.29), only elements of the basis $\mathcal{B}$. This leads to a system of n ordinary differential equations for unknowns $\xi_\alpha(t)$, $\alpha = 1,\dots,n$.

In practice, we also need a *time discretization* of the problem by a suitable method. In the implementation of problem (1.29) we have used the simplest approach, which is the *Euler forward scheme*. This has two disadvantages: it is only first order accurate, and a severe limitation on the time step has to be applied in order to respect a *CFL-like* stability condition. However, this method is simple from an implementational point of view and we are mostly interested in

the steady-state solution as $t \to \infty$, if this exists, therefore first order accuracy is not a drawback.

Let $0 = t_0 < t_1 < \dots$ be a partition of the time interval $[0, T]$, and $\tau_k = t_{k+1} - t_k$. The Euler forward scheme has the simple form:

$$\begin{aligned}(u_h^{k+1},\varphi_h) + \tau_k\big[b_h(u_h^k,\varphi_h) + \nu J_h(u_h^k,\varphi_h) + a_h(u_h^k,\varphi_h)\big] \\ = \tau_k l(\varphi_h)(t_k) + (u_h^k,\varphi_h), \quad \forall\varphi_h \in S_h,\ k = 0,1,\dots.\end{aligned} \tag{1.31}$$

Let $\boldsymbol{\xi}^k = (\xi_1^k, \dots, \xi_n^k)$, where ξ_α^k is an approximation of $\xi_\alpha(t_k)$. Then the Euler forward scheme can be written in the form of a system of n linear equations:

$$\mathbf{M}\boldsymbol{\xi}^{k+1} = \mathbf{M}\boldsymbol{\xi}^k + \tau\mathbf{a}(\boldsymbol{\xi}^k), \tag{1.32}$$

where $\mathbf{a}(\boldsymbol{\xi}^k)$ is a vector-valued mapping corresponding to the forms (1.21) - (1.28) and $\mathbf{M} = \{m_{ij}\}_{i,j=1}^n$ is a $n \times n$ matrix, the so-called *mass matrix*, with entries $m_{ij} = \int_\Omega w_i w_j\, dx$.

In the finite element method, we usually choose the basis functions with supports as small as possible. In the discontinuous case, the support of a basis function can be exactly one element. If functions w_α, w_β have different support elements, then $m_{\alpha\beta} = 0$. This is advantageous, because the mass matrix $\mathbf{M}$ will not only be sparse, but by "clustering" the basis functions with common support elements, we can achieve that $\mathbf{M}$ will be *block-diagonal* with $n_p \times n_p$ blocks, where $n_p = \dim P_p(K)$. If $\mathbf{M} = \mathrm{diag}\{\mathbf{B}_1, \dots, \mathbf{B}_n\}$, then $\mathbf{M}^{-1} = \mathrm{diag}\{\mathbf{B}_1^{-1}, \dots, \mathbf{B}_n^{-1}\}$ and the system (1.32) can be solved quickly by inverting each $\mathbf{B}_i$, the so-called *local* mass matrices corresponding to element K. This can be done beforehand and the inversions $\mathbf{B}_i^{-1}$ can be calculated explicitly, so that we need not include any iterative linear solvers in the process of solution.

1.5 Practical Implementation

1.5.1 Numerical flux

The numerical experiments in Section 1.6 were carried out with the following numerical flux:

$$H(u_1, u_2, \mathbf{n}) = \begin{cases} \sum_{s=1}^2 f_s(u_1)n_s, & \text{if } A > 0 \\ \sum_{s=1}^2 f_s(u_2)n_s, & \text{if } A \le 0, \end{cases} \tag{1.33}$$

where

$$A = \sum_{s=1}^2 f_s'(\overline{u}), \qquad \overline{u} = \frac{1}{2}(u_1 + u_2). \tag{1.34}$$

This flux is based on the concept of *upwinding* and the Vijayasundaram numerical flux, which will be used in the discretization of the Euler equations, can be viewed as a generalization of this scalar case.

j	$x_j^{(1)}$-coordinate	$x_j^{(2)}$-coordinate	α_j
1.	0.333333333333333	0.333333333333333	0.225
2.	0.470142064105115	0.470142064105115	0.132394152788506
3.	0.470142064105115	0.05971587178977	0.132394152788506
4.	0.05971587178977	0.470142064105115	0.132394152788506
5.	0.101286507323456	0.101286507323456	0.125939180544827
6.	0.101286507323456	0.797426985353087	0.125939180544827
7.	0.797426985353087	0.101286507323456	0.125939180544827

Table 1.1: Gauss seven point rule on the reference triangle $\hat{K}$.

When computing convective terms, it is necessary to define the meaning of $u|_{\Gamma_{ji}}$ if $j \in \gamma(i)$ (i.e. when $\Gamma_{ij} \subset \partial\Omega$). In [14], extrapolation is used: $u|_{\Gamma_{ji}} := u|_{\Gamma_{ij}}$. However, this gives unsatisfactory results on Γ_D. In the general case, numerical experiments show, that the solution exhibits spurious overshoots undershoots at Γ_D when using extrapolation. The solution to this problem is to set $u|_{\Gamma_{ji}} := u_D$ for $\Gamma_{ji} = \Gamma_{ij} \subset \Gamma_D$. When this is done, such undesired phenomena do not occur, as can be seen in Section 1.6.

1.5.2 Numerical integration

In practice, the DGFE formulation requires calculation of terms in the form $\int_{K_i} f(x)\,dx$ and $\int_{\Gamma_{ij}} s(x)\,dS$. It is useful to evaluate element integrals over a common *reference element* $\hat{K}$, which is the triangle with vertices $(0,0),(1,0),(0,1)$. Edge integrals are evaluated on the unit interval $\hat{\Gamma} = [0,1]$. This is realized by the substitution theorem, and the subject will be thoroughly treated in Section 3.5 in the case of an isoparametric approximation of the boundary.

For the evaluation of integrals over $\hat{\Gamma}$ and $\hat{K}$ we use 1D and 2D Gaussian quadrature formulae of high order of accuracy: both are accurate for polynomials with degree ≤ 5. In 2D it is the seven point rule,

$$\int_{\hat{K}} f(x)\,dx \approx \sum_{j=1}^{7} \alpha_j f(x_j), \tag{1.35}$$

where α_j and x_j are given in Table 1.1. In 1D we use the three point rule,

$$\int_0^1 s(x)\,dx \approx \sum_{j=1}^{3} \beta_j f(x_j), \tag{1.36}$$

where β_j and $x_j = (x_j^{(1)}, x_j^{(2)})$ are given in Table 1.2. Details on higher order quadrature rules applicable in finite element methods can be found in [34].

j	x_j	α_j
1.	$(1-\sqrt{3/5})/2$	5/18
2.	0.5	4/9
3.	$(1+\sqrt{3/5})/2$	5/18

Table 1.2: Gauss three point rule on the unit interval $\hat{\Gamma}$.

1.5.3 Basis functions

For P_1, i.e. linear elements, the basis $\{\varphi_{in} \in S_h; i \in I, n = 1,2,3\}$ is used, such that $\varphi_{in}(P_{i'}^{n'}) = \delta_{ii'}\delta_{nn'}$, where $P_i^n, n = 1,2,3$, are vertices of element K_i and δ is the Kronecker symbol. For P_2, i.e. quadratic elements, the basis $\{\psi_{in} \in S_h; i \in I, n = 1,\dots,6\}$ is used, such that $\psi_{in}(P_{i'}^{n'}) = \delta_{ii'}\delta_{nn'}$, where $P_i^n, n = 1,2,3$, are vertices of element K_i and $P_i^n, n = 4,5,6$, are midpoints of edges of K_i. These are standard local basis functions as known from the finite element method and they work quite well. Experiments were done with the simple monomial basis $1, x, y, x^2, y^2, xy, \dots$ as an alternative, for which evaluation is simpler. However, the latter basis is very "non-orthogonal" compared to the first one, and the local mass matrices $\mathbf{B}_i$ are ill-conditioned, causing a great loss of accuracy.

1.6 Numerical examples

The above DGFE space semidiscretization scheme is theoretically analyzed in Chapter 2, where error estimates are derived. The main theorem gives the following estimates for $e_h(t) := u_h(\cdot,t) - u(\cdot,t)$:

$$
\begin{aligned}
&||e_h(t)||_{L^2(\Omega)} = \begin{cases} O(h^{p+1}) & \text{for the } \textit{symmetric} \text{ variant,} \\ O(h^p) & \text{otherwise,} \end{cases} \\
&\int_0^t |e_h(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)}\, d\vartheta = O(h^{2p}), \\
&\int_0^t J_h(e_h(\vartheta), e_h(\vartheta))\, d\vartheta = O(h^{2p}),
\end{aligned}
\tag{1.37}
$$

where p is the order of polynomials from $P^p(K)$ and $t \in [0,T]$. However, the question of optimal L^2-error estimates for the *nonsymmetric* and *incomplete* variants is still open, since the following numerical experiments indicate that these variants converge faster than $O(h^p)$ in the L^2-norm when p is odd.

Let us consider the 2D viscous Burgers equation

$$
\frac{\partial u}{\partial t} + \frac{1}{2}\sum_{s=1}^{2} \frac{\partial (u^2)}{\partial x_s} = \varepsilon \Delta u + g \quad \text{in } \Omega \times (0,T), \tag{1.38}
$$

	NIPG	IIPG	SIPG
p=1	1.0	5.0	10.0
p=2	1.0	5.0	15.0

Table 1.3: Choice of C_W for individual variants applied to the viscous Burgers equation, $\varepsilon = 0.002$.

	NIPG	IIPG	SIPG
p=1	1.0	5.0	20.0
p=2	1.0	5.0	20.0

Table 1.4: Choice of C_W for individual variants applied to the viscous Burgers equation, $\varepsilon = 0.1$.

where $\Omega = (0,1)^2$, equipped with a Dirichlet boundary condition, i.e. $\Gamma_N = \emptyset$. The function g and the initial and boundary conditions are defined in such a way that the exact solution has the form

$$u(x_1, x_2, t) = \big[sin\big(4(x_1 + x_2 - x_1x_2)\big) + sin(5x_1x_2)\big](1 - e^{-t}). \tag{1.39}$$

We conduct two experiments, first we set $\varepsilon = 0.002$ and in the second case $\varepsilon = 0.1$.

We discretize equation (1.38) by all three variants of the scheme presented in Section 1.3 using piecewise linear ($p = 1$) and piecewise quadratic ($p = 2$) elements. The choice of the parameter C_W from (1.18) in each of the six cases is given in Table 1.3. In Section 2.2, where error estimates are derived, we assume only that $C_W > 0$ in the *nonsymmetric* variant. However, in the *incomplete* case, we need C_W larger than some constant (2.74) and in the *symmetric* case, the lower bound for C_W is twice as large (2.63). These results led us to the choices given in Tables 1.3 and 1.4. In the *symmetric* variant we have experienced that the lower bound for C_W increases with p, since for $p = 2$, the choice $C_W = 10$ sufficient for $p = 1$, $\varepsilon = 0.002$ proved to be insufficient and led to instabilities in other cases.

l	1	2	3	4	5	6
h_l	3,95E-01	2,54E-01	1,78E-01	1,27E-01	8,83E-02	5,72E-02
$\#\mathcal{T}_{h_l}$	126	289	597	1177	2354	5938

Table 1.5: Triangulation data.

The resulting system of ordinary differential equations is solved by the Euler forward method presented in Section 1.4 with a very small time step $\approx \tau = 10^{-5}$,

in order to guarantee stability and sufficiently accurate resolution with respect to time. We investigate the error at time $t_0 = 1$ for $\varepsilon = 0.002$ and $t_0 = 0.1$ for $\varepsilon = 0.1$, respectively – this is due to the an extremely restrictive conditions on the time step of the explicit method. We take $E_h = ||e_h(t_0)||_{L^2(\Omega)}$ or $|e_h(t_0)|_{H^1(\Omega,\mathcal{T}_h)}$ or $J_h(e_h(t_0), e_h(t_0))^{1/2}$ and suppose that

$$E_h \approx Ch^\alpha, \tag{1.40}$$

where $C > 0$ is a constant independent of h and α is the *order of accuracy* of the method. The numerical solution was computed on 6 unstructured meshes $(\mathcal{T}_{h_l}, l = 1, \ldots, 6)$ with descending h_l. The values of h_l and the number of elements $\#\mathcal{T}_{h_l}$ are given in Table 1.5. We define the *local order of accuracy* by

$$\alpha_l = \frac{\log\left(e_{h_l}/e_{h_{l-1}}\right)}{\log\left(h_l/h_{l-1}\right)}, \quad l = 2, \ldots, 6 \tag{1.41}$$

and $\overline{\alpha}$ is defined as the average of $\alpha_l,\ l = 2, \ldots, 6$.

For $\varepsilon = 0.002$, tables 1.6 - 1.8 show the results for piecewise linear elements, and tables 1.9 - 1.11 show the results for piecewise quadratic elements. One can see that, in the case of dominating convection, all three variants have an experimental order of accuracy in the L^2-norm $O(h^{p+1})$, and the *incomplete* variant has the smallest error in this norm.

For $\varepsilon = 0.1$, results are given in tables 1.12 - 1.14 (piecewise linear elements) and 1.15 - 1.17 (piecewise quadratic elements). One can see that, in the piecewise linear case, all three variants have an optimal experimental order of accuracy in the L^2-norm as $O(h^{p+1})$, and again, the *incomplete* variant has the smallest error in this norm among all three variants. However, in the piecewise quadratic case, only the *symmetric* variant exhibits optimal convergence. This is in agreement with previous results for the Poisson equation, which indicate the nonsymmetric variant is suboptimal in the L^2-norm if p is even. Since the *incomplete* variant can be viewed as an average between symmetric and nonsymmetric, we can expect the same behavior.

Finally we note that the *symmetric* variant exhibited worse stability properties with respect to the explicit time discretization than the other two cases. Generally, the time step had to be chosen two to five times smaller than in the other variants to achieve stability. In Figure 1.1 the exact solution and its contours are depicted (top) along with three approximate solutions in the following order from top to bottom: piecewise linear elements on $\mathcal{T}_{h_1}$, piecewise quadratic elements on $\mathcal{T}_{h_3}$ and piecewise quadratic elements on $\mathcal{T}_{h_6}$. We can observe that the piecewise polynomial discontinuous solution tends to the continuous solution, since interelement jumps diminish with $h \to 0$. This is a consequence of the error estimate for $J_h(e_h, e_h)$ in (1.37). The depicted approximate solutions were obtained by the IIPG scheme with $\varepsilon = 0.002$

1.7 Conclusion

We have introduced a discontinuous Galerkin discretization of a model scalar convection-diffusion equation with three possibilities how to treat the second order (diffusion) term. Numerical experiments have verified theoretical error estimates for the $H^1(\Omega, \mathcal{T}_h)$ seminorm and penalty terms $J_h(\cdot,\cdot)^{1/2}$. As for the $L^2(\Omega)$ norm, the *incomplete* and *nonsymmetric* variants exhibit optimal order of convergence when $p = 1$ is odd, and suboptimal for $p = 2$. This phenomenon has been conjectured for the Poisson equation and, to the authors knowledge, has not yet been proved. Among the three tested variants, the incomplete case seems to be the best compromise. As opposed to the symmetric variant, it is more robust with respect to the choice of parameter C_W and the choice of the time step. The $L^2(\Omega)$ norm of the error was smaller than in the other two cases, whenever optimal convergence rate was attained (dominating convection and/or p odd). And finally, unlike the (non)symmetric case, the generalization to systems of equations is straightforward – Chapter 4.

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	1.28E-01	—	3.12E-00	—	6.55E-02	—
2	5.23E-02	2.028	2.02E-00	0.990	4.33E-02	0.941
3	2.42E-02	2.166	1.39E-00	1.056	2.89E-02	1.133
4	1.28E-02	1.907	1.00E-00	0.971	2.14E-02	0.897
5	6.76E-03	1.739	7.18E-01	0.908	1.57E-02	0.845
6	2.69E-03	2.125	4.27E-01	1.194	9.27E-03	1.216
$\overline{\alpha}$		1,993		1,024		1,006

Table 1.6: Computational errors for $\varepsilon = 0.002$, $p = 1$, *nonsymmetric* variant (NIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	1.30E-01	—	3.15E-00	—	6.24E-02	—
2	5.21E-02	2.071	2.04E-00	0.987	3.94E-02	1.047
3	2.34E-02	2.248	1.41E-00	1.046	2.52E-02	1.258
4	1.21E-02	1.968	1.02E-00	0.959	1.81E-02	0.993
5	6.11E-03	1.867	7.34E-01	0.902	1.26E-02	0.999
6	2.26E-03	2.287	4.38E-01	1.186	6.85E-03	1.390
$\overline{\alpha}$		2,088		1,016		1,137

Table 1.7: Computational errors for $\varepsilon = 0.002$, $p = 1$, *incomplete* variant (IIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	1.33E-01	—	3.20E-00	—	5.95E-02	—
2	5.32E-02	2.081	2.07E-00	0.986	3.64E-02	1.115
3	2.38E-02	2.263	1.44E-00	1.034	2.27E-02	1.337
4	1.24E-02	1.959	1.04E-00	0.949	1.60E-02	1.043
5	6.19E-03	1.887	7.52E-01	0.894	1.06E-02	1.115
6	2.31E-03	2.273	4.51E-01	1.176	5.52E-03	1.507
$\overline{\alpha}$		2,093		1,008		1,224

Table 1.8: Computational errors for $\varepsilon = 0.002$, $p = 1$, *symmetric* variant (SIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	1.90E-02	—	7.91E-01	—	1.32E-02	—
2	5.46E-03	2.834	3.82E-01	1.649	6.68E-03	1.550
3	1.65E-03	3.368	1.72E-01	2.245	2.90E-03	2.351
4	6.73E-04	2.678	8.89E-02	1.975	1.50E-03	1.968
5	2.59E-04	2.600	4.40E-02	1.917	7.61E-04	1.850
6	8.37E-05	2.603	1.62E-02	2.307	2.77E-04	2.329
$\overline{\alpha}$		2,817		2,019		2,010

Table 1.9: Computational errors for $\varepsilon = 0.002$, $p = 2$, *nonsymmetric* variant (NIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	1.86E-02	—	8.01E-01	—	1.21E-02	—
2	4.78E-03	3.087	3.81E-01	1.689	5.72E-03	1.696
3	1.32E-03	3.627	1.72E-01	2.233	2.45E-03	2.389
4	5.00E-04	2.895	8.91E-02	1.966	1.27E-03	1.951
5	1.64E-04	3.041	4.41E-02	1.922	6.20E-04	1.967
6	6.52E-05	2.124	1.62E-02	2.301	2.25E-04	2.337
$\overline{\alpha}$		2,955		2,022		2,068

Table 1.10: Computational errors for $\varepsilon = 0.002$, $p = 2$, *incomplete* variant (IIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	1.91E-02	—	8.12E-01	—	1.04E-02	—
2	5.26E-03	2.933	3.90E-01	1.665	4.70E-03	1.809
3	1.56E-03	3.416	1.79E-01	2.201	1.93E-03	2.510
4	5.94E-04	2.886	9.39E-02	1.918	9.87E-04	1.995
5	2.14E-04	2.789	4.71E-02	1.883	4.56E-04	2.110
6	8.05E-05	2.247	1.75E-02	2.283	1.54E-04	2.497
$\overline{\alpha}$		2,854		1,990		2,184

Table 1.11: Computational errors for $\varepsilon = 0.002$, $p = 2$, *symmetric* variant (SIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	2.02E-02	—	4.26E-01	—	1.28E-01	—
2	8.95E-03	1.851	2.78E-01	0.965	8.46E-02	0.932
3	4.18E-03	2.145	1.90E-01	1.073	5.86E-02	1.036
4	2.22E-03	1.884	1.39E-01	0.943	4.31E-02	0.914
5	1.15E-03	1.811	9.94E-02	0.906	3.10E-02	0.901
6	4.31E-04	2.249	6.07E-02	1.136	1.89E-02	1.138
$\overline{\alpha}$		1.988		1.005		0.984

Table 1.12: Computational errors for $\varepsilon = 0.1$, $p = 1$, *nonsymmetric* variant (NIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	1.46E-01	—	4.32E-1	—	7.33E-01	—
2	6.09E-03	1.980	2.81E-01	0.982	4.28E-02	1.224
3	2.77E-03	2.214	1.94E-01	1.048	2.71E-02	1.286
4	1.52E-03	1.793	1.41E-01	0.937	1.94E-02	0.992
5	7.61E-04	1.890	1.03E-01	0.870	1.34E-02	1.013
6	2.77E-04	2.326	6.21E-02	1.162	7.91E-03	1.216
$\overline{\alpha}$		2.040		1.000		1.146

Table 1.13: Computational errors for $\varepsilon = 0.1$, $p = 1$, *incomplete* variant (IIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	2.48E-02	—	4.77E-01	—	9.71E-02	—
2	8.41E-03	2.458	3.12E-01	0.962	2.28E-02	3.291
3	3.88E-03	2.179	2.16E-01	1.046	1.19E-02	1.829
4	2.07E-03	1.873	1.57E-01	0.940	8.49E-03	1.010
5	1.08E-03	1.779	1.15E-01	0.864	5.01E-03	1.440
6	3.92E-04	2.331	6.86E-02	1.180	2.77E-03	1.362
$\overline{\alpha}$		2.124		0.998		1.786

Table 1.14: Computational errors for $\varepsilon = 0.1$, $p = 1$, *symmetric* variant (SIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	5.17E-03	—	1.13E-01	—	2.88E-02	—
2	1.96E-03	2.205	5.50E-02	1.632	1.32E-02	1.775
3	8.42E-04	2.376	2.59E-02	2.121	6.08E-03	2.175
4	4.35E-04	1.970	1.36E-02	1.930	3.17E-03	1.947
5	2.11E-04	1.980	6.91E-03	1.842	1.60E-03	1.871
6	8.35E-05	2.132	2.60E-03	2.247	5.93E-04	2.279
$\overline{\alpha}$		2.133		1.955		2.010

Table 1.15: Computational errors for $\varepsilon = 0.1$, $p = 2$, *nonsymmetric* variant (NIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	2.42E-03	—	1.03E-01	—	1.81E-02	—
2	1.05E-03	1.893	4.96E-02	1.662	8.53E-03	1.709
3	4.21E-04	2.581	2.31E-02	2.150	4.06E-03	2.094
4	2.26E-04	1.852	1.21E-02	1.923	2.17E-03	1.871
5	1.07E-04	2.041	6.19E-03	1.840	1.10E-03	1.855
6	4.32E-05	2.090	2.32E-03	2.258	4.22E-04	2.203
$\overline{\alpha}$		2.092		1.967		1.947

Table 1.16: Computational errors for $\varepsilon = 0.1$, $p = 2$, *incomplete* variant (IIPG).

l	$\|\|e_h(t)\|\|_{L^2(\Omega)}$	α_l	$\|e_h(t)\|_{H^1(\Omega,\mathcal{T}_h)}$	α_l	$J_h(e_h(t),e_h(t))^{1/2}$	α_l
1	2.13E-03	—	1.08E-01	—	6.94E-03	—
2	6.88E-04	2.562	5.23E-02	1.653	3.41E-03	1.617
3	2.26E-04	3.130	2.42E-02	2.169	1.52E-03	2.278
4	9.12E-05	2.714	1.28E-02	1.895	8.28E-04	1.804
5	3.28E-05	2.787	6.56E-03	1.833	3.99E-04	1.996
6	7.50E-06	3.400	2.44E-03	2.277	1.50E-04	2.244
$\overline{\alpha}$		2.919		1.965		1.988

Table 1.17: Computational errors for $\varepsilon = 0.1$, $p = 2$, *symmetric* variant (SIPG).

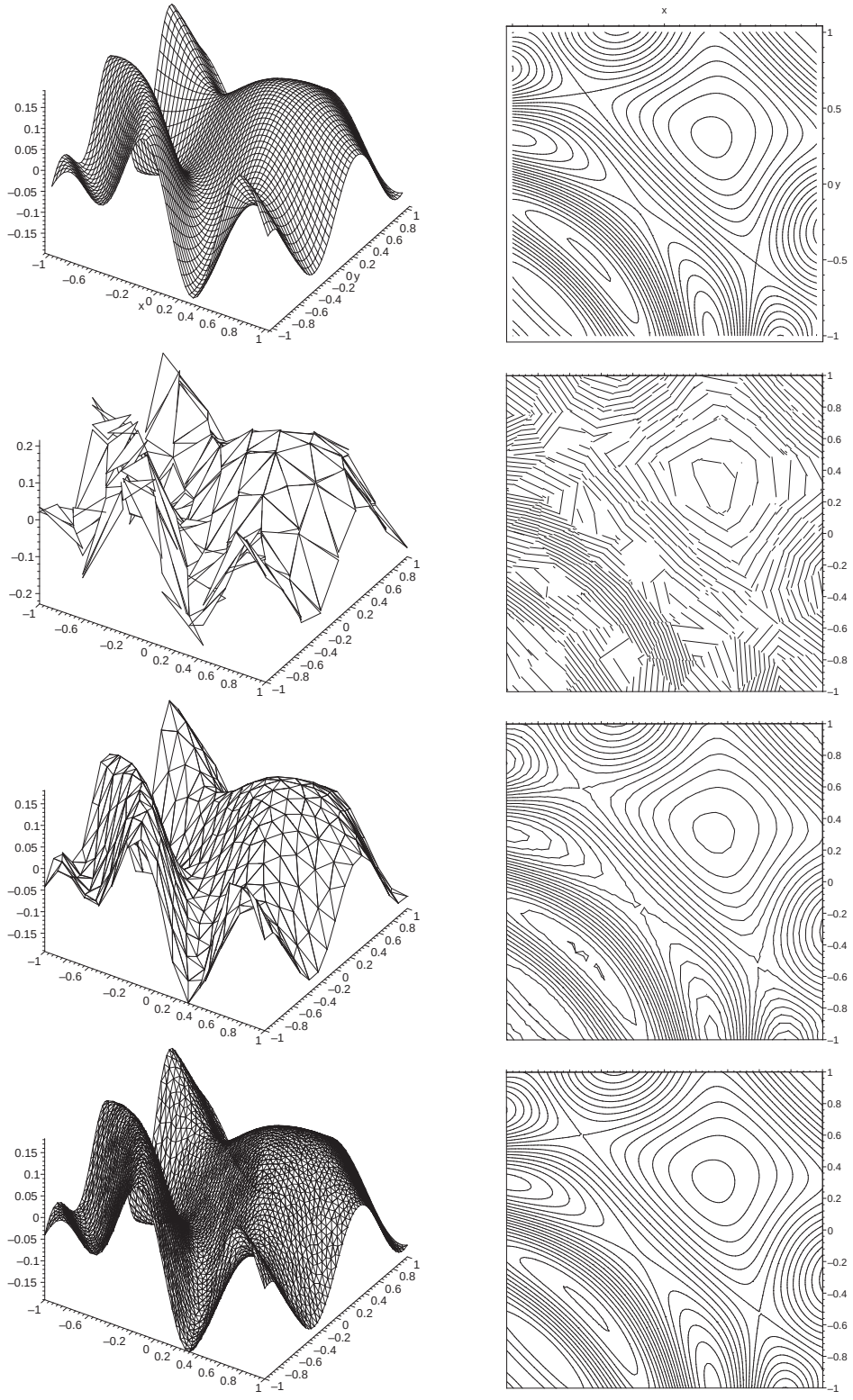


Figure 1.1: Function values and contours from top to bottom: 1) exact solution, 2) P^1 approximate solution on $\mathcal{T}_{h_1}$, 3) P^2 approximate solution on $\mathcal{T}_{h_3}$ and 4) P^2 approximate solution on $\mathcal{T}_{h_6}$.

Chapter 2

Error Estimates for DGFEM

In this chapter, we analyze the three variants of the discontinuous Galerkin method when applied to a scalar model problem. Section 2.2 deals with error estimates in the $L^2(H^1)$- and $L^\infty(L^2)$-norms. In this case we derive estimates for all three variants applied to a convection diffusion equation with nonlinear convection and *nonlinear diffusion. The derived estimates are optimal in the $L^2(H^1)$-norm, but suboptimal in the $L^\infty(L^2)$-norm. This problem is addressed in Section 2.3, where we derive $L^\infty(L^2)$-optimal error estimates. Since specific techniques are applied, the result holds only for the* symmetric *variant and for* linear diffusion*. Additional assumptions need to be imposed also on the computational domain: Ω is a* convex *polygonal domain and $\Gamma_N = \emptyset$. On the other hand, all presented results hold for rather general triangular meshes with hanging nodes, which improves the result of [17] concerning $L^\infty(L^2)$-optimal error estimates.*

2.1 Some necessary results and assumptions

First we shall introduce some notation. Let $G \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded domain with a Lipschitz-continuous boundary ∂G. By $\overline{G}$ we denote the closure of G. Further, let $k \in \{0, 1, 2, \dots\}$ and $p \in [1, \infty]$. We use the well-known Lebesgue and Sobolev spaces $L^p(G)$, $L^p(\partial G)$, $W^{k,p}(G)$, $H^k(G) = W^{k,2}(G)$, $W^{k,p}(\partial G)$, Bochner spaces $L^p(0, T; X)$ of functions defined in $(0, T)$ with values in a Banach space X and the spaces $C^k([0, T]; X)$ of k-times continuously differentiable mappings of the interval $[0, T]$ with values in X (see e.g. [28]). The symbols $\|\cdot\|_X$ and $|\cdot|_X$ will denote a norm and a seminorm in a space X, respectively.

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a polyhedral domain. Throughout this chapter we denote by C a generic constant independent of h and parameters of the problem, p denotes the order of approximation, where $S_h = \{v; v|_K \in P_p(K)\ \forall K \in \mathcal{T}_h\}$.

2.1.1 Geometry of the mesh

Let us consider a system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$, $h_0 > 0$, of partitions of the domain $\overline{\Omega}$ into a finite number of closed triangles (if d = 2) or tetrahedra (if d = 3) K with mutually disjoint interiors., i.e. $\mathcal{T}_h = \{K_i\}_{i\in I}, I \subset Z^+$.

Let us assume that the system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ has the following properties:

(A1) The system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ is regular: there exists a constant $C_1 > 0$ such that

$$\frac{h_K}{\rho_K} \le C_1 \quad \forall K \in \mathcal{T}_h \quad \forall h \in (0, h_0). \tag{2.1}$$

(A2) There exists a constant $C_2 > 0$ such that

$$h_{K_i} \le C_2\, d(\Gamma_{ij}), \quad \forall i \in I, \quad \forall j \in S(i), \quad \forall h \in (0, h_0). \tag{2.2}$$

Let us note that we do not require the usual conforming properties from the finite element method, particularly, hanging nodes are allowed. Condition (A2) means that the faces Γ_{ij} do not degenerate with respect to h_{K_i} for $h \to 0+$

2.1.2 Some auxiliary results

Now we can state two necessary results from [19] needed in the following analysis:

Lemma 2.1.1 (Multiplicative trace inequality) *There exists a constant $C_M > 0$ independent of h, K such that*

$$\begin{aligned} ||v||^2_{L^2(\partial K)} &\le C_M\big(||v||_{L^2(K)}|v|_{H^1(K)} + h_K^{-1}||v||^2_{L^2(K)}\big), \\ &\forall K \in \mathcal{T}_h,\ v \in H^1(K),\ h \in h_0. \end{aligned} \tag{2.3}$$

Lemma 2.1.2 (Inverse inequality) *There exists a constant $C_I > 0$ independent of h, K such that*

$$|v|_{H^1(K)} \le C_I h_K^{-1}||v||_{L^2(K)}, \quad \forall K \in \mathcal{T}_h,\ v \in P^p(K).$$

The proof is a consequence of standard scaling arguments – see, e. g. [7], proof of Theorem 3.2.5, from which it follows that the constant C_I depends on the polynomial degree p in such a way that it is an increasing function of p. Using [32], Theorem 4.76 and standard scaling arguments, we can find that $C_I = C_I^* p^2$, where C_I^* is a constant independent of v, h, K and p.

Now, for $v \in L^2(\Omega)$ we denote by $\Pi_h v$ the $L^2(\Omega)$-projection of v on S_h:

$$\Pi_h v \in S_h, \quad (\Pi_h v - v,\, \varphi_h) = 0 \qquad \forall\, \varphi_h \in S_h. \tag{2.4}$$

Obviously, if $K \in \mathcal{T}_h$, then the function $\Pi_h v|_K$ is the $L^2(K)$-projection of $v|_K$ on $P^p(K)$. Let $k \in [1, p]$ be an integer. It is possible to show ([22, Lemma 4.1]) that

Lemma 2.1.3 *There exists a constant $C > 0$ independent of h, K such that*

$$||\Pi_h v - v||_{L^2(K)} \leq C h_K^{k+1} |v|_{H^{k+1}(K)}, \tag{2.5}$$

$$|\Pi_h v - v|_{H^1(K)} \leq C h_K^{k} |v|_{H^{k+1}(K)}, \tag{2.6}$$

$$|\Pi_h v - v|_{H^2(K)} \leq C h_K^{k-1} |v|_{H^{k+1}(K)}, \tag{2.7}$$

for all $v \in H^{k+1}(K)$, $K \in T_h$ and $h \in (0, h_0)$, where $k \in [1, p]$ is an integer.

Let $[0, T]$ be a given time interval, $u \in C([0,T]; H^{p+1}(\Omega))$ such that $\frac{\partial u}{\partial t} \in L^2([0,T]; H^{p+1}(\Omega))$. We set $\eta(t) = \Pi_h u(t) - u(t) \in H^{p+1}(\Omega, \mathcal{T}_h)$ for a.a. $t \in (0, T)$.

We have the following Lemma:

Lemma 2.1.4 *There exists a constant $C > 0$ independent of h, K such that for all $h \in (0, h_0)$*

$$a)\ ||\eta||_{L^2(\Omega,\mathcal{T}_h)} \leq C h^{p+1} |u|_{H^{p+1}(\Omega)}, \tag{2.8}$$

$$b)\ |\eta|_{H^1(\Omega,\mathcal{T}_h)} \leq C h^{p} |u|_{H^{p+1}(\Omega)}, \tag{2.9}$$

$$c)\ |\eta|_{H^2(\Omega,\mathcal{T}_h)} \leq C h_K^{k-1} |v|_{H^{p+1}(\Omega)}, \tag{2.10}$$

$$d)\ \left|\left|\frac{\partial \eta}{\partial t}\right|\right| \leq C h^{p+1} \left|\frac{\partial u}{\partial t}\right|_{H^{p+1}(\Omega)}. \tag{2.11}$$

Proof: These results follow immediately from Lemma 2.1.3 and the fact that $h_K \leq h$, $\forall K \in \mathcal{T}_h$. □

Lemma 2.1.5 (Properties of the form J_h) *For all $v, w \in H^1(\Omega, \mathcal{T}_h)$ we have*

$$a)\ J_h(v,w) \leq \big(J_h(v,v)\big)^{1/2} \big(J_h(w,w)\big)^{1/2}, \tag{2.12}$$

$$b)\ J_h(\eta,\eta) \leq C h^{2p} |u|^2_{H^{p+1}(\Omega)}. \tag{2.13}$$

Proof: Case a): the Cauchy inequality gives us

$$\begin{aligned} J_h^\sigma(w,v) &= \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sigma [w]\,[v]\,\mathrm{d}S + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma\, w\, v\,\mathrm{d}S \\ &\leq \left(\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sigma [w]^2\,\mathrm{d}S + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma\, w^2\,\mathrm{d}S \right)^{1/2} \cdot \\ &\quad \times \left(\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sigma [v]^2\,\mathrm{d}S + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma\, v^2\,\mathrm{d}S \right)^{1/2} \\ &= \left(J_h^\sigma(w,w)\right)^{1/2} \left(J_h^\sigma(v,v)\right)^{1/2}. \end{aligned}$$

Case b):

$$J_h(\eta,\eta) = \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \frac{C_W}{d(\Gamma_{ij})}[\eta]^2 \,\mathrm{d}S + \sum_{i\in I}\sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \frac{C_W}{d(\Gamma_{ij})}\,\eta^2 \,\mathrm{d}S$$
$$\leq C\sum_{i\in I}\frac{1}{h_{K_i}}\int_{\partial K_i}\eta^2\,\mathrm{d}S = C\sum_{i\in I}\frac{1}{h_{K_i}}||\eta||^2_{L^2(\partial K_i)}$$

Now the multiplicative trace inequality and Lemma 2.1.3 give us

$$J_h(\eta,\eta) \leq C\sum_{i\in I}\frac{1}{h_{K_i}}\big(||\eta||_{L^2(K_i)}|\eta|_{H^1(K_i)} + h_{K_i}^{-1}||\eta||^2_{L^2(K_i)}\big)$$
$$\leq C\,h^{2p}|u|^2_{H^{p+1}(\Omega)}$$

□

Lemma 2.1.6 *There exist constants $C>0$ independent of h, K such that*

$$\textit{a)}\ \sum_{i\in I} h_{K_i}||\eta||_{L^2(\partial K_i)} \leq Ch^{2p+2}|u|^2_{H^{p+1}(\Omega)}, \tag{2.14}$$

$$\textit{b)}\ \sum_{i\in I} h_{K_i}||\nabla\eta||_{L^2(\partial K_i)} \leq Ch^{2p}|u|^2_{H^{p+1}(\Omega)}, \tag{2.15}$$

$$\textit{c)}\ \sum_{i\in I} h_{K_i}||\varphi_h||_{L^2(\partial K_i)} \leq C\|\varphi_h\|^2_{L^2(\Omega)},\ \forall\varphi_h\in S_h, \tag{2.16}$$

$$\textit{d)}\ \sum_{i\in I} h_{K_i}||\nabla\varphi_h||_{L^2(\partial K_i)} \leq C\|\varphi_h\|^2_{H^1(\Omega,\mathcal{T}_h)},\ \forall\varphi_h\in S_h. \tag{2.17}$$

Proof: We prove only *a)* and *c)*, since *b)* and *d)* are analogous. a) follows from Lemmas 2.1.1 and 2.1.3:

$$\sum_{i\in I} h_{K_i}||\eta||_{L^2(\partial K_i)} \leq C\sum_{i\in I} h_{K_i}\big(\|\eta\|_{L^2(K_i)}|\eta|_{H^1(K)} + h_{K_i}^{-1}\|\eta\|^2_{L^2(K_i)}\big)$$
$$\leq Ch^{2p+2}|u|^2_{H^{p+1}(\Omega)}.$$

Case *c)* follows from Lemmas 2.1.1 and 2.1.2:

$$\sum_{i\in I} h_{K_i}||\varphi_h||_{L^2(\partial K_i)} \leq C\sum_{i\in I} h_{K_i}\big(\|\varphi_h\|_{L^2(K_i)}|\varphi_h|_{H^1(K)} + h_{K_i}^{-1}\|\varphi_h\|^2_{L^2(K_i)}\big)$$
$$\leq C\sum_{i\in I} h_{K_i}\big(\|\varphi_h\|_{L^2(K_i)}h_{K_i}^{-1}\|\varphi_h\|_{L^2(K_i)} + h_{K_i}^{-1}\|\varphi_h\|^2_{L^2(K_i)}\big) \leq C\|\varphi_h\|^2_{L^2(\Omega)}.$$

□

In the error estimates of the following sections, we will apply the following version of Gronwall's lemma:

Lemma 2.1.7 (Gronwall's lemma) *Let $y, q, z, r \in C([0,T]),\ r \geq 0$ and let*

$$y(t) + q(t) \leq z(t) + \int_0^t r(s)y(s)\,\mathrm{d}s, \quad t \in [0,T].$$

Then

$$\begin{aligned} &y(t) + q(t) + \int_0^t r(\vartheta)q(\vartheta)\exp\Big(\int_\vartheta^t r(s)\,\mathrm{d}s\Big)\,\mathrm{d}\vartheta \\ &\quad \leq z(t) + \int_0^t r(\vartheta)z(\vartheta)\exp\Big(\int_\vartheta^t r(s)\,\mathrm{d}s\Big)\,\mathrm{d}\vartheta, \quad t \in [0,T]. \end{aligned} \tag{2.18}$$

Proof: Can be carried out in a similar way as in [20], Par. 8.2.29. □

2.2 Error estimates for model nonlinear convection-diffusion equation

2.2.1 Definitions

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded polygonal (if $d = 2$) or polyhedral (if $d = 3$) domain with Lipschitz-continuous boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and $T > 0$. Let us assume that the $(d-1)$-dimensional measure of Γ_D is positive. We treat the following nonlinear problem:

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{2} \frac{\partial f_s(u)}{\partial x_s} - \operatorname{div}(\beta(u)\nabla u)) = g \quad \text{in } Q_T, \tag{2.19}$$

$$u|_{\Gamma_D \times (0,T)} = u_D, \tag{2.20}$$

$$\beta(u)\frac{\partial u}{\partial n}\Big|_{\Gamma_N \times (0,T)} = g_N, \tag{2.21}$$

$$u(x, 0) = u^0(x), \quad x \in \Omega, \tag{2.22}$$

where the function β satisfies

$$\beta : I\!R \to [\beta_0, \beta_1], \quad 0 < \beta_0 < \beta_1 < \infty, \tag{2.23}$$

$$|\beta(u_1) - \beta(u_2)| \le L|u_1 - u_2|, \quad \forall u_1, u_2 \in I\!R. \tag{2.24}$$

Let $g : Q_T \to \mathbb{R}$, $u_D : \Gamma_D \times (0, T) \to \mathbb{R}$, $g_N : \Gamma_N \times (0, T) \to \mathbb{R}$ and $u^0 : \Omega \to \mathbb{R}$ be given functions, and $f_s \in C^1(\mathbb{R})$, $s = 1, \ldots, d$, be prescribed Lipschitz-continuous fluxes. Without the loss of generality let $f_s(0) = 0, s = 1, \ldots, d$. We assume that the weak solution u is sufficiently regular, namely

$$\frac{\partial u}{\partial t} \in L^2\big([0, T]; H^{p+1}(\Omega)\big),$$

where $p \ge 1$ denotes the given degree of approximation. It is possible to show that, under these conditions, u satisfies equation (2.19) pointwise and $u \in C\big([0, T]; H^{p+1}(\Omega)\big)$.

To treat the nonlinear diffusion terms, we need one more regularity assumption on the solution u of the continuous problem:

$$\|\nabla u(t)\|_{L^\infty(\Omega)} \le C_R. \quad \text{for a.a. } t \in (0, T). \tag{2.25}$$

We define the discontinuous Galerkin solution of our problem using the formulation from Chapter 1.

Using forms (1.21)-(1.28) we can introduce the discrete problem (space semidiscretization with continuous time). For simplicity of notation, we use a generic diffusion form $a_h(u, \varphi)$ and right-hand side $l_h(u, \varphi)$. In our framework, these can be either the *symmetric*, *nonsymmetric* or *incomplete* variants. In the following,

if we work with only one of the three variants we shall explicitly state this fact, otherwise for results which hold for any of these variants, we use the generic notation without the superscripts N, S or I.

The form b_h approximates convective terms with the aid of a numerical flux $H(u, v, \mathbf{n})$. We assume that H has the following properties:

Assumptions (H):

(H1) $H(u, v, \boldsymbol{n})$ is defined in $\mathbb{R}^2 \times B_1$, where $B_1 = \{\boldsymbol{n} \in \mathbb{R}^d; |\boldsymbol{n}| = 1\}$, and is *Lipschitz-continuous* with respect to u, v:

$$|H(u, v, \boldsymbol{n}) - H(u^*, v^*, \boldsymbol{n})| \le C_L(|u - u^*| + |v - v^*|),$$
$$u,\ v,\ u^*,\ v^* \in \mathbb{R},\ \boldsymbol{n} \in B_1.$$

(H2) $H(u, v, \boldsymbol{n})$ is *consistent*:

$$H(u, u, \boldsymbol{n}) = \sum_{s=1}^{d} f_s(u)\, n_s, \quad u \in \mathbb{R},\ \boldsymbol{n} = (n_1, \dots, n_d) \in B_1.$$

(H3) $H(u, v, \boldsymbol{n})$ is *conservative*:

$$H(u, v, \boldsymbol{n}) = -H(v, u, -\boldsymbol{n}), \quad u,\ v \in \mathbb{R},\ \boldsymbol{n} \in B_1.$$

In virtue of assumptions (H1) and (H2), we have $2C_L \ge L_f$, where L_f is the Lipschitz-continuity constant of the functions $f_s,\ s = 1, \dots, d$.

Due to the assumption that $f_s(0) = 0$ for $s = 1, \dots, d$, we have

$$H(0, 0, \boldsymbol{n}) = 0 \quad \forall\, \boldsymbol{n} \in B_1. \tag{2.26}$$

Definition 2.2.1 *We say that u_h is a DGFE solution of the convection-diffusion problem (2.19) - (2.22), if*

$$\begin{aligned}
&a)\ u_h \in C^1([0, T]; S_h),\\
&b)\ \frac{d}{dt}(u_h(t), \varphi_h) + b_h(u_h(t), \varphi_h) + \beta_0 J_h(u_h(t), \varphi_h) + a_h(u_h(t), \varphi_h)\\
&\qquad = l_h(u_h, \varphi_h)(t), \quad \forall \varphi_h \in S_h,\ \forall t \in (0, T),\\
&c)\ u_h(0) = u_h^0,
\end{aligned} \tag{2.27}$$

where $a_h = a_h^N,\ l_h = l_h^N$ (nonsymmetric variant) or $a_h = a_h^S,\ l_h = l_h^S$ (symmetric variant) or $a_h = a_h^I,\ l_h = l_h^I$ (incomplete variant). By u_h^0 we denote an S_h approximation of the initial condition u^0.

Due to the consistency of the numerical flux we see that the exact solution u satisfies

$$\begin{aligned}\frac{d}{dt}(u(t),\varphi_h) + b_h(u(t),\varphi_h) + \beta_0 J_h(u(t),\varphi_h) + a_h(u(t),\varphi_h) \\ = l_h(u,\varphi_h)(t), \quad \forall \varphi_h \in S_h,\ \forall t \in (0,T),\end{aligned} \tag{2.28}$$

which implies the *Galerkin orthogonality property* of the error.

2.2.2 Properties of the convective term

We use the following notation:

$$\|w\|_{DG} = \left(\frac{1}{2}\left(|w|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(w,w)\right)\right)^{1/2}. \tag{2.29}$$

Since the $(d-1)$-dimensional measure of Γ_D is positive, $\|\cdot\|_{DG}$ is a norm in $H^1(\Omega,\mathcal{T}_h)$.

Now, we shall be concerned with the properties of the form b_h. Under assumptions (H) and (A) the convective form b_h is Lipschitz continuous in the following sense:

Lemma 2.2.1 *Let $u, \hat{u}, v \in H^1(\Omega,\mathcal{T}_h)$ and $h \in (0,h_0)$. Then there exists a constant $C>0$ independent of $u,\hat{u},v,h$ such that*

$$\begin{aligned}|b_h(u,v) - b_h(\bar{u},v)| \le C\left(J_h^\sigma(v,v)^{1/2} + |v|_{H^1(\Omega,\mathcal{T}_h)}\right) \\ \times\left(\|u-\bar{u}\|_{L^2(\Omega)} + \Big(\sum_{i\in I} h_{K_i}\|u-\bar{u}\|^2_{L(\partial K_i)}\Big)^{1/2}\right).\end{aligned} \tag{2.30}$$

Proof: By the definitions of b_h (1.28) and the state $u|_{\Gamma_{ij}}$ (1.20), we have for $u,\ \bar{u},\ v \in H^1(\Omega,\mathcal{T}_h)$,

$$\begin{aligned}b_h(u,v) - b_h(\bar{u},v) \ = \ & \underbrace{-\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}\left(f_s(u)-f_s(\bar{u})\right)\frac{\partial v}{\partial x_s}\,\mathrm{d}x}_{:=\sigma_1} \\ & + \underbrace{\sum_{i\in I}\sum_{j\in s(i)}\int_{\Gamma_{ij}}\left(H(u|_{\Gamma_{ij}},u|_{\Gamma_{ji}},\boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}},\bar{u}|_{\Gamma_{ji}},\boldsymbol{n}_{ij})\right)v|_{\Gamma_{ij}}\,\mathrm{d}S}_{:=\sigma_2} \\ & + \underbrace{\sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}}\left(H(u|_{\Gamma_{ij}},u|_{\Gamma_{ij}},\boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}},\bar{u}|_{\Gamma_{ij}},\boldsymbol{n}_{ij})\right)v|_{\Gamma_{ij}}\,\mathrm{d}S}_{:=\sigma_3} \\ & + \underbrace{\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\left(H(u|_{\Gamma_{ij}},u_D,\boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}},u_D,\boldsymbol{n}_{ij})\right)v|_{\Gamma_{ij}}\,\mathrm{d}S}_{:=\sigma_4}.\end{aligned}$$

From the Lipschitz-continuity of the functions $f_s,\ s=1,\dots,d$, we have

$$|\sigma_1| \le L_f \sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} |u-\bar u| \left|\frac{\partial v}{\partial x_s}\right| \,\mathrm{d}x \le \sqrt{d} L_f \|u-\bar u\|_{L^2(\Omega)} |v|_{H^1(\Omega,\mathcal{T}_h)}. \tag{2.31}$$

Using the conservativity (2.26) of H we find that

$$\sigma_2 = \frac{1}{2}\sum_{i\in I}\sum_{j\in s(i)} \int_{\Gamma_{ij}} \left(v|_{\Gamma_{ij}} - v|_{\Gamma_{ji}}\right)\left(H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) - H(\bar u|_{\Gamma_{ij}}, \bar u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij})\right) \mathrm{d}S.$$

This, the Lipschitz-continuity (2.26) of H and the Cauchy inequality imply that

$$\begin{aligned}
|\sigma_2+\sigma_3+\sigma_4| &\le C \sum_{i\in I} \left\{ \frac{1}{2}\sum_{j\in s(i)} \int_{\Gamma_{ij}} \left|[v]_{\Gamma_{ij}}\right| \left(|u-\bar u|_{\Gamma_{ij}} + |u-\bar u|_{\Gamma_{ji}}\right) \mathrm{d}S \right. \\
&\quad \left. + 2\sum_{j\in\gamma_N(i)} \int_{\Gamma_{ij}} |v|_{\Gamma_{ij}}|\,|u-\bar u|_{\Gamma_{ij}}\,\mathrm{d}S + \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} |v|_{\Gamma_{ij}}|\,|u-\bar u|_{\Gamma_{ij}}\,\mathrm{d}S \right\} \\
&\le C \sum_{i\in I} \left\{ \sum_{j\in s(i)} \int_{\Gamma_{ij}} \left|[v]_{\Gamma_{ij}}\right| |u-\bar u|_{\Gamma_{ij}}\,\mathrm{d}S + 2\sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} |v|_{\Gamma_{ij}}|\,|u-\bar u|_{\Gamma_{ij}}\,\mathrm{d}S \right\} \\
&\le C \sum_{i\in I} \left\{ \left(\sum_{j\in s(i)} \int_{\Gamma_{ij}} \frac{C_W}{d(\Gamma_{ij})} [v]^2\,\mathrm{d}S + \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \frac{C_W}{d(\Gamma_{ij})} v^2\,\mathrm{d}S \right)^{1/2} \right. \\
&\quad \left. \times \left(\sum_{j\in s(i)} \frac{d(\Gamma_{ij})}{C_W} \int_{\Gamma_{ij}} (u-\bar u)^2\,\mathrm{d}S + \sum_{j\in\gamma(i)} \frac{d(\Gamma_{ij})}{C_W} \int_{\Gamma_{ij}} (u-\bar u)^2\,\mathrm{d}S \right)^{1/2} \right\} \\
&\le C\, J_h^\sigma(v,v)^{1/2} \left(\sum_{i\in I} h_{K_i} \|u-\bar u\|^2_{L^2(\partial K_i)} \right)^{1/2}.
\end{aligned}$$

From this estimate and (2.31) we get (2.30). □

Lemma 2.2.2 *Let u be the solution of the continuous problem (2.19), u_h the solution of the discrete problem (2.27), $\Pi_h u$ be defined by (2.4), and $\xi\ (=\xi_h) = u_h - \Pi_h u \in S_h$. Then there exists a constant $C>0$, independent of $h\in(0,h_0)$, such that*

$$|b_h(u,\xi) - b_h(u_h,\xi)| \le C\|\xi\|_{DG} \left(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\right). \tag{2.32}$$

Proof: We can write

$$\begin{aligned}
&|b_h(u,\xi) - b_h(u_h,\xi)| \\
&\qquad \le |b_h(u,\xi) - b_h(\Pi_h u,\xi)| + |b_h(\Pi_h u,\xi) - b_h(u_h,\xi)|\,.
\end{aligned} \tag{2.33}$$

According to (2.30), Lemma 2.1.4, a) and Lemma 2.1.6, a),

$$\begin{aligned}
&|b_h(u,\xi) - b_h(\Pi_h u,\xi)| \\
&\quad \le C\|\xi\|_{DG}\left(\|u-\Pi_h u\|_{L^2(\Omega)} + \left(\sum_{i\in I} h_{K_i}\|u-\Pi_h u\|^2_{L_2(\partial K_i)}\right)^{1/2}\right) \qquad (2.34)\\
&\quad \le C\|\xi\|_{DG}\, h^{p+1}|u|_{H^{p+1}(\Omega)}.
\end{aligned}$$

Finally, it remains to estimate the second term in (2.33). We use Lemma 2.30 and Lemma 2.1.6, c), which yields

$$\begin{aligned}
&|b_h(\Pi_h u,\xi) - b_h(u_h,\xi)| \\
&\quad \le C\|\xi\|_{DG}\left(\|\xi\|_{L^2(\Omega)} + \left(\sum_{i\in I} h_{K_i}\|\xi\|^2_{L_2(\partial K_i)}\right)^{1/2}\right) \qquad (2.35)\\
&\quad \le C\|\xi\|_{DG}\,\|\xi\|_{L^2(\Omega)}.
\end{aligned}$$

Now, the combination of (2.33), (2.34) and (2.35) gives the desired estimate (2.32), which we wanted to prove. □

2.2.3 Error estimates

Let us remind that in this section we denote by C a generic constant independent of $h, \beta_0, \beta_1, L, C_R$. Let u be the exact solution of the continuous problem and u_h the solution of the discrete problem. We set

$$\xi(t) = u_h(t) - \Pi_h u(t) \in S_h, \quad \eta(t) = \Pi_h u(t) - u(t) \in H^{p+1}(\Omega, \mathcal{T}_h).$$

Then

$$e_h(t) := u_h(t) - u(t) = \xi(t) + \eta(t).$$

We subtract (2.28) from (2.27), set $\varphi_h := \xi$ and use Lemmas 2.1.4, 2.2.2, properties of the form J_h and the relation

$$\left(\frac{\partial \xi(t)}{\partial t}, \xi(t)\right) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}||\xi(t)||^2_{L^2(\Omega)}. \qquad (2.36)$$

Then we get:

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}||\xi(t)||^2_{L^2(\Omega)} + a_h(u_h,\xi) - a_h(u,\xi) - l_h(u_h,\xi) + l_h(u,\xi) + \beta_0 J_h(\xi,\xi)\\
&= b_h(u,\xi) - b_h(u_h,\xi) - \left(\frac{\partial \eta(t)}{\partial t}, \xi(t)\right) - \beta_0 J_h(\eta,\xi)\\
&\le C\Big\{\big(J_h(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\big)\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + ||\xi||_{L^2(\Omega)}\big) \qquad (2.37)\\
&\quad + h^{p+1}|\partial u/\partial t|_{H^p(\Omega)}||\xi||_{L^2(\Omega)} + \beta_0 h^p |u|_{H^{p+1}(\Omega)} J_h(\xi,\xi)^{1/2}\Big\}.
\end{aligned}$$

For the treatment of the left-hand side diffusion forms a_h and l_h in (2.37) we need the following results of Lemmas 2.2.3, 2.2.4 and 2.2.5, which treat individual variants of the diffusion and right-hand side forms. In Corollary 2.2.6, we unify these results using the generic notation for simplicity.

Lemma 2.2.3 (Nonsymmetric case) *Let the constant from (1.18) satisfy $C_W > 0$. For the nonsymmetric diffusion form $a_h = a_h^N$ and nonsymmetric right hand side $l_h = l_h^N$ we have*

$$a_h^N(u_h,\xi) - a_h^N(u,\xi) - l_h^N(u_h,\xi) + l_h^N(u,\xi) = A + B, \tag{2.38}$$

where

$$\begin{aligned} &A \geq \beta_0 |\xi|^2_{H^1(\Omega,\mathcal{T}_h)}, \\ &|B| \leq C\Bigg((2\beta_1 - \beta_0) h^p |u|_{H^{p+1}(\Omega)} + \\ &\quad + LC_R\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big)\Bigg)\big(|\xi|_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)^{1/2}\big) \end{aligned} \tag{2.39}$$

Proof: We break down $a_h^N(u,\xi) - l_h^N(u,\xi)$ into individual terms by setting

$$\begin{aligned} \sigma^1(u,\xi) &= \sum_{i\in I} \int_{K_i} \beta(u)\nabla u \cdot \nabla \xi \, dx, \\ \sigma^2(u,\xi) &= -\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \beta(u)\nabla u \rangle \cdot \mathbf{n}_{ij}[\xi] \, dS, \\ \sigma^3(u,\xi) &= \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \beta(u)\nabla \xi \rangle \cdot \mathbf{n}_{ij}[u] \, dS, \\ \sigma^4(u,\xi) &= -\sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \beta(u)\nabla u \cdot \mathbf{n}_{ij}\xi \, dS, \\ \sigma^5(u,\xi) &= \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \beta(u)\nabla \xi \cdot \mathbf{n}_{ij} u \, dS - \\ &\quad - \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \beta(u)\nabla \xi \cdot \mathbf{n}_{ij} u_D \, dS. \end{aligned} \tag{2.40}$$

Therefore

$$a_h^N(u_h,\xi) - a_h^N(u,\xi) - l_h^N(u_h,\xi) + l_h^N(u,\xi) = \sum_{i=1}^{5} (\sigma^i(u_h,\xi) - \sigma^i(u,\xi)) \tag{2.41}$$

and we shall treat these terms separately:

1) First term:

$$\begin{aligned}\sigma^1(u_h,\xi)-\sigma^1(u,\xi)&=\sum_{i\in I}\int_{K_i}\big(\beta(u_h)\nabla u_h-\beta(u)\nabla u\big)\cdot\nabla\xi\,dx\\&=\sum_{i\in I}\int_{K_i}\Big(\big(\beta(u_h)\nabla u_h-\beta(u_h)\nabla\Pi_h u\big)+\big(\beta(u_h)\nabla\Pi_h u-\beta(u)\nabla\Pi_h u\big)\\&\quad+\big(\beta(u)\nabla\Pi_h u-\beta(u)\nabla u\big)\Big)\cdot\nabla\xi\,dx=\sigma_1^1+\sigma_2^1+\sigma_3^1,\end{aligned}\tag{2.42}$$

and we estimate using (2.23)

$$\sigma_1^1=\sum_{i\in I}\int_{K_i}\beta(u_h)\nabla\xi\cdot\nabla\xi\,dx\geq\beta_0|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}.\tag{2.43}$$

Further, using (2.23), (2.24), the Cauchy inequality, (2.25) and Lemma 2.1.4, we get

$$\begin{aligned}|\sigma_2^1|&=\Big|\sum_{i\in I}\int_{K_i}\big(\beta(u_h)-\beta(u)\big)\nabla\Pi_h u\cdot\nabla\xi\,dx\Big|\\&=\Big|\sum_{i\in I}\int_{K_i}\Big(\big(\beta(u_h)-\beta(u)\big)\nabla\eta+\big(\beta(u_h)-\beta(u)\big)\nabla u\Big)\cdot\nabla\xi\,dx\Big|\\&\leq\sum_{i\in I}\int_{K_i}\Big((\beta_1-\beta_0)|\nabla\eta|+L|u_h-u||\nabla u|\Big)|\nabla\xi|\,dx\\&\leq(\beta_1-\beta_0)|\eta|_{H^1(\Omega,\mathcal{T}_h)}|\xi|_{H^1(\Omega,\mathcal{T}_h)}+L\sum_{i\in I}\Big(\|\nabla u\|_{L^\infty(K_i)}\int_{K_i}|u_h-u||\nabla\xi|\,dx\Big)\\&\leq(\beta_1-\beta_0)|\eta|_{H^1(\Omega,\mathcal{T}_h)}|\xi|_{H^1(\Omega,\mathcal{T}_h)}+L\|\nabla u\|_{L^\infty(\Omega)}\sum_{i\in I}\int_{K_i}|\eta||\nabla\xi|+|\xi||\nabla\xi|\,dx\\&\leq\Big(\big((\beta_1-\beta_0)+LC_Rh\big)Ch^p|u|_{H^{p+1}(\Omega)}+LC_R\|\xi\|_{L^2(\Omega)}\Big)|\xi|_{H^1(\Omega,\mathcal{T}_h)}.\end{aligned}\tag{2.44}$$

Finally, using (2.23), the Cauchy inequality and Lemma 2.1.4,

$$\begin{aligned}|\sigma_3^1|&=\Big|\sum_{i\in I}\int_{K_i}\beta(u)\big(\nabla\Pi_h u-\nabla u\big)\cdot\nabla\xi\,dx\Big|\\&\leq\sum_{i\in I}\int_{K_i}\beta_1|\nabla\eta||\nabla\xi|\,dx\leq\beta_1Ch^p|u|_{H^{p+1}(\Omega)}|\xi|_{H^1(\Omega,\mathcal{T}_h)}.\end{aligned}\tag{2.45}$$

2) Second term:

$$\sigma^2(u_h,\xi)-\sigma^2(u,\xi)=-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\big\langle\beta(u_h)\nabla u_h-\beta(u)\nabla u\big\rangle\cdot\mathbf{n}_{ij}[\xi]\,dS$$

$$=-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\big\langle\big(\beta(u_h)\nabla u_h-\beta(u_h)\nabla\Pi_h u\big)+\big(\beta(u_h)\nabla\Pi_h u-\beta(u)\nabla\Pi_h u\big)$$
$$+\big(\beta(u)\nabla\Pi_h u-\beta(u)\nabla u\big)\big\rangle\cdot\mathbf{n}_{ij}[\xi]\,dS=\sigma_1^2+\sigma_2^2+\sigma_3^2, \tag{2.46}$$

where

$$\sigma_1^2=-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\beta(u_h)\nabla\xi\rangle\cdot\mathbf{n}_{ij}[\xi]\,dS, \tag{2.47}$$

we do not estimate σ_1^2, since it will cancel out a similar term in the following. After applying (2.23) and (2.24), the estimates follow:

$$|\sigma_2^2|=\left|\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\left\langle\big(\beta(u_h)-\beta(u)\big)\nabla\Pi_h u\right\rangle\cdot\mathbf{n}_{ij}[\xi]\,dS\right|$$

$$=\left|\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\left\langle\big(\beta(u_h)-\beta(u)\big)\nabla\eta+\big(\beta(u_h)-\beta(u)\big)\nabla u\right\rangle\cdot\mathbf{n}_{ij}[\xi]\,dS\right| \tag{2.48}$$

$$\le\sum_{i\in I}\sum_{j\in s(i)}\int_{\Gamma_{ij}}(\beta_1-\beta_0)|\nabla\eta||[\xi]|+L|u_h-u||\nabla u||[\xi]|\,dS=:RHS.$$

Next, we apply the Cauchy inequality and (2.25):

$$RHS\le$$
$$\le(\beta_1-\beta_0)\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{d(\Gamma_{ij})}{C_W}||\nabla\eta||^2_{L^2(\Gamma_{ij})}\right)^{1/2}\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{C_W}{d(\Gamma_{ij})}|||[\xi]|||^2_{L^2(\Gamma_{ij})}\right)^{1/2}$$
$$+L\|\nabla u\|_{L^\infty(\Omega)}\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{d(\Gamma_{ij})}{C_W}||u_h-u||^2_{L^2(\Gamma_{ij})}\right)^{1/2}\times$$
$$\times\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{C_W}{d(\Gamma_{ij})}|||[\xi]|||^2_{L^2(\Gamma_{ij})}\right)^{1/2}$$
$$\le\frac{(\beta_1-\beta_0)}{C_W}\left(\sum_{i\in I}h_{K_i}||\nabla\eta||^2_{L^2(\partial K_i)}\right)^{1/2}J_h(\xi,\xi)^{1/2}$$
$$+\frac{L\|\nabla u\|_{L^\infty(\Omega)}}{C_W}\left(\sum_{i\in I}h_{K_i}||u_h-u||^2_{L^2(\partial K_i)}\right)^{1/2}J_h(\xi,\xi)^{1/2}. \tag{2.49}$$

Now we apply Lemma 2.1.6 a), c) and conclude the estimate of $|\sigma_2^2|$:

$$\begin{aligned}
|\sigma_2^2| &\le \frac{(\beta_1-\beta_0)}{C_W} C h^p |u|_{H^{p+1}(\Omega)} J_h(\xi,\xi)^{1/2} \\
&+ \frac{LC_R}{C_W}\left\{ \left(\sum_{i\in I} h_{K_i} ||\eta||^2_{L^2(\partial K_i)} \right)^{1/2} + \left(\sum_{i\in I} h_{K_i} ||\xi||^2_{L^2(\partial K_i)} \right)^{1/2} \right\} J_h(\xi,\xi)^{1/2} \qquad (2.50)\\
&\le C\left(\big((\beta_1-\beta_0) + LC_R h\big) h^p |u|_{H^{p+1}(\Omega)} + LC_R ||\xi||_{L^2(\Omega)} \right) J_h(\xi,\xi)^{1/2}.
\end{aligned}$$

The last term is estimated using (2.23), the Cauchy inequality and Lemma 2.1.6 b):

$$\begin{aligned}
|\sigma_3^2| &= \left| \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \big\langle \beta(u)\nabla\eta \big\rangle \cdot \mathbf{n}_{ij} [\xi]\, dS \right| \\
&\le \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} \beta_1 |\nabla\eta| |[\xi]|\, dS \\
&\le \beta_1 \left(\sum_{i\in I} \sum_{j\in s(i)} \frac{d(\Gamma_{ij})}{C_W} ||\nabla\eta||^2_{L^2(\Gamma_{ij})} \right)^{1/2} \left(\sum_{i\in I} \sum_{j\in s(i)} \frac{C_W}{d(\Gamma_{ij})} ||[\xi]||^2_{L^2(\Gamma_{ij})} \right)^{1/2} \qquad (2.51)\\
&\le \frac{\beta_1}{C_W} \left(\sum_{i\in I} h_{K_i} ||\nabla\eta||^2_{L^2(\partial K_i)} \right)^{1/2} J_h(\xi,\xi)^{1/2} \\
&\le \beta_1 C h^p |u|_{H^{p+1}(\Omega)} J_h(\xi,\xi)^{1/2}
\end{aligned}$$

3) Third term:

$$\begin{aligned}
&\sigma^3(u_h,\xi) - \sigma^3(u,\xi) \\
&= \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \beta(u_h)\nabla\xi \rangle \cdot \mathbf{n}_{ij} [u_h] - \langle \beta(u)\nabla\xi \rangle \cdot \mathbf{n}_{ij} [u]\, dS \\
&= \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \big\langle \beta(u_h)\nabla\xi \big\rangle \cdot \mathbf{n}_{ij} [u_h - \Pi_h u] \qquad (2.52)\\
&\quad + \big\langle \big(\beta(u_h) - \beta(u)\big)\nabla\xi \big\rangle \cdot \mathbf{n}_{ij} [\Pi_h u] + \big\langle \beta(u)\nabla\xi \big\rangle \cdot \mathbf{n}_{ij} [\Pi_h u - u]\, dS \\
&= \sigma_1^3 + \sigma_2^3 + \sigma_3^3,
\end{aligned}$$

Using (2.47) we get

$$\sigma_1^3 = \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \beta(u_h)\nabla\xi \rangle \cdot \mathbf{n}_{ij} [\xi]\, dS = -\sigma_1^2. \qquad (2.53)$$

Due to the regularity condition (2.25), the function u is continuous and, thus, $[u] = 0$ and $[\Pi_h u] = [\eta]$. We get the estimate:

$$\begin{aligned}
|\sigma_2^3| &= \left| \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \Big\langle \big(\beta(u_h) - \beta(u)\big)\nabla\xi \Big\rangle \cdot \mathbf{n}_{ij}[\Pi_h u]\, dS \right| \\
&= \left| \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \Big\langle \big(\beta(u_h) - \beta(u)\big)\nabla\xi \Big\rangle \cdot \mathbf{n}_{ij}[\eta]\, dS \right| \\
&\le \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} (\beta_1 - \beta_0)|\nabla\xi||[\eta]|\, dS \\
&\le (\beta_1 - \beta_0)\Big(\sum_{i\in I} \sum_{j\in s(i)} \frac{d(\Gamma_{ij})}{C_W} ||\nabla\xi||^2_{L^2(\Gamma_{ij})} \Big)^{1/2} \Big(\sum_{i\in I} \sum_{j\in s(i)} \frac{C_W}{d(\Gamma_{ij})} |||[\eta]|||^2_{L^2(\Gamma_{ij})} \Big)^{1/2} \\
&\le \frac{(\beta_1 - \beta_0)}{C_W} \Big(\sum_{i\in I} h_{K_i} ||\nabla\xi||^2_{L^2(\partial K_i)} \Big)^{1/2} J_h(\eta,\eta)^{1/2} \\
&\le (\beta_1 - \beta_0) C |\xi|_{H^1(\Omega,\mathcal{T}_h)} h^p |u|_{H^{p+1}(\Omega)},
\end{aligned} \tag{2.54}$$

where relation (2.23), the Cauchy inequality and Lemma 2.1.6 d) were applied. Finally we have

$$\begin{aligned}
|\sigma_3^3| &= \left| \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \big\langle \beta(u)\nabla\xi \big\rangle \cdot \mathbf{n}_{ij}[\eta]\, dS \right| \\
&\le \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} \beta_1 |\nabla\xi||[\eta]|\, dS \\
&\le \beta_1 \Big(\sum_{i\in I} \sum_{j\in s(i)} \frac{d(\Gamma_{ij})}{C_W} ||\nabla\xi||^2_{L^2(\Gamma_{ij})} \Big)^{1/2} \Big(\sum_{i\in I} \sum_{j\in s(i)} \frac{C_W}{d(\Gamma_{ij})} |||[\eta]|||^2_{L^2(\Gamma_{ij})} \Big)^{1/2} \\
&\le \beta_1 C |\xi|_{H^1(\Omega,\mathcal{T}_h)} h^p |u|_{H^{p+1}(\Omega)}.
\end{aligned} \tag{2.55}$$

4) Fourth term:

$$\begin{aligned}
\sigma^4(u_h,\xi) - \sigma^4(u,\xi) &= -\sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \big(\beta(u_h)\nabla u_h - \beta(u)\nabla u\big)\cdot \mathbf{n}_{ij}\xi\, dS \\
&= -\sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \big(\beta(u_h)\nabla\xi + \big(\beta(u_h) - \beta(u)\big)\nabla\Pi_h u + \beta(u)\nabla\eta\big)\cdot \mathbf{n}_{ij}\xi\, dS \\
&= \sigma_1^4 + \sigma_2^4 + \sigma_3^4
\end{aligned} \tag{2.56}$$

and these terms can be treated similarly as $\sigma_1^2,\ \sigma_2^2$ and σ_3^2 to obtain

$$\begin{aligned}
&\sigma_1^4 = -\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u_h)\nabla\xi\cdot\mathbf{n}_{ij}\xi\,dS,\\
&|\sigma_2^4| \le C\bigg(\big((\beta_1-\beta_0)+LC_Rh\big)h^p|u|_{H^{p+1}(\Omega)}+LC_R||\xi||_{L^2(\Omega)}\bigg)J_h(\xi,\xi)^{1/2},\\
&|\sigma_3^4| \le \beta_1 Ch^p|u|_{H^{p+1}(\Omega)}J_h(\xi,\xi)^{1/2}.
\end{aligned} \tag{2.57}$$

5) Fifth term:

$$\begin{aligned}
\sigma^5(u_h,\xi)-\sigma^5(u,\xi) &= \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u_h)\nabla\xi\cdot\mathbf{n}_{ij}u_h\\
&\quad-\beta(u)\nabla\xi\cdot\mathbf{n}_{ij}u-\big(\beta(u_h)-\beta(u)\big)\nabla\xi\cdot\mathbf{n}_{ij}u_D\,dS\\
&=\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\beta(u_h)\nabla\xi\cdot\mathbf{n}_{ij}\xi\\
&\quad+\big(\beta(u_h)-\beta(u)\big)\nabla\xi\cdot\mathbf{n}_{ij}\big(\Pi_hu-u_D\big)+\beta(u)\nabla\xi\cdot\mathbf{n}_{ij}\eta\,dS\\
&=\sigma_1^5+\sigma_2^5+\sigma_3^5
\end{aligned} \tag{2.58}$$

and it follows, due to (2.57) that

$$\sigma_1^5 = -\sigma_1^4, \tag{2.59}$$

Further

$$\begin{aligned}
\sigma_2^5 &= \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\big(\beta(u_h)-\beta(u)\big)\nabla\xi\cdot\mathbf{n}_{ij}\big(\Pi_hu-u_D\big)\,dS\\
&=\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\big(\beta(u_h)-\beta(u)\big)\nabla\xi\cdot\mathbf{n}_{ij}\eta\\
&\quad+\big(\beta(u_h)-\beta(u)\big)\nabla\xi\cdot\mathbf{n}_{ij}(u-u_D)\,dS\\
&=\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\big(\beta(u_h)-\beta(u)\big)\nabla\xi\cdot\mathbf{n}_{ij}\eta\,dS,
\end{aligned} \tag{2.60}$$

since $u=u_D$ on Γ_D. This, the Cauchy inequality and Lemma 2.1.6 d) give us

$$\begin{aligned}
|\sigma_2^5| &\le \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}(\beta_1-\beta_0)|\nabla\xi||\eta|\,dS\\
&\le(\beta_1-\beta_0)\bigg(\sum_{i\in I}\frac{h_{K_i}}{C_W}||\nabla\xi||^2_{L^2(\partial K_i)}\bigg)^{1/2}J_h(\eta,\eta)^{1/2}\\
&\le(\beta_1-\beta_0)C|\xi|_{H^1(\Omega,\mathcal{T}_h)}h^p|u|_{H^{p+1}(\Omega)}.
\end{aligned} \tag{2.61}$$

Similarly

$$\begin{aligned} |\sigma_3^5| &= \left| \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \beta(u)\nabla\xi\cdot\mathbf{n}_{ij}\eta\, dS \right| \\ &\le \beta_1 \left(\sum_{i\in I} \frac{h_{K_i}}{C_W} ||\nabla\xi||^2_{L^2(\partial K_i)} \right)^{1/2} J_h(\eta,\eta)^{1/2} \\ &\le \beta_1 C |\xi|_{H^1(\Omega,\mathcal{T}_h)} h^p |u|_{H^{p+1}(\Omega)}. \end{aligned} \tag{2.62}$$

Finally the proof is completed by taking

$$A = \sigma_1^1$$

using the appropriate cancellations (2.53), (2.59), which imply that

$$B = \sum_{i=1}^{5}\sum_{j=1}^{3} \sigma_j^i - \sigma_1^1 = \sum_{i=1}^{5}\sum_{j=2}^{3} \sigma_j^i.$$

Applying the derived inequalities to individual terms give us the sought estimates (2.39). □

We can derive similar results for the symmetric forms, however we must impose a condition on the constant C_W from (2.92).

Lemma 2.2.4 (Symmetric case) *Let*

$$C_W \ge 4\left(\frac{\beta_1}{\beta_0}\right)^2 C_M(1+C_I), \tag{2.63}$$

where C_M and C_I are constants from Lemmas 2.1.1 and 2.1.2, respectively, then for the symmetric diffusion form $a_h = a_h^S$ and symmetric right hand side $l_h = l_h^S$ we have

$$a_h^S(u_h,\xi) - a_h^S(u,\xi) - l_h^S(u_h,\xi) + l_h^S(u,\xi) = A + B, \tag{2.64}$$

where

$$\begin{aligned} A &\ge \frac{\beta_0}{2}\big(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} - J_h(\xi,\xi)\big), \\ |B| &\le C\Big((2\beta_1-\beta_0)h^p|u|_{H^{p+1}(\Omega)} + \\ &\quad + LC_R\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big)\Big)\big(|\xi|_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)^{1/2}\big). \end{aligned} \tag{2.65}$$

Proof: Using the notation from Lemma 2.2.3 we see that

$$\begin{aligned} a_h^S(u_h,\xi) - a_h^S(u,\xi) - l_h^S(u_h,\xi) + l_h^S(u,\xi) = \\ = \sigma^1(u_h,\xi) - \sigma^1(u,\xi) + \sum_{i=2}^{5}(-1)^i\big(\sigma^i(u_h,\xi) - \sigma^i(u,\xi)\big). \end{aligned} \tag{2.66}$$

Using (2.53), (2.59), we can take

$$\begin{aligned} A &= \sigma_1^1 + \sum_{i=2}^{5}(-1)^i\sigma_1^i = \sigma_1^1 + 2\sigma_1^2 + 2\sigma_1^4, \\ B &= \sum_{i=1}^{5}\sum_{j=2}^{3}\sigma_j^i. \end{aligned} \tag{2.67}$$

The estimate for $|B|$ is the same as in Lemma 2.2.3. However, the estimation of A is more difficult, since terms that have cancelled out in the nonsymmetric form are present in this case.

Using the estimate (2.43), the Cauchy and Young inequality, for any $\delta > 0$ we get

$$\begin{aligned} A \geq \sigma_1^1 - 2|\sigma_1^2| - 2|\sigma_1^4| \geq \beta_0|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} \\ - 2\beta_1\Big(\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\frac{d(\Gamma_{ij})}{\delta}|\langle\nabla\xi\rangle|^2\,dS\Big)^{1/2}\Big(\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\frac{\delta}{d(\Gamma_{ij})}[\xi]^2\,dS\Big)^{1/2} \\ - 2\beta_1\Big(\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\frac{d(\Gamma_{ij})}{\delta}|\nabla\xi|^2\,dS\Big)^{1/2}\Big(\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\frac{\delta}{d(\Gamma_{ij})}|\xi|^2\,dS\Big)^{1/2}. \end{aligned} \tag{2.68}$$

Now using the fact that $\forall\alpha,\beta,\gamma,\delta\in\mathbb{R}$, $2(\alpha\gamma+\beta\delta)\leq\alpha^2+\beta^2+\gamma^2+\delta^2$ applied to the previous, we get

$$A \geq \beta_0|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} - \beta_1\omega - \beta_1\frac{\delta}{C_W}J_h(\xi,\xi), \tag{2.69}$$

where

$$\omega = \frac{1}{\delta}\sum_{i\in I}\Big(\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\frac{d(\Gamma_{ij})}{\delta}|\langle\nabla\xi\rangle|^2\,dS + \sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\frac{d(\Gamma_{ij})}{\delta}|\nabla\xi|^2\,dS\Big). \tag{2.70}$$

Further using Lemma 2.1.1 and 2.1.2, we get

$$\begin{aligned}\omega &\le \frac{1}{\delta}\sum_{i\in I} h_{K_i}\int_{\partial K_i}|\nabla\xi|^2\,dS\\ &\le \frac{C_M}{\delta}\sum_{i\in I} h_{K_i}\big(|\xi|_{H^1(K_i)}|\nabla\xi|_{H^1(K_i)}+h_{K_i}^{-1}|\xi|^2_{H^1(K_i)}\big)\\ &\le \frac{C_M(1+C_I)}{\delta}|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}.\end{aligned}\tag{2.71}$$

Now if we choose

$$\delta=\frac{\beta_1}{\beta_0}2C_M(1+C_I),\tag{2.72}$$

and use the condition (2.63), we get

$$\begin{aligned}A &\ge \frac{\beta_0}{2}|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}-\frac{2C_M(1+C_I)\beta_1^2}{C_W\beta_0}J_h(\xi,\xi)\\ &\ge \frac{\beta_0}{2}\big(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}-J_h(\xi,\xi)\big).\end{aligned}\tag{2.73}$$

□

Lemma 2.2.5 (Incomplete case) *Let*

$$C_W\ge 2\left(\frac{\beta_1}{\beta_0}\right)^2 C_M(1+C_I),\tag{2.74}$$

where C_M and C_I are constants from Lemmas 2.1.1 and 2.1.2, respectively, then for the incomplete *diffusion form $a_h=a_h^I$ and* incomplete *right hand side $l_h=l_h^I$ we have*

$$a_h^I(u_h,\xi)-a_h^I(u,\xi)-l_h^I(u_h,\xi)+l_h^I(u,\xi)=A+B,\tag{2.75}$$

where

$$\begin{aligned}&A\ge\frac{\beta_0}{2}\big(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}-J_h(\xi,\xi)\big),\\ &|B|\le C\Bigg((2\beta_1-\beta_0)h^p|u|_{H^{p+1}(\Omega)}+\\ &\quad+LC_R\big(h^{p+1}|u|_{H^{p+1}(\Omega)}+\|\xi\|_{L^2(\Omega)}\big)\Bigg)\big(|\xi|_{H^1(\Omega,\mathcal{T}_h)}+J_h(\xi,\xi)^{1/2}\big).\end{aligned}\tag{2.76}$$

Proof: The proof is almost identical to the proof of the previous Lemma 2.2.4. We see that

$$a_h^I(u_h,\xi)-a_h^I(u,\xi)-l_h^I(u_h,\xi)+l_h^I(u,\xi)=\sum_{i\in\{1,2,4\}}\big(\sigma^i(u_h,\xi)-\sigma^i(u,\xi)\big).\tag{2.77}$$

We can take

$$A = \sigma_1^1 + \sigma_1^2 + \sigma_1^4, \quad \text{and } B = \sum_{i\in\{1,2,4\}} \sum_{j=2}^{3} \sigma_j^i. \tag{2.78}$$

The estimate for $|B|$ is the same as in Lemma 2.2.3, and the estimate of A as in Lemma 2.2.4.

Using the estimate (2.43), the Cauchy and Young's inequality, for any $\delta > 0$ we get

$$\begin{aligned} A &\geq \sigma_1^1 - |\sigma_1^2| - |\sigma_1^5| \geq \beta_0 |\xi|^2_{H^1(\Omega,\mathcal{T}_h)} \\ &- \beta_1 \Big(\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \frac{d(\Gamma_{ij})}{\delta} |\langle \nabla \xi \rangle|^2 \, dS \Big)^{1/2} \Big(\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \frac{\delta}{d(\Gamma_{ij})} [\xi]^2 \, dS \Big)^{1/2} \\ &- \beta_1 \Big(\sum_{i\in I} \sum_{j\in \gamma_D(i)} \int_{\Gamma_{ij}} \frac{d(\Gamma_{ij})}{\delta} |\nabla \xi|^2 \, dS \Big)^{1/2} \Big(\sum_{i\in I} \sum_{j\in \gamma_D(i)} \int_{\Gamma_{ij}} \frac{\delta}{d(\Gamma_{ij})} |\xi|^2 \, dS \Big)^{1/2}. \end{aligned} \tag{2.79}$$

Now using the fact that $2(AC+BD) \leq A^2+B^2+C^2+D^2$ holds for all $A, B, C, D \in \mathbb{R}$,we get

$$A \geq \beta_0 |\xi|^2_{H^1(\Omega,\mathcal{T}_h)} - \frac{\beta_1}{2}\omega - \beta_1 \frac{\delta}{2C_W} J_h(\xi,\xi), \tag{2.80}$$

where ω is defined in (2.70) and inequality (2.71) holds. Now if we choose

$$\delta = \frac{\beta_1}{\beta_0} C_M (1 + C_I), \tag{2.81}$$

and use the condition (2.74), we get

$$A \geq \frac{\beta_0}{2} \big(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} - J_h(\xi,\xi) \big). \tag{2.82}$$

□

Now we can put together the results of the previous three lemmas into one corollary, which enables us to treat all three variants of the diffusion and right-hand side forms simultaneously. We note the requirements on C_W in order to obtain 'coercivity' of the individual variants: in the *nonsymmetric* case, it suffices to take $C_W > 0$, whereas in the *incomplete* case, we need C_W to satisfy (2.74) and in the *symmetric* case, the lower bound for C_W is twice as large as for the incomplete variant.

Corollary 2.2.6 *Due to Lemma 2.2.3, Lemma 2.2.4 and 2.2.5 we have*

$$a_h(u_h,\xi) - a_h(u,\xi) - l_h(u_h,\xi) + l_h(u,\xi) + \beta_0 J_h(\xi,\xi) = A + B, \tag{2.83}$$

where

$$\begin{aligned} A &\geq \frac{\beta_0}{2}\big(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)\big), \\ |B| &\leq C\Bigg((2\beta_1 - \beta_0)h^p |u|_{H^{p+1}(\Omega)} + \\ &\quad + LC_R\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big)\Bigg)\big(|\xi|_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)^{1/2}\big). \end{aligned} \tag{2.84}$$

This estimate holds for $a_h = a_h^N$ with $C_W > 0$, further for $a_h = a_h^S$, provided the constant C_W satisfies (2.63) and for $a_h = a_h^I$, provided the constant C_W satisfies (2.74).

2.2.4 Main theorem

Theorem 2.2.7 *Let $e_h = u_h - u$ be the error of the method presented. Then it satisfies the estimate*

$$\begin{aligned} &\max_{t\in[0,T]} ||e_h(t)||^2_{L^2(\Omega)} + \frac{\beta_0}{2}\int_0^t \big(|e_h(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(e_h(\vartheta), e_h(\vartheta))\big)\, d\vartheta \\ &\qquad \leq C\, h^{2p}, \end{aligned} \tag{2.85}$$

with a constant $C > 0$ independent of h.

Proof: From (2.37) in combination with Corollary 2.2.6 it follows that

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}||\xi(t)||^2_{L^2(\Omega)} + \frac{\beta_0}{2}\big(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)\big) \\ &\leq C\Bigg\{ \big(J_h(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\big)\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + ||\xi||_{L^2(\Omega)}\big) \\ &\quad + h^{p+1}|\partial u/\partial t|_{H^{p+1}(\Omega)}||\xi||_{L^2(\Omega)} + \beta_0 h^p |u|_{H^{p+1}(\Omega)} J_h(\xi,\xi)^{1/2} \\ &\quad + \Bigg((2\beta_1 - \beta_0)h^p |u|_{H^{p+1}(\Omega)} \\ &\quad + LC_R\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big)\Bigg)\big(|\xi|_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)\big)\Bigg\} \end{aligned}$$

Applying Young's inequality gives us

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}||\xi(t)||^2_{L^2(\Omega)} + \beta_0\big(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)\big) \\ &\leq \frac{\beta_0}{2}\big(J_h(\xi,\xi) + |\xi|^2_{H^1(\Omega,\mathcal{T}_h)}\big) + C\Bigg\{ \left(1 + \frac{1 + L^2C_R^2}{\beta_0}\right)||\xi||^2_{L^2(\Omega)} \\ &\quad + \frac{1}{\beta_0}\big(h^{2p+2} + \beta_0^2 h^{2p} + (2\beta_1 - \beta_0)^2 h^{2p} + L^2C_R^2 h^{2p+2}\big)|u|^2_{H^{p+1}(\Omega)} \\ &\quad + h^{2p+2}|\partial u/\partial t|^2_{H^{p+1}(\Omega)}\Bigg\}. \end{aligned}$$

After integrating from 0 to $t \in [0,T]$ and noticing that $\xi(0) = u_h^0 - \Pi_h u^0 = 0$, we obtain

$$
\begin{aligned}
&||\xi(t)||^2_{L^2(\Omega)} + \beta_0 \int_0^t \left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi(\vartheta),\xi(\vartheta))\right) d\vartheta \\
&\leq C\Bigg\{ \left(1 + \frac{1+L^2C_R^2}{\beta_0}\right) \int_0^t ||\xi(\vartheta)||^2_{L^2(\Omega)}\, d\vartheta + \frac{1}{\beta_0} h^{2p} \int_0^t \big(h^2 + \beta_0^2 \\
&+ (2\beta_1 - \beta_0)^2 + L^2C_R^2h^2\big)|u(\vartheta)|^2_{H^{p+1}(\Omega)}\, d\vartheta + h^{2p+2} \int_0^t |\partial u/\partial t|^2_{H^p(\Omega)}\, d\vartheta \Bigg\},
\end{aligned}
$$

Now the application of Gronwall's Lemma 2.1.7, where we set

$$
\begin{aligned}
y(t) &= ||\xi(t)||^2_{L^2(\Omega)}, \\
q(t) &= \beta_0 \int_0^t \left(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi,\xi)\right) d\vartheta, \\
r(t) &= C\frac{\beta_0 + 1 + L^2C_R^2}{\beta_0}, \\
z(t) &= Ch^{2p}\Bigg(\frac{1}{\beta_0}\big(h^2 + \beta_0^2 + (2\beta_1 - \beta_0)^2 + L^2C_R^2h^2\big)\|u\|^2_{L^2(0,T;H^{p+1}(\Omega))} \\
&\quad + h^2\|\partial u/\partial t\|^2_{L^2(0,T;H^{p+1}(\Omega))}\Bigg),
\end{aligned}
$$

implies that

$$
\begin{aligned}
&||\xi(t)||^2_{L^2(\Omega)} + \beta_0 \int_0^t \left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(\xi(\vartheta),\xi(\vartheta))\right) d\vartheta \\
&\leq Ch^{2p}\frac{\beta_0 + 1 + L^2C_R^2}{\beta_0}\Bigg(\frac{1}{\beta_0}\big(h^2 + \beta_0^2 + (2\beta_1 - \beta_0)^2 + L^2C_R^2h^2\big) \\
&\quad \times \|u\|^2_{L^2(0,T;H^{p+1}(\Omega))} + h^2\|\partial u/\partial t\|^2_{L^2(0,T;H^{p+1}(\Omega))}\Bigg) \\
&\quad \times \exp\left(C\frac{\beta_0 + 1 + L^2C_R^2}{\beta_0}\, t\right), \quad t \in [0,T].
\end{aligned}
$$

Finally, since $e_h = \xi + \eta$, the above estimates and estimates from Lemma 2.1.4 yield the sought result. □

2.3 Optimal $L^\infty(L^2)$ error estimates.

This section is concerned with the analysis of the discontinuous Galerkin space semidiscretization of a nonstationary convection-diffusion problem with a linear diffusion and nonlinear convection, equipped by mixed Dirichlet-Neumann boundary conditions and an initial condition. We prove the optimal error estimate of order $O(h^{p+1})$ in the $L^\infty(0,T;L^2(\Omega))$-norm for the symmetric interior penalty *(SIPG) method, when piecewise polynomial approximation of degree p are used. We use the so-called Aubin-Nitsche technique, which involves the use of an auxiliary* dual *problem. This imposes additional requirements on the domain Ω, namely that Ω is* convex *and we have no Neumann boundary condition, i.e. $\Gamma_N = \emptyset$. Otherwise, throughout this section we assume that the assumption* (H) *on the numerical flux (page 35), and the assumption* (A) *on the finite element mesh (page 30) hold. Namely, we allow nonconforming meshes with hanging nodes, which improves the result of [17].*

2.3.1 Continuous problem

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded polygonal (if $d = 2$) or polyhedral (if $d = 3$) *convex* domain with Lipschitz-continuous boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and $T > 0$. Let us assume that the $(d-1)$-dimensional measure of Γ_D is positive. We are concerned with the following nonstationary nonlinear convection-diffusion problem:
Find $u : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

$$\begin{aligned}
&\text{a)} \quad \frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \varepsilon \Delta u + g \quad \text{in } Q_T, \\
&\text{b)} \quad u\big|_{\Gamma_D \times (0,T)} = u_D, \\
&\text{c)} \quad \varepsilon \frac{\partial u}{\partial n}\Big|_{\Gamma_N \times (0,T)} = g_N, \\
&\text{d)} \quad u(x,0) = u^0(x), \quad x \in \Omega.
\end{aligned} \tag{2.86}$$

The diffusion coefficient $\varepsilon > 0$ is a given constant, $g : Q_T \to \mathbb{R}$, $u_D : \Gamma_D \times (0,T) \to \mathbb{R}$, $g_N : \Gamma_N \times (0,T) \to \mathbb{R}$ and $u^0 : \Omega \to \mathbb{R}$ are given functions, and $f_s \in C^1(\mathbb{R})$, $s = 1, \ldots, d$, are prescribed Lipschitz-continuous fluxes. Without the loss of generality we assume that $f_s(0) = 0, s = 1, \ldots, d$.

In what follows, we shall assume that problem (2.86) has a unique sufficiently regular solution u such that

$$u_t = \frac{\partial u}{\partial t} \in L^2(0,\, T;\, H^{p+1}(\Omega)). \tag{2.87}$$

Hence, $u \in C([0,T];\, H^{p+1}(\Omega))$.

DGFE formulation

In the following we use the notation of Chapter 1. In order to introduce the space semidiscretization of problem (2.86) over the mesh $\mathcal{T}_h$ by the DGFEM, we use the following forms defined in Chapter 1:

$$\begin{aligned} a_h(u,\varphi) = &\sum_{i\in I}\int_{K_i} \varepsilon\,\nabla u\cdot\nabla\varphi\,\mathrm{d}x \\ &-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}} \varepsilon\langle\nabla u\rangle\cdot\boldsymbol{n}_{ij}[\varphi]\,\mathrm{d}S - \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}} \varepsilon\langle\nabla\varphi\rangle\cdot\boldsymbol{n}_{ij}[u]\,\mathrm{d}S \\ &-\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}} \varepsilon\,\nabla\,u\cdot\boldsymbol{n}_{ij}\,\varphi\,\mathrm{d}S - \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}} \varepsilon\,\nabla\,\varphi\cdot\boldsymbol{n}_{ij}\,u\,\mathrm{d}S, \end{aligned} \tag{2.88}$$

$$J_h(u,\varphi) = \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}} \sigma[u]\,[\varphi]\,\mathrm{d}S + \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}} \sigma\,u\,\varphi\,\mathrm{d}S, \tag{2.89}$$

$$\begin{aligned} \ell_h(\varphi)(t) = &\int_\Omega g(t)\,\varphi\,\mathrm{d}x + \sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_N} g_N(t)\,\varphi\,\mathrm{d}S \\ &-\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}} \varepsilon\,\nabla\,\varphi\cdot\boldsymbol{n}_{ij}\,u_D(t)\,\mathrm{d}S + \varepsilon\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}} \sigma\,u_D(t)\,\varphi\,\mathrm{d}S, \end{aligned} \tag{2.90}$$

$$\begin{aligned} b_h(u,\varphi) = &-\sum_{i\in I}\int_K \sum_{s=1}^{d} f_s(u)\,\frac{\partial\varphi}{\partial x_s}\,\mathrm{d}x \\ &+\sum_{i\in I}\sum_{j\in S(i)}\int_{\Gamma_{ij}} H\left(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}\right)\varphi|_{\Gamma_{ij}}\,\mathrm{d}S. \end{aligned} \tag{2.91}$$

Again, the constant σ in (2.89) and (2.90) is defined by

$$\sigma|_{\Gamma_{ij}} = \frac{C_W}{d(\Gamma_{ij})}\,, \tag{2.92}$$

where

$$C_W \geq 4C_M(1+C_I) \tag{2.93}$$

and C_M and C_I are constants from (2.1.1) and (2.1.2), respectively. In this Section, we deal only with the *symmetric* variant and therefore we omit the superscript S in the diffusion form a_h^S and right-hand side form ℓ_h^S.

As in the previous section, we assume that the geometry of the mesh satisfies conditions of Section 2.1.1 and the numerical flux $H(u,v,\mathbf{n})$ is *Lipschitz-continuous*, *consistent* and *conservative* (page 35).

Now we can introduce the *discrete problem*.

Definition 1 *Let $u_h^0 \in S_h$ be the $L^2(\Omega)$-projection of the initial condition u^0 onto S_h, i.e. a function defined by*

$$(u_h^0 - u^0, \varphi_h) = 0 \qquad \forall \varphi_h \in S_h. \tag{2.94}$$

We say that u_h is a DGFE solution of the convection-diffusion problem (2.86)*, if*

$$\begin{aligned}
&\text{a)} \quad u_h \in C^1([0,T]; S_h),\\
&\text{b)} \quad \left(\frac{\partial u_h(t)}{\partial t}, \varphi_h\right) + b_h(u_h(t), \varphi_h) + a_h(u_h(t), \varphi_h) + \varepsilon J_h(u_h(t), \varphi_h) = \ell_h(\varphi_h)\,(t)\\
&\qquad\qquad \forall\, \varphi_h \in S_h,\ \forall\, t \in (0,T),\\
&\text{c)} \quad u_h(0) = u_h^0.
\end{aligned} \tag{2.95}$$

It is possible to show that a sufficiently regular exact solution u satisfies condition (2.95), b):

$$\left(\frac{\partial u(t)}{\partial t}, \varphi_h\right) + b_h(u(t), \varphi_h) + a_h(u(t), \varphi_h) + \varepsilon J_h^\sigma(u(t), \varphi_h) = \ell_h(\varphi_h)\,(t) \tag{2.96}$$

$$\text{for all } \varphi_h \in S_h \text{ and for a.a. } t \in (0,T),$$

which implies the *Galerkin orthogonality property* of the error.

2.3.2 Error analysis

In the following we shall use all the assumptions and results from Section 2.1. Let us deal with properties of the forms a_h and J_h. In [15, Corollary 3.10] it was shown that under this assumption on C_W the forms a_h and J_h have the following coercivity property:

$$\begin{aligned}
a_h(\varphi_h, \varphi_h) + J_h(\varphi_h, \varphi_h) \geq \frac{\varepsilon}{2}\left(|\varphi_h|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(\varphi_h, \varphi_h)\right)\\
\forall\, \varphi_h \in S_h, \quad \forall h \in (0, h_0).
\end{aligned} \tag{2.97}$$

Moreover, in the same way as in [15, estimates (3.22) and (3.28)] we can prove the existence of a constant $C > 0$ such that for any $u \in H^{p+1}(\Omega)$, $\varphi_h \in S_h$ and $h \in (0, h_0)$ we have

$$|a_h(\Pi_h u - u, \varphi_h)| \leq \varepsilon C\, h^p |u|_{H^{p+1}(\Omega)} \left(J_h(\varphi_h, \varphi_h)^{1/2} + |\varphi_h|_{H^1(\Omega,\mathcal{T}_h)}\right), \tag{2.98}$$

$$|J_h(\Pi_h u - u,\, \Pi_h u - u)| \leq C\, h^{2p} |u|^2_{H^{p+1}(\Omega)}. \tag{2.99}$$

Now, let us define the form

$$A_h(w, v) = a_h(w, v) + \varepsilon J_h(w, v) \quad v,\ w \in H^2(\Omega, \mathcal{T}_h), \tag{2.100}$$

and set

$$\|w\|_{DG} = \left(\frac{1}{2} \left(|w|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h(w,w) \right) \right)^{1/2}. \tag{2.101}$$

Since the $(d-1)$-dimensional measure of Γ_D is positive, $\|\cdot\|_{DG}$ is a norm in $H^1(\Omega,\mathcal{T}_h)$. With the use of these notations we can rewrite estimates (2.97) and (2.98) in the form

$$A_h(\varphi_h,\varphi_h) \geq \varepsilon \|\varphi_h\|^2_{DG} \tag{2.102}$$

and

$$|a_h(\Pi_h u - u,\, \varphi_h)| \leq \varepsilon C\, h^p |u|_{H^{p+1}(\Omega)}\, \|\varphi_h\|_{DG}, \tag{2.103}$$

respectively.

For each $h \in (0,h_0)$ and $t \in [0,T]$ we define the function $u^*(t)$ $(= u^*_h(t))$ as the "A_h-projection" of $u(t)$ on S_h, i. e. a function satisfying the conditions

$$u^*(t) \in S_h, \qquad A_h(u^*(t),\varphi_h) = A_h(u(t),\varphi_h) \quad \forall\, \varphi_h \in S_h. \tag{2.104}$$

(In what follows we shall usually omit the argument t of the functions u and u^*.)

First, we shall derive estimates of the functions $\chi = u - u^*$ and $\chi_t = \frac{\partial \chi}{\partial t}$ in the norm $\|\cdot\|_{DG}$ and in the $L^2(\Omega)$-norm.

Lemma 2.3.1 *There exists a constant $C > 0$ such that*

$$\|\chi\|_{DG} \leq C\, h^p |u|_{H^{p+1}(\Omega)}, \tag{2.105}$$

$$\|\chi_t\|_{DG} \leq C\, h^p |u_t|_{H^{p+1}(\Omega)} \tag{2.106}$$

for all $h \in (0,h_0)$.

Proof: Let us set $\hat{u} = \Pi_h u$, the L^2-projection of u onto the space S_h. By (2.102) and (2.104) and the definition of u^* we obtain

$$\begin{aligned} \varepsilon \|\hat{u} - u^*\|^2_{DG} &\leq A_h(\hat{u} - u^*,\, \hat{u} - u^*) \\ &= A_h(\hat{u} - u^*,\, \hat{u} - u^*) + A_h(u^* - u,\, \hat{u} - u^*) \\ &= A_h(\hat{u} - u,\, \hat{u} - u^*) \\ &= a_h(\hat{u} - u,\, \hat{u} - u^*) + \varepsilon J_h(\hat{u} - u,\, \hat{u} - u^*). \end{aligned} \tag{2.107}$$

From (2.103) we have

$$a_h(\hat{u} - u,\, \hat{u} - u^*) \leq C\varepsilon h^p |u|_{H^{p+1}(\Omega)} \|\hat{u} - u^*\|_{DG}. \tag{2.108}$$

If we combine Lemma 2.1.5 with the definition of the norm $\|\,.\,\|_{DG}$ and with (2.99), we obtain

$$\begin{aligned} J_h(\hat{u} - u,\, \hat{u} - u^*) &\leq \left(J_h(\hat{u} - u,\, \hat{u} - u)\right)^{1/2} \left(J_h(\hat{u} - u^*,\, \hat{u} - u^*)\right)^{1/2} \\ &\leq C\, h^p |u|_{H^{p+1}(\Omega)} \|\hat{u} - u^*\|_{DG}. \end{aligned} \tag{2.109}$$

From (2.107) – (2.109) we get

$$\|\hat{u} - u^*\|_{DG} \leq C\, h^p |u|_{H^{p+1}(\Omega)}. \tag{2.110}$$

Further, in virtue of the regularity of u, Lemma 2.1.4, b) and (2.99), we have

$$\|u - \hat{u}\|_{DG} = \|\eta\|_{DG} \leq Ch^p |u|_{H^{p+1}(\Omega)}.$$

Now it is sufficient to use the triangle inequality

$$\|u - u^*\|_{DG} \leq \|u - \hat{u}\|_{DG} + \|\hat{u} - u^*\|_{DG},$$

which implies that

$$\|\chi\|_{DG} = \|u - u^*\|_{DG} \leq C\, h^p |u|_{H^{p+1}(\Omega)}.$$

Hence, (2.105) is proven.

Let us deal now with the norm $\|\chi_t\|_{DG}$. As

$$A_h\left(u(t) - u^*(t),\, \varphi_h\right) = 0 \qquad \forall\, \varphi_h \in S_h,\ \forall\, t \in (0, T),$$

from the definitions of a_h and J_h, for all $\varphi_h \in S_h$ we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left(A_h\left(u(t) - u^*(t),\, \varphi_h\right)\right) = A_h\left(\frac{\partial(u(t) - u^*(t))}{\partial t},\, \varphi_h\right),$$

i. e.

$$A_h(\chi_t, \varphi_h) = 0 \qquad \forall\, \varphi_h \in S_h.$$

Obviously

$$\Pi_h u_t = (\Pi_h u)_t = \hat{u}_t \in S_h \quad \text{and} \quad u_t^* \in S_h.$$

In the same way as in the estimation of the norm $\|\chi\|_{DG}$ we now obtain

$$\begin{aligned} \varepsilon\|\hat{u}_t - u_t^*\|_{DG}^2 &\leq A_h\left(\hat{u}_t - u_t^*,\, \hat{u}_t - u_t^*\right) + A_h\left(u_t^* - u_t,\, \hat{u}_t - u_t^*\right) \\ &= A_h\left(\hat{u}_t - u_t,\, \hat{u}_t - u_t^*\right) \\ &\leq \varepsilon C\, h^p |u_t|_{H^{p+1}(\Omega)} \|\hat{u}_t - u_t^*\|_{DG}, \end{aligned}$$

which implies that

$$\|\hat{u}_t - u_t^*\|_{DG} \leq C\, h^p |u_t|_{H^{p+1}(\Omega)},$$

and thus,

$$\|u_t - u_t^*\|_{DG} \leq \|u_t - \hat{u}_t\|_{DG} + \|\hat{u}_t - u_t^*\|_{DG} \leq C\, h^p |u_t|_{H^{p+1}(\Omega)}.$$

Hence, we have obtained the desired estimate (2.106). □

In what follows, for an arbitrary $z \in L^2(\Omega)$ we shall consider the dual problem: Given $z \in L^2(\Omega)$, find ψ such that

$$\begin{aligned} -\Delta\psi &= z \quad \text{in } \Omega, \\ \psi|_{\Gamma_D} &= 0, \\ \frac{\partial\psi}{\partial n}\Big|_{\Gamma_N} &= 0. \end{aligned} \tag{2.111}$$

Under the notation

$$V = \left\{ v \in C^\infty(\overline{\Omega}),\, \operatorname{supp} v \subset \Omega \cup \Gamma_N \right\},$$

the weak formulation of (2.111) reads: Find $\psi \in H^1(\Omega)$ such that $\psi|_{\Gamma_D} = 0$ and

$$(\nabla\psi,\, \nabla v) = (z, v) \quad \forall\, v \in V. \tag{2.112}$$

Let us assume that $\psi \in H^2(\Omega)$ and there exists a constant $C > 0$, independent of z such that

$$\|\psi\|_{H^2(\Omega)} \le C\|z\|_{L^2(\Omega)}. \tag{2.113}$$

As the domain Ω is convex, this is true, e.g. provided $\Gamma_N = \emptyset$, as follows from [25]. Let us note that $H^2(\Omega) \subset C(\overline{\Omega})$.

Further, let ψ_h be the piecewise linear L^2-projection of the function ψ, i.e. $\psi|_K \in P^1(K)$ and

$$\big(\psi - \psi_h, \varphi_h\big)_{L^2(K)} = 0, \quad \forall \varphi_h \in P^1(K),\ \forall K \in \mathcal{T}_h. \tag{2.114}$$

Lemma 2.3.2 *The following estimates hold:*

$$\begin{aligned} J_h(\psi - \psi_h, \psi - \psi_h) &\le C\, h^2 |\psi|^2_{H^2(\Omega)} \\ \|\psi - \psi_h\|^2_{DG} &\le C\, h^2 |\psi|^2_{H^2(\Omega)} \end{aligned} \tag{2.115}$$

Proof: Obviously $\psi_h \in S_h$. Due to Lemma 2.1.1 and estimates (2.6) we have

$$\begin{aligned} &J_h(\psi - \psi_h, \psi - \psi_h) \\ &\quad = \sum_{i\in I} \sum_{\substack{j\in s(i) \\ j<i}} \int_{\Gamma_{ij}} \frac{C_W}{d(\Gamma_{ij})} [\psi - \psi_h]^2 \,\mathrm{d}S + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \frac{C_W}{d(\Gamma_{ij})} (\psi - \psi_h)^2 \,\mathrm{d}S \\ &\quad \le C \sum_{i\in I} \int_{\partial K_i} \frac{1}{h_{K_i}} (\psi - \psi_h)^2 \,\mathrm{d}x \\ &\quad \le C \sum_{i\in I} \frac{C_M}{h_{K_i}} \Big(||\psi - \psi_h||_{L^2(K_i)} |\psi - \psi_h|_{H^1(K_i)} + h_{K_i}^{-1} ||\psi - \psi_h||_{L^2(K_i)} \Big) \\ &\quad \le C \sum_{i\in I} \frac{1}{h_{K_i}} h_{K_i}^3 |\psi|^2_{H^2(K_i)} \le C h^2 |\psi|^2_{H^2(\Omega)} \end{aligned}$$

The second inequality in (2.115) follows from the preceding and estimate (2.6):

$$\|\psi-\psi_h\|_{DG}^2=\frac{1}{2}\Big(|\psi-\psi_h|^2_{H^1(\Omega,\mathcal{T}_h)}+J_h(\psi-\psi_h,\psi-\psi_h)\Big)\le C\,h^2|\psi|^2_{H^2(\Omega)}$$

□

Now we shall use the dual problem (2.111) to obtain L^2-optimal error estimates for χ and χ_t.

Lemma 2.3.3 *There exists a constant $C>0$ such that*

$$\|\chi\|_{L^2(\Omega)} \le Ch^{p+1}|u|_{H^{p+1}(\Omega)}, \tag{2.116}$$
$$\|\chi_t\|_{L^2(\Omega)} \le Ch^{p+1}|u_t|_{H^{p+1}(\Omega)} \tag{2.117}$$

for all $h\in(0,h_0)$.

Proof: We have

$$\|\chi\|_{L^2(\Omega)}=\sup_{z\in L^2(\Omega)}\frac{(\chi,z)}{\|z\|_{L^2(\Omega)}}. \tag{2.118}$$

Now, using (2.111), Green's theorem, the homogeneous Neumann condition and the fact that the continuity of functions from the space $H^2(\Omega)$ yields

$$[\psi]_{\Gamma_{ij}}=0 \qquad \forall i\in I,\ j\in s(i), \tag{2.119}$$

we obtain

$$\begin{aligned}
(\chi,z)&=\int_\Omega z\chi\,\mathrm{d}x=-\int_\Omega\Delta\psi\chi\,\mathrm{d}x\\
&=\sum_{i\in I}\int_{K_i}\nabla\psi\cdot\nabla\chi\,\mathrm{d}x-\sum_{i\in I}\sum_{j\in s(i)}\int_{\Gamma_{ij}}\nabla\psi\cdot\boldsymbol{n}_{ij}\,\chi\,\mathrm{d}S\\
&\quad-\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\nabla\psi\cdot\boldsymbol{n}_{ij}\,\chi\,\mathrm{d}S-\sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}}\nabla\psi\cdot\boldsymbol{n}_{ij}\,\chi\,\mathrm{d}S\\
&=\sum_{i\in I}\int_{K_i}\nabla\psi\cdot\nabla\chi\,\mathrm{d}x\\
&\quad-\left(\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\nabla\psi\rangle\cdot\boldsymbol{n}_{ij}\,[\chi]\,\mathrm{d}S+\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\nabla\chi\rangle\cdot\boldsymbol{n}_{ij}\,[\psi]\,\mathrm{d}S\right)\\
&\quad-\left(\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\nabla\psi\cdot\boldsymbol{n}_{ij}\,\chi\,\mathrm{d}S+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\nabla\chi\cdot\boldsymbol{n}_{ij}\,\psi\,\mathrm{d}S\right)\\
&\quad+\left(\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sigma\,[\psi]\,[\chi]\,\mathrm{d}S+\sum_{i\in I}\sum_{j\in\Gamma_D(i)}\int_{\Gamma_{ij}}\sigma\,\psi\,\chi\,\mathrm{d}S\right),
\end{aligned}$$

i. e.,

$$(\chi, z) = \frac{1}{\varepsilon} A_h(\psi, \chi). \tag{2.120}$$

Further, the symmetry of A_h and (2.104) give

$$A_h(\psi_h, \chi) = A_h(\chi, \psi_h) = A_h(u - u^*, \psi_h) = 0, \tag{2.121}$$

and thus

$$(\chi, z) = \frac{1}{\varepsilon} A_h(\psi - \psi_h, \chi) = A_1 + A_2 + A_3 + A_4,$$

where

$$A_1 := \sum_{i\in I} \int_{K_i} \nabla(\psi - \psi_h) \cdot \nabla\chi \,\mathrm{d}x$$

$$A_2 := -\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \nabla(\psi - \psi_h)\rangle \cdot \boldsymbol{n}_{ij}[\chi] \,\mathrm{d}S - \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \nabla(\psi - \psi_h) \cdot \boldsymbol{n}_{ij}\chi \,\mathrm{d}S$$

$$A_3 := -\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \nabla\chi\rangle \cdot \boldsymbol{n}_{ij}[\psi - \psi_h] \,\mathrm{d}S - \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \nabla\chi \cdot \boldsymbol{n}(\psi - \psi_h) \,\mathrm{d}S$$

$$A_4 := \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sigma[\psi - \psi_h]\,[\chi] \,\mathrm{d}S + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma(\psi - \psi_h)\,\chi \,\mathrm{d}S$$

Now we estimate these terms individually. First, due to Lemmas 2.3.1 and 2.3.2, we have

$$\begin{aligned} A_1 &\le |\psi - \psi_h|_{H^1(\Omega)} |\chi|_{H^1(\Omega,\mathcal{T}_h)} \le C\, h |\psi|_{H^2(\Omega)} \,\|\chi\|_{DG} \\ &\le C\, h^{p+1} |u|_{H^{p+1}(\Omega)} \|z\|_{L^2(\Omega)}. \end{aligned}$$

Next we write

$$\begin{aligned} A_2 = &-\sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \left(\frac{d(\Gamma_{ij})}{C_W}\right)^{1/2} \langle \nabla(\psi - \psi_h)\rangle \cdot \boldsymbol{n}_{ij} \left(\frac{C_W}{d(\Gamma_{ij})}\right)^{1/2} [\chi] \,\mathrm{d}S \\ &-\sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \left(\frac{d(\Gamma_{ij})}{C_W}\right)^{1/2} \nabla(\psi - \psi_h) \cdot \boldsymbol{n}_{ij} \left(\frac{C_W}{d(\Gamma_{ij})}\right)^{1/2} \chi \,\mathrm{d}S \\ &\le \frac{1}{C_W^{1/2}} \left(\sum_{i\in I} h_{K_i} \|\nabla(\psi - \psi_h)\|^2_{L^2(\partial K_i)}\right)^{1/2} \left(J_h(\chi,\chi)\right)^{1/2}. \end{aligned}$$

According to the multiplicative trace inequality (2.1.1), [7, Theorem 3.1.6], Lemma

2.3.1, (2.113) and (2.105), we can write

$$\begin{aligned}
A_2 &\le \frac{1}{C_W^{1/2}} \left(\sum_{i\in I} h_{K_i} \|\nabla(\psi-\psi_h)\|^2_{L^2(\partial K_i)} \right)^{1/2} (J_h(\chi,\chi))^{1/2} \\
&\le \frac{\sqrt{2}\|\chi\|_{DG}}{C_W^{1/2}} \Bigg(\sum_{i\in I} h_{K_i} \Big(|\nabla(\psi-\psi_h)|_{H^1(K_i)} \, \|\nabla(\psi-\psi_h)\|_{L^2(K_i)} + \\
&\qquad + h_{K_i}^{-1} \|\nabla(\psi-\psi_h)\|^2_{L^2(K_i)} \Big) \Bigg)^{1/2} \\
&\le C\|\chi\|_{DG} \left(\sum_{i\in I} h_{K_i}^2 |\psi|^2_{H^2(K_i)} \right)^{1/2} \le C\, h^{p+1} |u|_{H^{p+1}(\Omega)} \, \|z\|_{L^2(\Omega)}.
\end{aligned}$$

Now we estimate A_3:

$$\begin{aligned}
A_3 &= -\sum_{i\in I} \sum_{\substack{j\in s(i) \\ j<i}} \int_{\Gamma_{ij}} \langle \nabla\chi \rangle \cdot \boldsymbol{n}_{ij} [\psi-\psi_h] \, \mathrm{d}S - \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \nabla\chi \cdot \boldsymbol{n} (\psi-\psi_h) \, \mathrm{d}S \\
&\le \frac{1}{C_W^{1/2}} \Big(\sum_{i\in I} h_{K_i} ||\nabla\chi||^2_{L^2(\partial K_i)} \Big)^{1/2} \big(J_h(\psi-\psi_h, \psi-\psi_h) \big)^{1/2} \\
&\le Ch|\psi|_{H^2(\Omega)} \Big(\sum_{i\in I} h_{K_i} |\nabla\chi|_{H^1(K_i)} ||\nabla\chi||_{L^2(K_i)} + ||\nabla\chi||^2_{L^2(K_i)} \Big)^{1/2} \\
&\le Ch|\psi|_{H^2(\Omega)} \Big(\sum_{i\in I} h_{K_i} \big(|\nabla(\hat{u}-u^*)|_{H^1(K_i)} + |\nabla(u-\hat{u})|_{H^1(K_i)} \big) ||\nabla\chi||_{L^2(K_i)} \\
&\qquad + ||\nabla\chi||^2_{L^2(K_i)} \Big)^{1/2},
\end{aligned}$$

where $\hat{u} = \Pi_h u$ is the L^2 projection of u onto the space S_h. Now since $\hat{u} - u^* \in S_h$, we can apply the inverse inequality, result (2.110) and estimate (2.7) and write

$$\begin{aligned}
A_3 &\le Ch|\psi|_{H^2(\Omega)} \Bigg(\sum_{i\in I} h_{K_i} \Big(\frac{C_I}{h_{K_i}} ||\nabla(\hat{u}-u^*)||_{L^2(K_i)} + |\nabla(u-\hat{u})|_{H^1(K_i)} \Big) ||\nabla\chi||_{L^2(K_i)} \\
&\qquad + ||\nabla\chi||^2_{L^2(K_i)} \Bigg)^{1/2} \\
&\le Ch|\psi|_{H^2(\Omega)} \left(\Big(||\nabla(\hat{u}-u^*)||_{L^2(\Omega)} + h|u-\hat{u}|_{H^2(\Omega)} \Big) ||\nabla\chi||_{L^2(\Omega)} + ||\nabla\chi||^2_{L^2(\Omega)} \right)^{1/2} \\
&\le Ch\|z\|_{L^2(\Omega)} \, h^p |u|_{H^{p+1}(\Omega)} = C\, h^{p+1} \|z\|_{L^2(\Omega)} |u|_{H^{p+1}(\Omega)}
\end{aligned}$$

To estimate the final term A_4, we can apply Lemmas 2.1.5, 2.3.1 and 2.3.2 to

obtain:

$$A_4 = J_h(\psi - \psi_h, \chi) \le \big(J_h(\psi - \psi_h, \psi - \psi_h)\big)^{1/2} \big(J_h(\chi, \chi)\big)^{1/2}$$
$$\le Ch|\psi|_{H^2(K_i)}||\chi||_{DG} \le C\, h^{p+1}|u|_{H^{p+1}(\Omega)}\, \|z\|_{L^2(\Omega)}.$$

Combining the previous estimates, we find that

$$(\chi, z) = A_1 + A_2 + A_3 + A_4 \le C\, h^{p+1}|u|_{H^{p+1}(\Omega)}\|z\|_{L^2(\Omega)}.$$

Hence,

$$\|\chi\|_{L^2(\Omega)} = \sup_{z \in L^2(\Omega)} \frac{(\chi, z)}{\|z\|_{L^2(\Omega)}} \le C\, h^{p+1}|u|_{H^{p+1}(\Omega)},$$

which completes the proof of (2.116).

In the derivation of the estimate of the norm $\|\chi_t\|_{L^2(\Omega)}$ we proceed similarly as in the estimations of the norms $\|\chi_t\|_{DG}$ and $\|\chi\|_{L^2(\Omega)}$. □

Let us note that the assumption of the symmetry of the form A_h is crucial in the presented proof. It enables us to exchange arguments in (2.121). This is the reason, why we are unable to prove optimal error estimates for the nonsymmetric and incomplete variants of the DG scheme using the presented technique.

Now, we shall be concerned with the properties of the form b_h. Let assumptions (A) and (H) be satisfied. We formulate the following result, which is similar to Lemma 2.2.2, only, instead of $\xi = \Pi_h u - u_h$ we use $\zeta = u^* - u_h$.

Lemma 2.3.4 (Properties of the convective terms) *Let u be the solution of the continuous problem* (2.86)*, u_h the solution of the discrete problem* (2.95)*, u^* be defined by* (2.104)*, and ζ $(= \zeta_h) = u^* - u_h \in S_h$. Then there exists a constant $C > 0$, independent of $h \in (0, h_0)$, such that*

$$|b_h(u, \zeta) - b_h(u_h, \zeta)| \le C\|\zeta\|_{DG} \left(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\zeta\|_{L^2(\Omega)}\right). \tag{2.122}$$

Proof: We can write

$$\begin{aligned} &\big|b_h(u, \zeta) - b_h(u_h, \zeta)\big| \\ &\qquad \le |b_h(u, \zeta) - b_h(u^*, \zeta)| + |b_h(u^*, \zeta) - b_h(u_h, \zeta)|\,. \end{aligned} \tag{2.123}$$

According to (2.30), (2.101) and (2.1.1),

$$\begin{aligned}
&\big|b_h(u,\zeta)-b_h(u^*,\zeta)\big| \\
&\quad\le C\|\zeta\|_{DG}\left(\|u-u^*\|_{L^2(\Omega)}+\left(\sum_{i\in I}h_{K_i}\|u-u^*\|^2_{L_2(\partial K_i)}\right)^{1/2}\right) \\
&\quad= C\|\zeta\|_{DG}\left(\|\chi\|_{L^2(\Omega)}+\left(\sum_{i\in I}h_{K_i}\|\chi\|^2_{L^2(\partial K_i)}\right)^{1/2}\right) \qquad (2.124)\\
&\quad\le C\|\zeta\|_{DG}\Bigg(\|\chi\|_{L^2(\Omega)}+ \\
&\qquad+\left(C_M\sum_{i\in I}h_{K_i}\Big(\|\chi\|_{L^2(K_i)}|\chi|_{H^1(K_i)}+h_{K_i}^{-1}\|\chi\|^2_{L^2(K_i)}\Big)\right)^{1/2}\Bigg).
\end{aligned}$$

Further, by the Cauchy inequality and Lemmas 2.3.1 and 2.3.3, we have

$$\begin{aligned}
&\sum_{i\in I}h_{K_i}\Big(\|\chi\|_{L^2(K_i)}|\chi|_{H^1(K_i)}+h_{K_i}^{-1}\|\chi\|^2_{L^2(K_i)}\Big) \\
&\quad\le\left(\sum_{i\in I}\|\chi\|^2_{L^2(K_i)}\right)^{1/2}\left(\sum_{i\in I}h^2_{K_i}|\chi|^2_{H^1(K_i)}\right)^{1/2}+\sum_{i\in I}\|\chi\|^2_{L^2(K_i)} \qquad (2.125)\\
&\quad\le\|\chi\|_{L^2(\Omega)}\,h\,|\chi|_{H^1(\Omega,\mathcal{T}_h)}+\|\chi\|^2_{L^2(\Omega)} \\
&\quad\le\|\chi\|_{L^2(\Omega)}\,\sqrt{2}h\|\chi\|_{DG}+\|\chi\|^2_{L^2(\Omega)} \\
&\quad\le C\,h^{2(p+1)}|u|^2_{H^{p+1}(\Omega)}.
\end{aligned}$$

Hence, by (2.124), (2.125) and Lemma 2.3.3,

$$|b_h(u,\zeta)-b_h(u^*,\zeta)|\le C\|\zeta\|_{DG}\,h^{p+1}|u|_{H^{p+1}(\Omega)}. \qquad (2.126)$$

Finally, it remains to estimate the second term in (2.123). We use Lemma 2.30 and Lemma 2.1.6, c), which yields

$$\begin{aligned}
&|b_h(u^*,\zeta)-b_h(u_h,\zeta)| \\
&\quad\le C\|\zeta\|_{DG}\Big(\|\zeta\|_{L^2(\Omega)}+\Big(\sum_{i\in I}h_{K_i}\|\zeta\|^2_{L_2(\partial K_i)}\Big)^{1/2}\Big) \qquad (2.127)\\
&\quad\le C\|\zeta\|_{DG}\,\|\zeta\|_{L^2(\Omega)}.
\end{aligned}$$

Now, the combination of (2.123), (2.126) and (2.127) gives the desired estimate (2.122), which we wanted to prove. □

Optimal error estimate for the method of lines

Now we can proceed to the *main result* of this section, which is an optimal error estimate of the method in the norm of the space $L^\infty(0,T;L^2(\Omega))$. In the proof we shall apply the following simplified version of Gronwall's lemma (2.1.7).

Lemma 2.3.5 *Let y, $q \in C([0,T])$, y, $q \geq 0$ in $[0,T]$, Z, $R \in \mathbb{R}$, $R \geq 0$ and*

$$y(t) + q(t) \leq Z + R \int_0^t y(s)\,\mathrm{d}S, \quad t \in [0,T].$$

Then

$$y(t) + q(t) \leq Z \exp(Rt), \quad t \in [0,T].$$

Proof: In Lemma (2.1.7) we simply set $z(t) := Z$ and $r(t) := R$ and write inequality (2.18) for $t \in [0,T]$:

$$y(t) + q(t) + \underbrace{\int_0^t Rq(\vartheta) \exp\Big(\int_\vartheta^t R\,\mathrm{d}s\Big)\,\mathrm{d}\vartheta}_{\geq 0} \leq Z + \int_0^t RZ \exp\Big(\int_\vartheta^t R\,\mathrm{d}s\Big)\,\mathrm{d}\vartheta.$$

Now we can neglect the positive term on the left-hand side and write

$$\begin{aligned} y(t) + q(t) &\leq Z + \int_0^t RZ \exp\Big(\int_\vartheta^t R\,\mathrm{d}s\Big)\,\mathrm{d}\vartheta \\ &= Z + RZ \int_0^t e^{R(t-\vartheta)}\,\mathrm{d}\vartheta = Z + Z\big(e^{Rt} - 1\big) = Ze^{Rt}. \end{aligned}$$

□

Theorem 2.3.1 (Main theorem) *Let assumptions* (H) *and* (A) *be satisfied and* (2.93) *hold. Let u be the exact solution of problem* (2.86) *satisfying the regularity condition* (2.87) *and let u_h be the approximate solution defined by* (2.95). *Moreover, let the solution of the dual problem* (2.111) *satisfy* (2.113). *Then the error $e_h = u - u_h$ satisfies the estimate*

$$\|e_h\|_{L^\infty(0,T;L^2(\Omega))} \leq Ch^{p+1}, \tag{2.128}$$

with a constant $C > 0$ independent of h.

Proof: Let u^* be defined by (2.104) and let χ and ζ be as in Lemmas 2.3.1, 2.3.3 and 2.3.4, i. e. $\chi = u - u^*$, $\zeta = u^* - u_h$. Then $e_h = u - u_h = \chi + \zeta$. Let us subtract (2.96) from (2.95, b), substitute $\zeta \in S_h$ for φ_h and use the relation

$$\left(\frac{\partial \zeta(t)}{\partial t}, \zeta(t)\right) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\zeta(t)\|^2_{L^2(\Omega)}.$$

Then we get

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\zeta(t)\|^2_{L^2(\Omega)} + A_h(\zeta(t),\zeta(t)) \qquad (2.129)\\
&\quad = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\zeta(t)\|^2_{L^2(\Omega)} + a_h(\zeta(t),\zeta(t)) + \varepsilon J_h(\zeta(t),\zeta(t))\\
&\quad = b_h(u(t),\zeta(t)) - b_h(u_h(t),\zeta(t)) - (\chi_t(t),\zeta(t)) - A_h\left(u(t)-u^*(t),\zeta(t)\right)\\
&\quad = [b_h(u(t),\zeta(t)) - b_h(u_h(t),\zeta(t))] - (\chi_t(t),\zeta(t)),
\end{aligned}$$

because $A_h(u(t)-u^*(t),\zeta(t)) = 0$. The first right-hand side term can be estimated by Lemma 2.3.4 and Young's inequality as follows (we omit the argument t)

$$\begin{aligned}
b_h(u,\zeta) - b_h(u_h,\zeta) &\le C\|\zeta\|_{DG}\left(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\zeta\|_{L^2(\Omega)}\right) \qquad (2.130)\\
&\le \frac{\varepsilon}{2}\|\zeta\|^2_{DG} + \frac{C}{\varepsilon}\left(h^{2(p+1)}|u|^2_{H^{p+1}(\Omega)} + \|\zeta\|^2_{L^2(\Omega)}\right).
\end{aligned}$$

For the second right-hand side term in (2.129), by the Cauchy and Young's inequalities and Lemma 2.3.3, we have

$$\begin{aligned}
|(\chi_t,\zeta)| &\le \|\chi_t\|_{L^2(\Omega)}\|\zeta\|_{L^2(\Omega)} \le \frac{1}{2}\left(\|\chi_t\|^2_{L^2(\Omega)} + \|\zeta\|^2_{L^2(\Omega)}\right) \qquad (2.131)\\
&\le \frac{1}{2}\left(C\,h^{2(p+1)}|u_t|^2_{H^{p+1}(\Omega)} + \|\zeta\|^2_{L^2(\Omega)}\right).
\end{aligned}$$

Finally, the coercivity property (2.102) of A_h gives the estimate of the left-hand side of (2.129).

Hence, combining (2.129) – (2.131) and (2.102), we obtain

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\|\zeta\|^2_{L^2(\Omega)} + \varepsilon\|\zeta\|^2_{DG} \qquad (2.132)\\
\le\ & C\,h^{2(p+1)}\left(\frac{1}{\varepsilon}|u|^2_{H^{p+1}(\Omega)} + |u_t|^2_{H^{p+1}(\Omega)}\right) + C\left(1+\frac{1}{\varepsilon}\right)\|\zeta\|^2_{L^2(\Omega)}.
\end{aligned}$$

The integration of (2.132) from 0 to $t \in [0,T]$ yields

$$\begin{aligned}
&\|\zeta(t)\|^2_{L^2(\Omega)} + \varepsilon\int_0^t \|\zeta(\vartheta)\|^2_{DG}\,\mathrm{d}\vartheta \qquad (2.133)\\
\le\ & C\,h^{2(p+1)}\left(\frac{1}{\varepsilon}\int_0^t |u(\vartheta)|^2_{H^{p+1}(\Omega)}\,\mathrm{d}\vartheta + \int_0^t |u_t(\vartheta)|^2_{H^{p+1}(\Omega)}\,\mathrm{d}\vartheta\right)\\
&\quad + C\left(1+\frac{1}{\varepsilon}\right)\int_0^t \|\zeta(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta + C\,h^{2(p+1)}|u^0|^2_{H^{p+1}(\Omega)},
\end{aligned}$$

since

$$\|\zeta(0)\|^2_{L^2(\Omega)} \le C\,h^{2(p+1)}|u^0|^2_{H^{p+1}(\Omega)}. \qquad (2.134)$$

To prove (2.134), we first write

$$\zeta(0) = u_h(0) - u_h^*(0) = (u_h(0) - u^0) + \big(u^0 - u^*(0)\big) = (u_h^0 - u^0) + (u(0) - u^*(0)). \tag{2.135}$$

Then, by (2.5) and Lemma 2.3.3,

$$\begin{aligned} \|u_h^0 - u^0\|_{L^2(\Omega)} &= \|\Pi_h u^0 - u^0\|_{L^2(\Omega)} \le C\, h^{p+1} |u^0|_{H^{p+1}(\Omega)}, \\ \|u(0) - u^*(0)\|_{L^2(\Omega)} &= \|\chi(0)\|_{L^2(\Omega)} \le C\, h^{p+1} |u(0)|_{H^{p+1}(\Omega)} = C\, h^{p+1} |u^0|_{H^{p+1}(\Omega)}. \end{aligned}$$

This together with (2.135) already implies estimate (2.134).

Now we apply Gronwall's Lemma 2.3.5 with

$$\begin{aligned} y(t) &= \|\zeta(t)\|^2_{L^2(\Omega)}, \\ R &= C\left(1 + \frac{1}{\varepsilon}\right), \\ Z &= C h^{2(p+1)} N(\varepsilon, u), \\ q(t) &= \varepsilon \int_0^t \|\zeta(\vartheta)\|^2_{DG}\, \mathrm{d}\vartheta, \end{aligned}$$

where

$$N(\varepsilon, u) = \frac{1}{\varepsilon} \|u\|^2_{L^2(0,T;H^{p+1}(\Omega))} + \|u_t\|^2_{L^2(0,T;H^{p+1}(\Omega))} + |u^0|^2_{H^{p+1}(\Omega)}.$$

This yields

$$\|\zeta(t)\|^2_{L^2(\Omega)} + \varepsilon \int_0^t \|\zeta(\vartheta)\|^2_{DG}\, \mathrm{d}\vartheta \le C\, h^{2(p+1)}\, N(\varepsilon, u)\, \exp\left(\tilde{C}\left(1 + \frac{1}{\varepsilon}\right) t\right) \tag{2.136}$$

(C and $\tilde{C}$ are constants independent of t, h, ε, u).

Since $e_h = \chi + \zeta$, to complete the proof, it is sufficient now to combine (2.136) with the estimate of $\|\chi(t)\|_{L^2(\Omega)}$ from Lemma 2.3.3. □

The effect of numerical integration

In practical computations the integrals appearing in (2.88) – (2.91) are evaluated with the aid of numerical integration. We shall show how to choose quadrature formulae in order to preserve the accuracy of the method. For the sake of simplicity, we shall restrict ourselves to the case $d = 2$.

Let $F \in C(K)$ and $G \in C(\Gamma)$, where $K \in \mathcal{T}_h$ and Γ is a side of K. We consider the following approximations of integrals

$$\int_K F\, \mathrm{d}x \approx |K| \sum_{\alpha=1}^{n_K} \omega_\alpha^K F(x_\alpha^K), \tag{2.137}$$

$$\int_\Gamma G\, \mathrm{d}S \approx |\Gamma| \sum_{\alpha=1}^{m_\Gamma} \beta_\alpha^\Gamma G(x_\alpha^\Gamma). \tag{2.138}$$

The constants $\omega_\alpha^K, \beta_\alpha^\Gamma \in \mathbb{R}$ represent here integration weights and $x_\alpha^K \in K$, $x_\alpha^\Gamma \in \Gamma$ are integration points. In order to be able to evaluate the effect of the numerical integration, let us assume that:

(B1) The integration points x_α^Γ are chosen and numbered in such a way that for each $i \in I$ and $j \in s(i)$ the integration points on $\Gamma_{ij} = \Gamma_{ji}$ satisfy the condition

$$x_\alpha^{\Gamma_{ij}} = x_\alpha^{\Gamma_{ji}}. \tag{2.139}$$

(B2) There exist constants $\omega,\ \beta > 0$ such that $\forall K \in \mathcal{T}_h,\ \forall \Gamma \in \{\Gamma_{ij}; j \in S(i), i \in I\}$ and $\forall h \in (0, h_0)$ we have

$$\sum_{\alpha=1}^{n_K} |\omega_\alpha^K| \le \omega,\ \sum_{\alpha=1}^{m_\Gamma} |\beta_\alpha^\Gamma| \le \beta, \tag{2.140}$$

(B3) The quadrature formulae (2.137) and (2.138) are exact for polynomials of degree $\le 2p$ and $\le 2p+1$, respectively.

Using formulae (2.137) and (2.138), we obtain the approximations $(\cdot,\cdot)_I,\ a_I,\ J_I^\sigma,\ b_I,\ \ell_I$ of the forms $(\cdot,\cdot),\ a_h,\ J_h^\sigma,\ b_h,\ \ell_h$. With the aid of these forms we can formulate the *semidiscrete problem with numerical integration*:

Find u_I such that

$$\begin{aligned}
&\text{a)}\quad u_I \in C^1([0,T]; S_h),\\
&\text{b)}\quad \Big(\frac{\partial u_I(t)}{\partial t}, \varphi_h\Big)_I + b_I(u_I(t),\varphi_h) + a_I(u_I(t),\varphi_h) + \varepsilon J_I^\sigma(u_I(t),\varphi_h) = \ell_I(\varphi_h)\,(t)\\
&\qquad\qquad \forall\, \varphi_h \in S_h,\ \forall\, t \in (0,T),\\
&\text{c)}\quad u_I(0) = u_I^0,
\end{aligned} \tag{2.141}$$

where the function $u_I^0 \in S_h$ is defined by

$$(u_I^0 - u^0, \varphi_h)_I = 0 \quad \forall \varphi_h \in S_h. \tag{2.142}$$

In a similar way as in [33] it is possible to show that under some additional assumptions on the regularity of the exact solution and of data, the rate of convergence of the method with numerical integration is the same as in the case of exact evaluation of integrals. We have

Theorem 2.3.2 *Let assumptions* (H), (A) *and* (B1) - (B3) *be satisfied and* (2.93) *hold. Let u be the exact solution of problem* (2.86) *satisfying the regularity condition* (2.87) *and $f_\ell(u) \in L^2(0,T; W^{p+2,\infty}(\Omega))$, $\ell = 1, 2$, and let the solution of the dual problem* (2.111) *satisfy* (2.113). *Moreover, let u_I be the approximate solution obtained by scheme* (2.141) *with numerical integration. Let*

$g \in L^2(0,T;H^{p+1}(\Omega))$, $g_N \in L^2(0,T;H^{p+2}(\Gamma_N))$, $u_D \in L^2(0,T;H^{p+3}(\Gamma_D))$ and $u^0 \in H^{p+1}(\Omega)$. Then the error $e_I = u - u_I$ satisfies the estimate

$$\|e_I\|_{L^\infty(0,T;L^2(\Omega))} \le Ch^{p+1}, \tag{2.143}$$

with a constant $C > 0$ depending on u, g, g_N, u_D, T, ε and h_0, but independent of h.

Let us note that the Gauss quadrature formulae defined in Section 1.5.2 satisfy assumptions (B1)-(B3), if we take $p = 2$. The preceding theorem states that we can use these formulae to carry out the numerical experiments in Section 1.6, since we use piecewise linear and quadratic elements.

Conclusion

In this section we have derived optimal error estimates in the $L^\infty(0,T;L^2(\Omega))$-norm of the *symmetric interior penalty* (SIPG) discontinuous Galerkin space semidiscretization of a nonstationary convection-diffusion problem. There are several open problems connected with the analysis of optimal error estimates of the DGFEM for convection-diffusion problems:

- Derivation of optimal error estimates in the case of a weaker regularity of the exact solution of the considered convection-diffusion problem. This would be connected with the error analysis for solutions, which are elements of the Sobolev-Slobodetskii spaces of functions with "non-integer derivatives".
- Derivation of optimal error estimates in the case of a weaker regularity of the solution of the dual problem. Namely, we are interested in the case of a polygonal **nonconvex** domain Ω and/or $\Gamma_N \neq \emptyset$.
- An extension of optimal error estimates to nonstationary problems with nonlinear convection as well as diffusion.
- What can one say if the nonsymmetric or incomplete variants of the diffusion terms are applied? We cannot use the presented technique – does this mean that the $L^\infty(L^2)$-norm of the error is suboptimal in these cases? Numerical experiments conducted in Section 1.6 indicate that this may not be the case.

Chapter 3

Discontinuous Galerkin method for the Euler equations

In this chapter we shall be concerned with the discontinuous Galerkin finite element method applied to the solution of inviscid compressible flows. The discretization of the Euler equations is described along with important topics - the numerical flux, boundary conditions, shack-capturing and semi-implicit time discretization. In the last section, numerical experiments are presented.

3.1 System of Euler equations

We shall be concerned with inviscid compressible two-dimensional flow. Let $T > 0$, $\Omega \subset \mathbb{R}^2$ and Q_T be the same as in Section 1.1. Furthermore, we define disjoint boundary components $\Gamma_I, \Gamma_O, \Gamma_W$, the *inlet*, *outlet* and *impermeable wall* respectively, such that $\partial\Omega = \Gamma_I \cup \Gamma_O \cup \Gamma_W$. We also define $\Gamma_{IO} = \Gamma_I \cup \Gamma_O$. The system of Euler equations describing 2D inviscid compressible flow can be written in the form of a conservation law for the *state vector* $\mathbf{w}(x,t)$:

$$\frac{\partial \mathbf{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\mathbf{w})}{\partial x_s} = 0 \quad \text{in } Q_T, \tag{3.1}$$

where $\boldsymbol{f}_s, s = 1, 2$, are the *inviscid fluxes* and

$$\begin{aligned} \mathbf{w} &= (\rho, \rho v_1, \rho v_2, e)^{\mathrm{T}} \in \mathbb{R}^4, \\ \boldsymbol{f}_i(w) &= (f_{i1}(\mathbf{w}), \dots, f_{i4}(\mathbf{w}))^{\mathrm{T}} \\ &= (\rho v_i, \rho v_1 v_i + \delta_{1i} p, \rho v_2 v_i + \delta_{2i} p, (e+p) v_i)^{\mathrm{T}}. \end{aligned} \tag{3.2}$$

Here the following notation has been used: ρ - density, p - pressure, $\boldsymbol{v} = (v_1, v_2)$ - velocity, $\rho\boldsymbol{v} = (\rho v_1, \rho v_2)$ - momentum, e - total energy. The system in this form is not complete (four equations for five variables) and we need to add the following

relation derived from the equation of state:

$$p = (\gamma - 1)(e - \rho|\boldsymbol{v}|^2/2), \tag{3.3}$$

where $\gamma = c_p/c_v > 1$ is the Poisson adiabatic constant. For example, for air, $\gamma = 1.4$. System (3.1) is time-dependant, thus we prescribe the initial condition

$$\mathbf{w}(x, 0) = \mathbf{w}^0(x), \quad x \in \Omega, \tag{3.4}$$

and of course, we also need convenient boundary conditions - this subject, however, will be treated in Section 3.4. System (3.1) and (3.3) represents the conservation of mass, momentum and energy of a perfect compressible gas.

In the following, we will need a property of the fluxes $\boldsymbol{f}_s$ called *homogeneity*:

$$\boldsymbol{f}_s(\alpha\mathbf{w}) = \alpha\boldsymbol{f}_s(\mathbf{w}), \quad \alpha \in \mathbb{R},\ \alpha \neq 0,\ s = 1, 2. \tag{3.5}$$

This property implies the useful relation

$$\boldsymbol{f}_s(\mathbf{w}) = \mathbb{A}_s(\mathbf{w})\mathbf{w}, \quad \text{where } \mathbb{A}_s(\mathbf{w}) = \frac{D\boldsymbol{f}_s(\mathbf{w})}{D\mathbf{w}},\ s = 1, 2. \tag{3.6}$$

We define the *speed of sound* and the *Mach number*:

$$a = \sqrt{\gamma p/\varrho}, \quad M = \frac{|\mathbf{v}|}{a}. \tag{3.7}$$

The speed of sound is the velocity of the propagation of perturbations in density and pressure. In other words, the speed of sound is the highest speed that "information" travels in a compressible fluid. Further we define the *entropy*

$$S = c_v \ln \frac{p}{\rho^\gamma}. \tag{3.8}$$

Assume $\mathbf{n} = (n_1, n_2)^T \in I\!R^2$ with $|\mathbf{n}| = 1$. Let

$$\begin{aligned} \mathbb{A}_s(\mathbf{w}) &:= \frac{D\boldsymbol{f}_s}{D\mathbf{w}}, \quad s = 1, 2, \\ \mathbb{P}(\mathbf{w}, \mathbf{n}) &:= \sum_{s=1}^{2} \mathbb{A}_s(\mathbf{w}) n_s, \\ \mathcal{P}(\mathbf{w}, \mathbf{n}) &:= \sum_{s=1}^{2} f_s(\mathbf{w}) n_s, \end{aligned} \tag{3.9}$$

then we have the following result [21]:

Lemma 3.1.1 (Diagonal hyperbolicity) *Let $\mathbf{n} = (n_1, n_2)^T \in I\!R^2$ with $|\mathbf{n}| = 1$. Then the matrix $\mathbb{P}(\mathbf{w}, \mathbf{n})$ is diagonilazable with real eigenvalues, i.e. there exists a matrix $\mathbb{T} \in I\!R^{4,4}$ and $\lambda_1, \ldots, \lambda_4 \in I\!R$ such that*

$$\mathbb{P}(\mathbf{w}, \mathbf{n}) = \mathbb{T}\mathbb{D}\mathbb{T}^{-1}, \quad \mathbb{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_4). \tag{3.10}$$

3.2 Discretization

We proceed similarly as in the scalar case. We seek the approximate solution $\mathbf{w}_h$ in the finite dimensional space of vector valued piecewise polynomial functions $\mathbf{S}_h = [S_h]^4$, where S_h is the space from Section 1.2. We multiply (3.1) by a test function $\boldsymbol{\varphi} \in [H^1(\Omega, \mathcal{T}_h)]^4$ and integrate over $K_i \in \mathcal{T}_h$. With the aid of Green's theorem and summing over all $i \in I$, we obtain

$$\begin{aligned}\frac{d}{dt}\sum_{K_i\in\mathcal{T}_h}\int_{K_i}\mathbf{w}\cdot\boldsymbol{\varphi}\,dx &= \sum_{K_i\in\mathcal{T}_h}\int_{K_i}\sum_{s=1}^{2}\boldsymbol{f}_s(\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx \\ &\quad - \sum_{K_i\in\mathcal{T}_h}\sum_{j\in S(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\boldsymbol{f}_s(\mathbf{w})n_{ij}^{(s)}\cdot\boldsymbol{\varphi}\,dS,\end{aligned} \tag{3.11}$$

where $\mathbf{n}_{ij} = (n_{ij}^{(1)}, n_{ij}^{(2)})$ is the outer unit normal to ∂K_i on Γ_{ij}. In the second right-hand side term, we use the approximation

$$\int_{\Gamma_{ij}}\sum_{s=1}^{2}\boldsymbol{f}_s(\mathbf{w})n_s^{ij}\cdot\boldsymbol{\varphi}\,dS \approx \int_{\Gamma_{ij}}\mathbf{H}(\mathbf{w}|_{\Gamma_{ij}}, \mathbf{w}|_{\Gamma_{ji}}, \mathbf{n}_{ij})\cdot\boldsymbol{\varphi}\,dS, \tag{3.12}$$

incorporating a numerical flux $\mathbf{H}$, discussed in Section 3.3. Now we introduce the form

$$\begin{aligned}b_h(\mathbf{w},\boldsymbol{\varphi}) &= -\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\boldsymbol{f}_s(\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx \\ &\quad + \sum_{i\in I}\sum_{j\in S(i)}\int_{\Gamma_{ij}}\mathbf{H}(\mathbf{w}|_{\Gamma_{ij}}, \mathbf{w}|_{\Gamma_{ji}}, \mathbf{n}_{ij})\cdot\boldsymbol{\varphi}\,dS,\end{aligned} \tag{3.13}$$

where $\mathbf{w}|_{\Gamma_{ji}}$ for $\Gamma_{ij} \subset \partial\Omega$ is defined in Section 3.4 using appropriate boundary conditions. Finally we define the discrete problem:

Definition 3.2.1 *We say that* $\mathbf{w}_h$ *is the* approximate solution *of problem (3.1), if*

$$\begin{aligned}&a)\ \mathbf{w}_h \in C^1([0,T];\mathbf{S}_h),\\ &b)\ \frac{d}{dt}(\mathbf{w}_h(t),\boldsymbol{\varphi}_h) + b_h(\mathbf{w}_h(t),\boldsymbol{\varphi}_h) = 0, \quad \forall\boldsymbol{\varphi}_h \in \mathbf{S}_h,\ \forall t\in(0,T),\\ &c)\ \mathbf{w}_h(0) = \mathbf{w}_h^0,\end{aligned} \tag{3.14}$$

where $\mathbf{w}_h^0$ *is an* $\mathbf{S}_h$ *approximation of the initial condition* $\mathbf{w}^0$.

3.3 Numerical fluxes

In the implementation of the methods presented here, we use the following numerical fluxes. Details concerning these numerical fluxes can be found in [21]. These numerical fluxes have a convenient form for the semi-implicit linearization with respect to time presented in Section 3.6.2. Particularly, they can all be written in the form

$$\mathbf{H}(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = \mathbb{A}_L(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n})\mathbf{w}_L + \mathbb{A}_R(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n})\mathbf{w}_R \tag{3.15}$$

with some matrices $\mathbb{A}_L, \mathbb{A}_R : I\!R^4 \times I\!R^4 \times I\!R^2 \to I\!R^{4,4}$.

In the following we define four numerical fluxes: the numerical flux of Vijayasundaram, Steger-Warming, Lax-Friedrichs and Roe. In numerical tests, the Steger-Warming numerical flux exhibited instabilities, unlike the remaining three. The Vijayasundaram and Roe numerical fluxes provided nearly identical results, while solutions computed using the Lax-Friedrichs numerical flux contained small stationary oscillations near curved boundaries on coarse meshes. Therefore all numerical experiments included in Section 3.8 and 4.4 are computed using the Vijayasundaram numerical flux.

3.3.1 Vijayasundaram numerical flux $\mathbf{H}_{VS}$

This numerical flux is based on the flux vector splitting concept, and can be viewed as an extension of the flux from Section 1.5.1. We use Lemma 3.1.1 and define the "absolute value", "positive" and "negative" parts of matrix $\mathbb{P}$:

$$\begin{aligned} |\mathbb{P}|(\mathbf{w}, \mathbf{n}) &= \mathbb{T}|\mathbb{D}|\mathbb{T}^{-1}, \quad |\mathbb{D}| = \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_4|), \\ \mathbb{P}^{\pm}(\mathbf{w}, \mathbf{n}) &= \mathbb{T}\mathbb{D}^{\pm}\mathbb{T}^{-1}, \quad \mathbb{D}^{\pm} = \mathrm{diag}(\lambda_1^{\pm}, \ldots, \lambda_4^{\pm}) \end{aligned} \tag{3.16}$$

and define the Vijayasundaram numerical flux $\mathbf{H}_{VS}$:

$$\mathbf{H}_{VS}(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = \mathbb{P}^{+}\left(\frac{\mathbf{w}_L + \mathbf{w}_R}{2}, \mathbf{n}\right)\mathbf{w}_L + \mathbb{P}^{-}\left(\frac{\mathbf{w}_L + \mathbf{w}_R}{2}, \mathbf{n}\right)\mathbf{w}_R. \tag{3.17}$$

Explicit formulae for $\mathbb{T}, \mathbb{D}, \mathbb{T}^{-1}$ can be found in [21] or [20]. The eigenvalues λ_i have the form

$$\begin{aligned} \lambda_1 &= \lambda_2 - a, \\ \lambda_2 &= \lambda_3 = n_1 v_1 + n_2 v_2, \\ \lambda_4 &= \lambda_1 + a, \end{aligned} \tag{3.18}$$

where $a = \sqrt{\gamma p/\varrho}$ is the speed of sound.

3.3.2 Steger-Warming numerical flux $\mathbf{H}_{SW}$

This numerical flux is similar to the Vijayasundaram numerical flux, which can be viewed as a central form of the Steger-Warming numerical flux. We define $\mathbf{H}_{SW}$:

$$\mathbf{H}_{SW}(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = \mathbb{P}^+ (\mathbf{w}_L, \mathbf{n})\, \mathbf{w}_L + \mathbb{P}^- (\mathbf{w}_R, \mathbf{n})\, \mathbf{w}_R. \tag{3.19}$$

3.3.3 Lax-Friedrichs numerical flux $\mathbf{H}_{LF}$

According to (3.9), the Lax-Friedrichs numerical flux is defined as

$$\mathbf{H}_{LF}(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = \frac{1}{2}\big(\mathcal{P} (\mathbf{w}_L, \mathbf{n}) + \mathcal{P} (\mathbf{w}_R, \mathbf{n}) + \alpha(\mathbf{w}_L, \mathbf{w}_R)(\mathbf{w}_L - \mathbf{w}_R)\big), \tag{3.20}$$

where

$$\alpha(\mathbf{w}_L, \mathbf{w}_R) = \max_{\mathbf{w}=\mathbf{w}_L,\mathbf{w}_R} \big\{|\lambda_{max}(\mathbf{w})|\big\}, \tag{3.21}$$

where $\lambda_{max}(\mathbf{w})$ is the largest (in absolute value) eigenvalue of $\mathbb{P}(\mathbf{w}, \mathbf{n})$.

3.3.4 Roe numerical flux $\mathbf{H}_{Roe}$

Roe's numerical flux introduced in [31] is defined, using notation from (3.16), as

$$\mathbf{H}_{Roe}(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = \frac{1}{2}\big(\mathcal{P} (\mathbf{w}_L, \mathbf{n}) + \mathcal{P} (\mathbf{w}_R, \mathbf{n}) + |\mathbb{P}| (\widehat{\mathbf{w}}, \mathbf{n}) (\mathbf{w}_L - \mathbf{w}_R)\big), \tag{3.22}$$

where the state $\widehat{\mathbf{w}}$ is the so called Roe average state defined by the following relations:

$$\begin{aligned}
\sqrt{\widehat{\rho}} &= \frac{1}{2}(\sqrt{\rho_L} + \sqrt{\rho_L}),\\
\widehat{u} &= \frac{\sqrt{\rho_L}\sqrt{u_L} + \sqrt{\rho_R}\sqrt{u_R}}{\sqrt{\rho_L} + \sqrt{\rho_R}},\\
\widehat{v} &= \frac{\sqrt{\rho_L}\sqrt{v_L} + \sqrt{\rho_R}\sqrt{v_R}}{\sqrt{\rho_L} + \sqrt{\rho_R}},\\
\widehat{H} &= \frac{\sqrt{\rho_L}\sqrt{H_L} + \sqrt{\rho_R}\sqrt{H_R}}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \quad \text{where } H = \frac{E+p}{\rho}.
\end{aligned} \tag{3.23}$$

3.4 Boundary conditions

The choice of appropriate boundary conditions is a delicate problem which plays a key role in the presented algorithms. Boundary conditions are incorporated into the DGFEM via the choice of $H(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n})$ or $\mathbf{w}_R = \mathbf{w}|_{\Gamma_{ji}}$ for boundary edges.

Boundary	Character	Extrapolated	Prescribed
INLET	supersonic	—	ρ, v_1, v_2, p
	subsonic	p	ρ, v_1, v_2
OUTLET	supersonic	ρ, v_1, v_2, p	—
	subsonic	ρ, v_1, v_2	p

Table 3.1: Boundary conditions for 2D flow.

3.4.1 Solid impermeable wall

For $\Gamma \subset \Gamma_W$ we prescribe the so-called *no-stick* condition: $\mathbf{v} \cdot \mathbf{n} = 0$ on Γ. Taking this into account, the normal component of the inviscid flux has the form

$$\sum_{s=1}^{2} \boldsymbol{f}_s(\mathbf{w}) n_s = (\mathbf{v} \cdot \mathbf{n})\mathbf{w} + p(0, n_1, n_2, \mathbf{v} \cdot \mathbf{n})^{\mathrm{T}} = p(0, n_1, n_2, 0)^{\mathrm{T}}. \tag{3.24}$$

If we extrapolate the value of pressure by $p_R := p_L$, we can define the numerical flux

$$H(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = p(0, n_1, n_2, 0)^{\mathrm{T}}. \tag{3.25}$$

3.4.2 Inlet/outlet conditions

In the case of inlet and outlet conditions the problem is, which quantities should be prescribed (*Dirichlet* condition) and which should be extrapolated onto Γ from the adjacent element (*Neumann*-type condition). One possibility used often in practice is given in [21] and [20] using the method of characteristics. We shall discuss the derivation of these boundary conditions in Section 3.4.3, where new inlet and outlet conditions are presented. For 2D flow the conditions are given in Table 3.1.

Numerical results conducted on the GAMM channel (channel with 10% circular bump, Mach number at inlet =0.67) show that the inlet and outlet boundary conditions tend to reflect incoming density waves that are a result of starting the computation with a nonphysical initial condition. This effect is more pronounced at the inlet, which causes a cascade of reflecting and interfering of these waves that produce undesired "noisy" perturbations of the solution, since the solid walls also reflect density waves. Consequently, it takes a long time for the solution to stabilize as $t \to \infty$. Nonetheless in the common test case, when $M = 0.67$, the oscillations eventually diminish and the solution becomes stationary. However, when we lower the Mach number, the undesired density oscillations are of the same magnitude as the variations of ρ in the steady-state solution and obtaining a stationary state is difficult.

3.4.3 Characteristic-based transparent boundary conditions

In this section we present alternative boundary conditions derived in [27].

Let $\Gamma = \Gamma_{ij} \subset \Gamma_{IO}$ and $\mathbf{n} = \mathbf{n}_{ij}$ be the outer unit normal to K_i on Γ. In order to compute $\mathbf{H}(\mathbf{w}_i, \mathbf{w}_j, \mathbf{n})$, we need to specify the value $\mathbf{w}_j$, when $\mathbf{w}_i$ is known.

Let $\mathbf{n} = \mathbf{n}_{ij}$ be the outer unit normal to K_i on $\Gamma = \Gamma_{ij}$. Let us introduce a new Cartesian coordinate system $\tilde{x}_1, \tilde{x}_2$ in $\mathbb{R}^2$ with the origin at the center of gravity of edge Γ, the coordinate $\tilde{x}_1$ is oriented in the direction of the normal $\mathbf{n}$ and $\tilde{x}_2$ tangent to Γ. The Euler equations transformed into this new coordinate system have the form

$$\frac{\partial \mathbf{q}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\mathbf{q})}{\partial \tilde{x}_s} = 0, \tag{3.26}$$

as follows from the rotational invariance of the Euler equations. Here

$$\mathbf{q} = \mathbb{Q}(\mathbf{n})\mathbf{w} \tag{3.27}$$

where

$$\mathbb{Q}(\mathbf{n}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & n_1 & n_2 & 0 \\ 0 & -n_2 & n_1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.28}$$

Now we neglect the tangential derivative $\partial/\partial\tilde{x}_2$ and get the system with one space variable $\tilde{x}_1$ in the form

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \boldsymbol{f}_1(\mathbf{q})}{\partial \tilde{x}_1} = 0. \tag{3.29}$$

Using the *homogeneity* of the fluxes (3.6) we write system (3.29) in the nonconservative form

$$\frac{\partial \mathbf{q}}{\partial t} + \mathbb{A}_1(\mathbf{q}) \frac{\partial \mathbf{q}}{\partial \tilde{x}_1} = 0. \tag{3.30}$$

Finally we linearize this system around the state $\mathbf{q}_i = \mathbb{Q}(\mathbf{n})\mathbf{w}_i$ and obtain the linear system

$$\frac{\partial \mathbf{q}}{\partial t} + \mathbb{A}_1(\mathbf{q}_i) \frac{\partial \mathbf{q}}{\partial \tilde{x}_1} = 0, \tag{3.31}$$

which will be considered in the set $(-\infty, 0) \times (0, \infty)$ and equipped with the initial condition

$$\mathbf{q}(\tilde{x}_1, 0) = \mathbf{q}_i, \quad \tilde{x}_1 \in (-\infty, 0) \tag{3.32}$$

and the boundary condition

$$\mathbf{q}(0, t) = \mathbf{q}_j, \quad t > 0. \tag{3.33}$$

The goal is to choose $\mathbf{q}_j$ in such a way that the initial-boundary problem (3.31)–(3.33) is well posed, i.e. has a unique solution. The solution can be written in the form

$$\mathbf{q}(\tilde{x}_1, t) = \sum_{s=1}^{4} \mu(\bar{x}_1, t)\mathbf{r}_s, \tag{3.34}$$

where $\mathbf{r}_s = \mathbf{r}_s(\mathbf{q}_i)$ are the eigenvectors of the matrix $\mathbb{A}_1(\mathbf{q}_i)$ corresponding to its eigenvalues $\tilde{\lambda}_s = \tilde{\lambda}_s(\mathbf{q}_i)$ and creating a basis in $\mathbb{R}^4$. Moreover,

$$\mathbf{q}_i = \sum_{s=1}^{4} \alpha_s \mathbf{r}_s, \quad \mathbf{q}_j = \sum_{s=1}^{4} \beta_s \mathbf{r}_s. \tag{3.35}$$

Substituting (3.34) into (3.31) and using the relation $\mathbb{A}_1(\mathbf{q}_i)\mathbf{r}_s = \tilde{\lambda}_s \mathbf{r}_s$, we find that problem (3.31)–(3.33) is equivalent to 4 mutually independent linear initial-boundary value scalar problems for $s = 1, \ldots, 4$:

$$\begin{aligned} &\frac{\partial \mu_s}{\partial t} + \tilde{\lambda}_s \frac{\partial \mu_s}{\partial \tilde{x}_1} = 0 \quad \text{in}(-\infty, 0) \times (0, \infty), \\ &\mu_s(\tilde{x}_1, 0) = \alpha_s, \quad \tilde{x}_1 \in (-\infty, 0), \\ &\mu_s(0, t) = \beta_s, \quad t \in (0, \infty), \end{aligned} \tag{3.36}$$

which can be solved by the method of characteristics. The solution is

$$\mu_s(\tilde{x}_1, t) = \begin{cases} \alpha_s, & \tilde{x}_1 - \tilde{\lambda}_s t < 0, \\ \beta_s, & \tilde{x}_1 - \tilde{\lambda}_s t > 0. \end{cases} \tag{3.37}$$

The conclusion is that if

a) $\tilde{\lambda}_s > 0$, then $\beta_s = \alpha_s$ (β_s is not prescribed, but it is obtained by the extrapolation of μ_s to the boundary $\tilde{x}_1 = 0$),

b) if $\tilde{\lambda}_s = 0$, then β_s is not prescribed (but can be defined as $\beta_s = \alpha_s$ by the continuous extension of μ_s to the boundary $\tilde{x}_1 = 0$),

c) if $\tilde{\lambda}_s < 0$, then β_s must be prescribed.

Furthermore, we incorporate the fact that

$$\tilde{\lambda}_s(\mathbf{q}_i) = \lambda_s(\mathbf{w}_i, \mathbf{n}), \quad s = 1, \ldots, 4, \tag{3.38}$$

where $\lambda_s(\mathbf{w}_i, \mathbf{n})$ are the eigenvalues of the Jacobi matrix $\mathbb{P}(\mathbf{w}_i, \mathbf{n})$ defined in (3.9). In [21] the conclusion is drawn, that we prescribe n_{pr} quantities characterizing $\mathbf{w}$, where n_{pr} is the number of negative eigenvalues λ_s, and extrapolate $n_{ex} = 4 - n_{pr}$ quantities. One choice of these quantities is given in Table 3.1. We propose to prescribed variables based on the local linearized problem.

We shall take some state $\mathbf{q}_j^0 = \mathbb{Q}(\mathbf{n})\mathbf{w}_j^0$. The state $\mathbf{w}_j^0$ is the state vector of the far-field flow, or the incoming fluid at the inlet, or the initial condition, depending on the situation and interpretation. With this state we repeat the

presented derivation. We calculate the eigenvectors $\mathbf{r}_s, s = 1, \ldots, 4$ from the state $\mathbf{q}_i$ and α_s, β_s such that

$$\mathbf{q}_i = \sum_{s=1}^{4} \alpha_s \mathbf{r}_s, \quad \mathbf{q}_j^0 = \sum_{s=1}^{4} \beta_s \mathbf{r}_s. \tag{3.39}$$

This is simple, since we have explicit formulae for the matrix $\mathbb{T}$, which has $\mathbf{r}_s$ for its columns, and the inverse $\mathbb{T}^{-1}$, as given in [21] or [20]. This is the same matrix that is used in the evaluation of the Vijayasundaram numerical flux in Section 3.3.1. We can thus see that for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)^{\mathrm{T}}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)^{\mathrm{T}}$ we have

$$\begin{aligned} \mathbf{q}_i = \mathbb{T}\boldsymbol{\alpha} \;&\Rightarrow\; \boldsymbol{\alpha} = \mathbb{T}^{-1}\mathbf{q}_i, \\ \mathbf{q}_j^0 = \mathbb{T}\boldsymbol{\beta} \;&\Rightarrow\; \boldsymbol{\beta} = \mathbb{T}^{-1}\mathbf{q}_j^0. \end{aligned} \tag{3.40}$$

Now we calculate the state $\mathbf{q}_j$ according to the presented process:

$$\mathbf{q}_j := \sum_{s=1}^{4} \gamma_s \mathbf{r}_s = \mathbb{T}\boldsymbol{\gamma}, \tag{3.41}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_4)^{\mathrm{T}}$ and

$$\gamma_s = \begin{cases} \alpha_s, & \lambda_s \geq 0, \\ \beta_s, & \lambda_s < 0. \end{cases} \tag{3.42}$$

Finally $\mathbf{w}_j = \mathbb{Q}^{-1}(\mathbf{n})\mathbf{q}_j$ and we can use this to calculate $\mathbf{H}(\mathbf{w}_i, \mathbf{w}_j, \mathbf{n})$.

In the framework of the presented theory, these boundary conditions seem to give the natural choice for $\mathbf{w}_j$. However, we must keep in mind two simplifications that we have made during the derivation:

a) we have neglected tangential derivatives of the solution in order to get a simplified equation (3.29),

b) we have avoided the nonlinearity of problem (3.29) by local linearization.

Nonetheless, experiments show that this method applied to the approximation of inlet and outlet boundary conditions lets density and pressure waves pass through the boundaries without reflection.

3.5 Approximation of the boundary

So far we have worked only with polygonal domains $\Omega \subset \mathbb{R}^2$. This is rather limiting, when we approach practical problems, in which we seldom meet completely polygonal (polyhedral) shapes. In practice, this means that we have a domain Ω with a curved boundary and have to approximate it with some Ω_h, which is polygonal. In the finite volume method this works well, since we seek piecewise

constant solutions. Also in the conforming finite element method with P^1 elements applied to elliptic or parabolic problems, polygonal approximations of the boundary yield optimal error estimates. However in the case of DGFE higher-order approximations, numerical experiments show, that this method does not give good results in the vicinity of curved parts of $\partial\Omega$. As stated in [21], refining the mesh locally does not help and undesired phenomena occur - for instance non-physical entropy production. In order to get good behavior near curved segments of the boundary when using higher orders discretizations, it is necessary to introduce a higher order approximation of the boundary $\partial\Omega$ and adjacent elements. This is discussed for the case of bilinear mappings of the reference element.

Isoparametric elements

Let $\Omega \subset \mathbb{R}^2$ and $\mathcal{T}_h$ be its partition formed by triangles $K_i, i \in I$. Let $\hat{K}$ be the reference triangle mentioned in Section 1.5.2. Let

$$\hat{P}^0 = (0;0), \quad \hat{P}^1 = (1;0), \quad \hat{P}^2 = (0;1) \tag{3.43}$$

be the vertices of $\hat{K}$ and

$$\hat{P}^{12} = (1/2;1/2). \tag{3.44}$$

Let $\{K_i, i \in I_c\}$ with $I_c \subset I$ be a set of triangles adjacent to a curved part of $\partial\Omega$. For $i \in I_c$ let $P_i^k, k = 0,1,2$, be the vertices of K_i such that $P_i^0 \in \Omega$, $P_i^1, P_i^2 \in \partial\Omega$. We suppose that the the center P_i^{12} of the curved side with endpoints P_i^1, P_i^2 is close to the center of the linear segment $P_i^1 P_i^2$ - this is natural for triangulations that are dense enough. Under these assumptions we can find a unique bilinear mapping F_i defined on $\hat{K}$, $F_i = (F_i^1, F_i^2)$ such that

$$\begin{aligned} F_i(\hat{P}^k) &= P_i^k, \quad k = 0,1,2, \\ F_i(\hat{P}^{12}) &= P_i^{12}. \end{aligned} \tag{3.45}$$

Triangles $K_i, i \in I_c$, are replaced by the curved triangles defined by

$$\bar{K}_i := F_i(\hat{K}), \tag{3.46}$$

which have two straight sides and one curved side approximating the curved segment of $\partial\Omega$ adjacent to K_i, Figure 3.1. If $i \notin I_c$ then F_i is a linear mapping and therefore $\bar{K}_i = K_i$.

In the described discretization we need to evaluate volume and boundary integrals over elements and their boundaries - here we describe the modification of the method for curved elements $\bar{K}_i, i \in I_c$ - the simpler case when $i \notin I_c$ is treated in the same manner, only the mapping F_i is linear. We denote by

$$J_{F_i}(\hat{x}) := \frac{DF_i}{D\hat{x}}(\hat{x}), \quad \hat{x} \in \hat{K}, \tag{3.47}$$

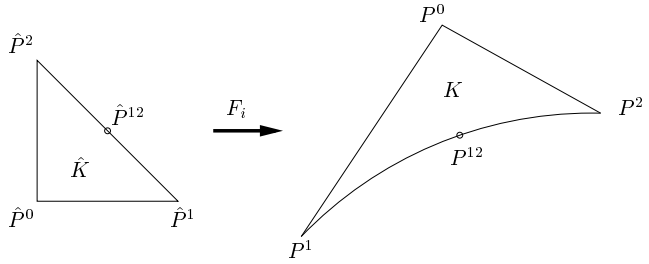


Figure 3.1: Bilinear mapping $F_i : \hat{K}_i \to K_i$

the Jacobi matrix of the mapping F_i. Test functions $\boldsymbol{\varphi}$ and the approximate solution $\mathbf{w}(\cdot, t)$ are defined on $\hat{K}_i$ as

$$\begin{aligned} \boldsymbol{\varphi}(x) &= \hat{\boldsymbol{\varphi}}(F_i^{-1}(x)), \quad x \in \bar{K}_i, \\ \mathbf{w}_h(x,t) &= \hat{\mathbf{w}}_i(F_i^{-1}(x), t), \quad x \in \bar{K}_i,\ t \in [0,T], \end{aligned} \tag{3.48}$$

where $\hat{\boldsymbol{\varphi}}, \hat{\mathbf{w}}(\cdot, t) \in [P^p(\hat{K})]^m$.

The forms in (3.14) are evaluated in the following way: The $L^2(K_i)$-scalar product is expressed, using the substitution as

$$\int_{\bar{K}_i} \mathbf{w}_h(x,t) \cdot \boldsymbol{\varphi}_h(x)\, dx = \int_{\hat{K}} \hat{\mathbf{w}}_i(\hat{x}, t) \cdot \hat{\boldsymbol{\varphi}}_h(\hat{x}) \det J_{F_i}(\hat{x})\, d\hat{x}, \quad i \in I. \tag{3.49}$$

In the inviscid volume terms in b_h we have to use the fact that

$$(\hat{\nabla}\hat{\boldsymbol{\varphi}}_h)(\hat{x}) = J_{F_i}(\hat{x})(\nabla \boldsymbol{\varphi}_h)(F_i(\hat{x})), \tag{3.50}$$

thus

$$(\nabla \boldsymbol{\varphi}_h)(F_i(\hat{x})) = [J_{F_i}(\hat{x})]^{-1}(\hat{\nabla}\hat{\boldsymbol{\varphi}}_h)(\hat{x}) \tag{3.51}$$

and

$$\begin{aligned} &\int_{\bar{K}_i} \sum_{s=1}^{2} \boldsymbol{f}_s(\mathbf{w}_h(x,t)) \cdot \frac{\partial \boldsymbol{\varphi}_h(x)}{\partial x_s}\, dx \\ &= \int_{\hat{K}} \big(\boldsymbol{f}_1(\hat{\mathbf{w}}_i), \boldsymbol{f}_2(\hat{\mathbf{w}}_i)\big)[J_{F_i}(\hat{x})]^{-1}\hat{\nabla}\hat{\boldsymbol{\varphi}}_h(\hat{x}) \det J_{F_i}(\hat{x})\, d\hat{x} \\ &= \int_{\hat{K}} \sum_{s=1}^{2} \boldsymbol{f}_s(\hat{\mathbf{w}}_i(x,t)) \cdot \sum_{j=1}^{2} \frac{\partial \hat{\boldsymbol{\varphi}}_h(x)}{\partial x_j} \frac{\partial (F_i^{-1})^j}{\partial x_s}(F_i(\hat{x}))\, \det J_{F_i}(\hat{x})\, d\hat{x}, \quad i \in I, \end{aligned} \tag{3.52}$$

where $(F_i^{-1})^j$ denotes the j-th component of the inverse mapping F_i^{-1}. However, the evaluation using the inverse $[J_{F_i}(\hat{x})]^{-1}$ is simpler then calculating the inverse

F_i^{-1} and than its Jacobi matrix. One can see that these two approaches are equivalent since

$$\frac{DF_i^{-1}}{Dx}(F_i(\hat{x})) = \left[\frac{DF_i}{D\hat{x}}(\hat{x})\right]^{-1} \tag{3.53}$$

following from the identity $x = F_i(F_i^{-1}(x))$.

Boundary integrals over a curved side $\Gamma_{ij} \subset \partial K_i$ in the boundary terms of the form b_h are computed with the aid of a suitable parameterization of Γ_{ij} and the side $\hat{\Gamma}$ of $\hat{K}$ corresponding to Γ_{ij} in the mapping F_i:

$$x = x(\xi) = F_i(\hat{x}(\xi)), \quad \xi \in [0,1]. \tag{3.54}$$

If we put

$$u(x) := \mathbf{H}(\mathbf{w}|_{\Gamma_{ij}}(x,t), \mathbf{w}|_{\Gamma_{ji}}(x,t), \mathbf{n}_{ij}) \cdot \boldsymbol{\varphi}_h(x) \tag{3.55}$$

(for a fixed t), we get

$$\begin{aligned} \int_{\Gamma_{ij}} u(x)\, dS &= \int_0^1 u(x(\xi))|x'(\xi)|\, d\xi = \\ &\int_0^1 u(F_i(\hat{x}(\xi)))\left\{\sum_{j=1}^{2}\left(\sum_{k=1}^{2}\frac{\partial F_i^j(\hat{x}(\xi))}{\partial \hat{x}_k}\hat{x}_k'(\xi)\right)^2\right\}^{1/2} d\xi. \end{aligned} \tag{3.56}$$

The parametrization $\hat{x} = \hat{x}(\xi)$ of $\hat{\Gamma}$ is expressed in the form

$$\hat{x}(\xi) = A + \xi(B - A), \tag{3.57}$$

where A, B are the endpoints of $\hat{\Gamma}$. The integrals over $\hat{K}$ and $\hat{\Gamma}$ are evaluated using the quadrature formulae given in Section 1.5.2.

3.6 Time discretization

As in the scalar case, the DGFE discretization (3.14) represents a system of ordinary differential equations. If we want to solve this system, we need to use a time discretization. We describe two possibilities: the *Euler forward method* and a *semi-implicit* linearization of the backward Euler scheme.

3.6.1 Explicit time discretization

For the forward Euler scheme we proceed as in Section 1.4. The scheme has the form

$$(\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \tau_k b_h(\mathbf{w}_h^k, \boldsymbol{\varphi}_h) = (\mathbf{w}_h^k, \boldsymbol{\varphi}_h), \quad \forall \varphi_h \in \mathbf{S}_h,\ k = 0, 1, \ldots. \tag{3.58}$$

Let $\mathcal{B} = \{\mathbf{w}_\alpha\}_{\alpha=1}^n$ be a basis in the space $\mathbf{S}_h$, where $n = \dim\mathbf{S}_h$. We seek the approximate solution $\mathbf{w}_h \in \mathbf{S}_h$ in the form

$$\mathbf{w}_h(t) = \sum_{\alpha=1}^{n} \xi_\alpha(t)\mathbf{w}_\alpha. \tag{3.59}$$

Due to the linearity of the form b_h in the variable $\boldsymbol{\varphi}$, we can use, as test functions only elements of the basis $\mathcal{B}$.

Let $0 = t_0 < t_1 < \dots$ be a partition of the time interval $[0, T]$, and $\tau_k = t_{k+1} - t_k$. Let $\boldsymbol{\xi}^k = (\xi_1^k, \dots, \xi_n^k)$, where ξ_α^k is an approximation of $\xi_\alpha(t_k)$. Then the Euler forward scheme can be written in the form of a system of n linear equations:

$$\mathbf{M}\boldsymbol{\xi}^{k+1} = \mathbf{M}\boldsymbol{\xi}^k + \tau_k \mathbf{a}(\boldsymbol{\xi}^k), \tag{3.60}$$

where $\mathbf{a}(\boldsymbol{\xi}^k)$ is a vector-valued mapping corresponding to the form b_h and $\mathbf{M} = \{m_{ij}\}_{i,j=1}^n$ is the $n \times n$ mass matrix with entries $m_{ij} = \int_\Omega \mathbf{w}_i \cdot \mathbf{w}_j \, dx$.

In the scalar case an appropriate choice of basis functions leads to the block-diagonality of $\mathbf{M}$. Here we must take into account that the basis functions are in the space $\mathbf{S}_h = [S_h]^4$. We use the P^1 and P^2 basis functions for S_h as in Section 1.5.3 'separately' for each component and get the basis for $[S_h]^4$. Then we write (3.59) as

$$\mathbf{w}_h^k(x) = \sum_{i\in I} \sum_{j=1}^{n_p} \sum_{l=1}^{4} \xi_{ijl}(t_k)\mathbf{w}_{ijl}(x), \tag{3.61}$$

where $\operatorname{supp} \mathbf{w}_{ijl} \subset K_i$, $n_p =$number of degrees of freedom for $P^p(K_i)$, $n_0 = 1, n_1 = 3, n_2 = 6$ and $\mathbf{w}_{ijl}^{k,(m)} = 0$ if $m \neq l$, where $\mathbf{u}^{(m)} = m$-th component of vector $\mathbf{u}$. Using this representation we are able to 'cluster' the basis functions with common support elements and representing the same unknown (i.e. with common nonzero component) thus achieving the block-diagonality of $\mathbf{M}$ - with $n_p \times n_p$ blocks. Here we can proceed as in Section 1.4. We need to calculate the inverse of the mass matrix, which is simple since $\mathbf{M}$ is block-diagonal. At the beginning of the calculation we can explicitly calculate the inversion of each block using a inversion algorithm based on Gaussian elimination. Let us note that in practice the order p of approximation can be chosen separately for every component of the state vector. Thus n_p becomes $n_{p(l)}$. This option was incorporated into the implementation of the presented scheme.

In order to guarantee the stability of scheme (3.58) we need to impose a limit on τ_k. In [16] the CFL condition in the case of P^1 approximations is proposed in the form

$$6\tau_k \max_{i\in I} \frac{1}{|K_i|} \left(\max_{j\in S(i)} d(\Gamma_{ij})\lambda_{ij}^{\max} \right) \leq CFL, \tag{3.62}$$

where $\lambda_{ij}^{\max} = \max_{x\in\Gamma_{ij}}(a(x) + |v(x)|)$, where $a(x)$ is the local speed of sound. CFL is a given constant ≤ 1, usually $CFL \approx 0.85$. This condition causes

problems if the Mach number is small, in this case the need arises for a semi-implicit or implicit scheme that would allow larger time steps.

3.6.2 Semi-implicit time discretization

We shall work with the DGFE discretization of the Euler equations as presented in Section 3.2. Thus, we seek a function $\mathbf{w}_h$ such that

$$\begin{aligned} &\text{a) } \mathbf{w}_h \in C^1([0,T];\mathbf{S}_h),\\ &\text{b) } \frac{d}{dt}(\mathbf{w}_h(t),\boldsymbol{\varphi}_h) + b_h(\mathbf{w}_h(t),\boldsymbol{\varphi}_h) = 0, \quad \forall \boldsymbol{\varphi}_h \in \mathbf{S}_h,\ \forall t \in (0,T),\\ &\text{c) } \mathbf{w}_h(0) = \tilde{\mathbf{w}}_h^0, \end{aligned} \tag{3.63}$$

where the inviscid form b_h is defined, using some numerical flux $\mathbf{H}$, as follows:

$$\begin{aligned} b_h(\mathbf{w},\boldsymbol{\varphi}) = &-\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\boldsymbol{f}_s(\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx\\ &+\sum_{i\in I}\sum_{j\in S(i)}\int_{\Gamma_{ij}}\mathbf{H}(\mathbf{w}|_{\Gamma_{ij}},\mathbf{w}|_{\Gamma_{ji}},\mathbf{n}_{ij})\cdot\boldsymbol{\varphi}\,dS, \end{aligned} \tag{3.64}$$

where $\mathbf{w}|_{\Gamma_{ji}}$ for $\Gamma_{ij} \subset \partial\Omega$ is defined in Section 3.4 using appropriate boundary conditions.

Relations (3.63) represent a system of ordinary differential equations, which must be, in practice, solved using an appropriate time discretization. In Section 3.6 an explicit forward Euler method is used. As stated earlier, we need a method which is not so limiting in terms of the time step τ_k. We therefore use the implicit *backward Euler method*.

Let $0 < t_0 < t_1 < \dots$ be a partition of the time interval $(0,T)$ and $\tau_k = t_{k+1} - t_k$. We seek $\mathbf{w}_h^k \approx \mathbf{w}_h(t_k)$ such that

$$\begin{aligned} &\text{a) } \mathbf{w}_h^{k+1} \in \mathbf{S}_h,\\ &\text{b) } \left(\frac{\mathbf{w}_h^{k+1}-\mathbf{w}_h^k}{\tau_k},\boldsymbol{\varphi}_h\right) + b_h(\mathbf{w}_h^{k+1},\boldsymbol{\varphi}_h) = 0\\ &\qquad\qquad \forall \boldsymbol{\varphi}_h \in \mathbf{S}_h,\ k = 0,1,\dots,\\ &\text{c) } \mathbf{w}_h^0 = \tilde{\mathbf{w}}_h^0. \end{aligned} \tag{3.65}$$

This scheme however leads to a large system of highly nonlinear equations due to the nonlinearity of the form b_h in the variable $\mathbf{w}_h^{k+1}$. The numerical solution of such a system is very complicated and time consuming, therefore in [16] a simplified linearization of problem 3.65 is presented in order to obtain a large (sparse) system of linear equations rather than solving the nonlinear system.

We shall treat the interior and boundary terms in (3.64) separately:

$$\begin{aligned} b_h(\mathbf{w}_h^{k+1},\boldsymbol{\varphi}_h) = &\underbrace{-\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\boldsymbol{f}_s(\mathbf{w}_h^{k+1})\cdot\frac{\partial\boldsymbol{\varphi}_h}{\partial x_s}\,dx}_{:=\widetilde{\sigma}_1} \\ &\underbrace{+\sum_{i\in I}\sum_{j\in S(i)}\int_{\Gamma_{ij}}\mathbf{H}(\mathbf{w}_h^{k+1}|_{\Gamma_{ij}},\mathbf{w}_h^{k+1}|_{\Gamma_{ji}},\mathbf{n}_{ij})\cdot\boldsymbol{\varphi}_h\,dS}_{:=\widetilde{\sigma}_2}. \end{aligned} \tag{3.66}$$

For $\widetilde{\sigma}_1$ we use the property of the Euler fluxes $\boldsymbol{f}_s$ given in (3.6). We set

$$\sigma_1 := \sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\mathbb{A}_s(\mathbf{w}_h^k)\,\mathbf{w}_h^{k+1}\cdot\frac{\partial\boldsymbol{\varphi}_h}{\partial x_s}\,dx. \tag{3.67}$$

In order to treat the term $\widetilde{\sigma}_2$ we must choose a numerical flux suitable for linearization. One possibility is the Vijayasundaram flux, as presented in Section 3.3.1, although any of the presented fluxes can be used, since they all have a similar form (3.15). The Vijayasundaram numerical flux is written in the form

$$\mathbf{H}_{VS}(\mathbf{w}_L,\mathbf{w}_R,\mathbf{n}) = \mathbb{P}^+\left(\frac{\mathbf{w}_L+\mathbf{w}_R}{2},\mathbf{n}\right)\mathbf{w}_L + \mathbb{P}^-\left(\frac{\mathbf{w}_L+\mathbf{w}_R}{2},\mathbf{n}\right)\mathbf{w}_R, \tag{3.68}$$

which is suitable for the linearization of the terms in $\widetilde{\sigma}_2$. For interior edges this reads:

$$\sum_{i\in I}\sum_{j\in s(i)}\int_{\Gamma_{ij}}\left[\,\mathbb{P}^+\big(\langle\mathbf{w}_h^k\rangle_{ij},\mathbf{n}_{ij}\big)\,\mathbf{w}_h^{k+1}|_{\Gamma_{ij}} + \mathbb{P}^-\big(\langle\mathbf{w}_h^k\rangle_{ij},\mathbf{n}_{ij}\big)\,\mathbf{w}_h^{k+1}|_{\Gamma_{ji}}\,\right]\cdot\boldsymbol{\varphi}_h\,dS.$$

For edges $\Gamma_{ij}\subset\Gamma_{IO}$ we cannot simply apply this linearization since we have no information about $\mathbf{w}_h^{k+1}|_{\Gamma_{ji}}$ – this is caused by the fact that the Inlet and Outlet are not a priori given and can change roles for complex flows. A simple solution is to treat these terms explicitly, i.e. $\mathbf{w}_h^{k+1}|_{\Gamma_{ji}}\approx\mathbf{w}_h^k|_{\Gamma_{ji}}$, where the latter state is calculated using a method from Section 3.4. In contrast to [16], we consider more suitable to use here $\mathbf{w}_{ij}^{k+1}$ (instead of $\mathbf{w}_{ij}^k$ from [16]). Thus, inlet and outlet terms have the form:

$$\sum_{i\in I}\sum_{j\in\gamma_{IO}(i)}\int_{\Gamma_{ij}}\left[\,\mathbb{P}^+\big(\langle\mathbf{w}_h^k\rangle_{ij},\mathbf{n}_{ij}\big)\,\mathbf{w}_h^{k+1}|_{\Gamma_{ij}} + \mathbb{P}^-\big(\langle\mathbf{w}_h^k\rangle_{ij},\mathbf{n}_{ij}\big)\,\mathbf{w}_h^{k}|_{\Gamma_{ji}}\,\right]\cdot\boldsymbol{\varphi}_h\,dS,$$

where $\gamma_{IO}(i)=\{j\in S(i);\Gamma_{ij}\subset\Gamma_{IO}\}$.

For $\Gamma_{ij}\subset\Gamma_W$ special treatment is needed – according to Section 3.4.1 we put

$$\begin{aligned} &\sum_{i\in I}\sum_{j\in\gamma_W(i)}\int_{\Gamma_{ij}}\mathbf{H}(\mathbf{w}_h^{k+1}|_{\Gamma_{ij}},\mathbf{w}_h^{k+1}|_{\Gamma_{ji}},\mathbf{n}_{ij})\cdot\boldsymbol{\varphi}_h\,dS \\ &\qquad\approx\sum_{i\in I}\sum_{j\in\gamma_W(i)}\int_{\Gamma_{ij}}\boldsymbol{F}_W(\mathbf{w}_h^{k+1},\mathbf{n}_{ij})\cdot\boldsymbol{\varphi}_h\,dS \end{aligned}$$

where $\boldsymbol{F}_W(\mathbf{w},\mathbf{n}) = (0, pn_1, pn_2, 0)^{\mathrm{T}}$ and $\gamma_W(i) = \{j \in S(i); \Gamma_{ij} \subset \Gamma_W\}$. To linearize these terms, we use the fact that $\boldsymbol{F}_W(\mathbf{w},\mathbf{n})$ is a homogeneous mapping with respect to $\mathbf{w}$. This is natural, since $\boldsymbol{F}_W(\mathbf{w},\mathbf{n})$ is derived from the homogeneous mapping $\sum_{s=1}^{2} \boldsymbol{f}_s n_s$ with the additional assumption that $\mathbf{v}\cdot\mathbf{n} = 0$. Similarly to (3.6), it follows that

$$\boldsymbol{F}_W(\mathbf{w}_h^{k+1}, \mathbf{n}_{ij}) = \frac{D\boldsymbol{F}_W}{D\mathbf{w}}(\mathbf{w}_h^{k+1}, \mathbf{n}_{ij})\, \mathbf{w}_h^{k+1}. \tag{3.69}$$

and we can use the semi-implicit approximation

$$\boldsymbol{F}_W(\mathbf{w}_h^{k+1}, \mathbf{n}_{ij}) \approx \frac{D\boldsymbol{F}_W}{D\mathbf{w}}(\mathbf{w}_h^{k}, \mathbf{n})\, \mathbf{w}_h^{k+1}. \tag{3.70}$$

Finally we can define the linearized edge terms as

$$\begin{aligned}
\sigma_2 \;:=\; & \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} \big[\, \mathbb{P}^+\big(\langle \mathbf{w}_h^k\rangle_{ij}, \mathbf{n}_{ij}\big)\, \mathbf{w}_h^{k+1}|_{\Gamma_{ij}} \\
& \qquad + \mathbb{P}^-\big(\langle \mathbf{w}_h^k\rangle_{ij}, \mathbf{n}_{ij}\big)\, \mathbf{w}_h^{k+1}|_{\Gamma_{ji}} \,\big] \cdot \boldsymbol{\varphi}_h \, dS. \\
& + \sum_{i\in I} \sum_{j\in \gamma_{IO}(i)} \int_{\Gamma_{ij}} \big[\, \mathbb{P}^+\big(\langle \mathbf{w}_h^k\rangle_{ij}, \mathbf{n}_{ij}\big)\, \mathbf{w}_h^{k+1}|_{\Gamma_{ij}} \\
& \qquad + \mathbb{P}^-\big(\langle \mathbf{w}_h^k\rangle_{ij}, \mathbf{n}_{ij}\big)\, \mathbf{w}_h^{k}|_{\Gamma_{ji}} \,\big] \cdot \boldsymbol{\varphi}_h \, dS, \\
& + \sum_{i\in I} \sum_{j\in \gamma_W(i)} \int_{\Gamma_{ij}} \frac{D\boldsymbol{F}_W}{D\mathbf{w}}(\mathbf{w}_h^k, \mathbf{n}_{ij})\, \mathbf{w}_h^{k+1} \cdot \boldsymbol{\varphi}_h \, dS.
\end{aligned} \tag{3.71}$$

Finally, we define the semi-implicitly linearized form as

$$b_h^{SI}(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = -\sigma_1 + \sigma_2, \tag{3.72}$$

where σ_1 and σ_2 are given in (3.67) and (3.71) respectively. We can now define the semi-implicit linearized scheme:

Definition 3.6.1 *For each $k = 0, 1, \ldots$ find $\mathbf{w}_h^{k+1}$ such that*

$$\begin{aligned}
& a)\ \mathbf{w}_h^{k+1} \in \mathbf{S}_h, \\
& b)\ (\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \tau_k b_h^{SI}(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = (\mathbf{w}_h^k, \boldsymbol{\varphi}_h) \\
& \qquad \forall \boldsymbol{\varphi}_h \in \mathbf{S}_h,\ k = 0, 1, \ldots, \\
& c)\ \mathbf{w}_h^0 = \tilde{\mathbf{w}}_h^0.
\end{aligned} \tag{3.73}$$

where $\mathbf{w}_h^0$ is an $\mathbf{S}_h$ approximation of the initial condition $\mathbf{w}^0$.

Matrix representation

The choice of appropriate basis functions for the space $\mathbf{S}_h$ as presented in Section 3.6 leads to the matrix representation of scheme (3.73):

$$\mathbf{A}(\boldsymbol{\xi}^k)\boldsymbol{\xi}^{k+1} = \mathbf{g}(\boldsymbol{\xi}^k), \tag{3.74}$$

where $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, is a $n \times n$ nonsymmetric matrices. This matrix has the form $\mathbb{A} = \mathbb{M} + \tau_k\mathbb{B}$, where $\mathbb{M}$ is the symmetric, positive definite block diagonal mass matrix and $\mathbb{B}$ represents the form b_h^{SI}. Since $\mathbb{M}$ is regular, we can expect that for small τ_k the matrix $\mathbb{A}$ is also regular. Furthermore, for sufficiently small τ_k the matrix will be close to the block diagonal matrix. One can therefore expect better behavior of the linear solver. On the other hand we want to avoid the limitations imposed on τ_k via the CFL-condition. We want to choose large τ_k, but we may expect slower convergence of an iterative linear solver.

Linear solver

System (3.74) needs to be efficiently solved on every time level. For this purpose we need a fast iterative or direct linear system solver. Since the resulting system is nonsymmetric, restarted preconditioned GMRES is used as an iterative solver. The block-Jacobi method is used as a left-hand preconditioner, where each block on the diagonal corresponds to all the variables on one element. The inverses of these blocks are explicitly calculated using an inversion method based on Gaussian elimination and stored on each time level. We restart GMRES after 10 iterations to avoid stagnation and permit up to 20 restarted GMRES cycles.

In situations where the iterative solver is not sufficient (e.g. low-Mach flows), we use the software package UMFPACK, which is an implementation of the Unsymmetric-pattern MultiFrontal method. This direct method uses graph algorithms to find a column pre-ordering which reduces fill-in during LU factorization [11], [12]. The method is suitable for sparse unsymmetric linear systems and the fact that the matrix $\mathbf{A}(\boldsymbol{\xi}^k)$ of system (3.74) has a symmetric nonzero structure may be used.

3.7 Shock capturing

From the theory of hyperbolic equations it is known that smooth initial conditions may lead in a finite time to solutions which have discontinuities. In the case of the Euler equations, one may expect for high speed flows the occurrence of so called shack waves and contact discontinuities. In this case the finite volume method works well, since it is only first order accurate in space and has a sufficient amount of numerical viscosity. However, when higher order schemes are used, the

so-called Gibbs phenomenon arises in the vicinity of discontinuities - oscillations, over- and under-shoots of the discrete solution. When using conforming finite elements, this behavior must be avoided by using additional stabilization techniques (streamline diffusion, Galerkin least squares, ...), otherwise these effects corrupt the solution. In the DGFEM, the situation is better, since these spurious over- and under-shoots remain localized in the vicinity of the discontinuity because we have relaxed inter-element continuity. However, additional shock capturing terms must be included in the scheme to avoid this phenomenon. In this section we present two possible techniques.

3.7.1 Limiting of the order of accuracy

In [21], the following approach is derived for the explicit time discretization. The idea is to preserve high order of accuracy in regions where the solution is regular and modify the scheme in a small neighbourhood of steep gradients and discontinuities. The local modification is based on the fact that first order methods (finite volume) have a sufficient amount of numerical viscosity to avoid the Gibbs phenomenon. Based on a discontinuity indicator, we locally project the higher order solution to a piecewise constant function in every time step, formally obtaining a first order scheme, where necessary.

Let us denote by u_h^k some scalar quantity characterizing the approximate solution $\mathbf{w}_h^k$. As follows from numerical experiments, in our case the density ρ is a good choice. Using the notation from Section 1.2, we define the jump function on ∂K_i as $[u_h^k]_{\partial K_i}(x) = [u_h^k]_{\Gamma_{ij}}(x)$ for $x \in \partial K_i \cap \Gamma_{ij}$. Numerical experiments show that interelement jumps are of the order $O(1)$ on discontinuities, but $O(h^2)$ in regions, where the solution is regular. On unstructured grids it is suitable to measure the interelement jumps in the integral form

$$\int_{\partial K_i} [u_h^k]^2 \, dS, \quad K_i \in \mathcal{T}_h. \tag{3.75}$$

In areas of regularity we have

$$g^1(i) = \int_{\partial K_i} [u_h^k]^2 \, dS/h^5 \approx \int_{\partial K_i} (O(h^2))^2 \, dS/h^5 \approx O(1), \tag{3.76}$$

whereas

$$g^2(i) = \int_{\partial K_i} [u_h^k]^2 \, dS/h \approx \int_{\partial K_i} (O(1))^2 \, dS/h \approx O(1), \tag{3.77}$$

for discontinuities or very steep gradients. These results lead to the idea that the switch between the higher and lower order scheme should be tested with the indicator

$$g^k(i) = \int_{\partial K_i} [u_h^k]^2 \, dS/h^\alpha, \quad K_i \in \mathcal{T}_h, \tag{3.78}$$

where $\alpha \in [0,5]$ - the natural choice being $\alpha = 5/2$. However on general unstructured grids it is suitable to define the indicator in terms of h_{K_i} and $|K_i|$. For $\alpha = 5/2$ we define the *discontinuity indicator*

$$g^k(i) = \int_{\partial K_i} [u_h^k]^2 \, dS/(h_{K_i}|K_i|^{3/4}), \quad K_i \in \mathcal{T}_h, \tag{3.79}$$

It was shown in [18] that the indicator $g^k(i)$ identifies discontinuities safely on unstructured and anisotropic meshes.

Now we can define an adaptive strategy for an automatic limiting of the order of accuracy of scheme (3.58):

$$\begin{aligned} &a)\ \mathbf{w}_h^{k+1} \in \mathbf{S}_h = \mathbf{S}_h^{p,-1}, \\ &b)\ (\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = (\tilde{\mathbf{w}}_h^k, \boldsymbol{\varphi}_h) - \tau_k b_h(\tilde{\mathbf{w}}_h^k, \boldsymbol{\varphi}_h), \quad \forall \varphi_h \in \mathbf{S}_h, \end{aligned} \tag{3.80}$$

where $\tilde{\mathbf{w}}_h^k$ is a modification of $\mathbf{w}_h^k$ defined in two steps:

$$\begin{aligned} &a)\ \text{Set } \tilde{\mathbf{w}}_h^k|_{K_i} := \mathbf{w}_h^k|_{K_i}, \quad \forall i \in I, \\ &b)\ \text{If } g^k(i) > 1 \text{ for some } i \in I, \text{ then } \tilde{\mathbf{w}}_h^k|_{K_i} := \Pi_h \mathbf{w}_h^k|_{K_i}, \end{aligned} \tag{3.81}$$

where Π_h is the $[L^2]^4$-projection operator, $\Pi_0 : [L^2]^4 \to [S^{0,-1}(\Omega, \mathcal{T}_h)]^4$ (= the space of piecewise constant vector functions) defined by (2.4). Therefore,

$$\pi_0 \mathbf{v}|_{K_i} = \int_{K_i} \mathbf{v} \, dx/|K_i|, \quad \forall i \in I. \tag{3.82}$$

Using this procedure, the solution is modified only on elements close to discontinuities, thus formally giving first order accuracy, but preserving a higher order elsewhere.

This approach gives good results in the explicit case, however it is not clear how to incorporate this technique into the semi-implicit scheme presented in Section 3.6.2. We propose a different approach presented in the following.

3.7.2 Artificial diffusion

The technique is motivated by the paper [26], on the basis of which the left-hand side of (3.63), b) is augmented by an artificial viscosity term of the form

$$\sum_{i \in I} h_{K_i} \int_{K_i} res_{K_i} \nabla \mathbf{w} \cdot \nabla \boldsymbol{\varphi} \, \mathrm{d}\boldsymbol{x}, \tag{3.83}$$

where res_{K_i} is some function of the residual of the equation, for instance the absolute value of the residual. The idea is that this quantity is small in regions where the solution is smooth and large near steep gradients and discontinuities, we therefore add artificial viscosity where it is needed.

However, since this form is nonzero also in regions, where the exact solution is regular, undesired effects, such as nonphysical entropy production, can appear in these regions. Therefore, we combine this technique with the approach described in the previous section. We use the discontinuity indicator $g^k(i)$ defined in (3.79) and introduce the discrete discontinuity indicator

$$G^k(i) = \begin{cases} 0 & \text{if } g^k(i) < 1, \\ 1 & \text{otherwise.} \end{cases} \tag{3.84}$$

To the left-hand side of (3.73), b) we add the artificial viscosity form

$$\Phi_h^1(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}) = \nu_1 \sum_{i \in I} h_{K_i} G^k(i) \int_{K_i} \nabla \mathbf{w}_h^{k+1} \cdot \nabla \boldsymbol{\varphi} \, \mathrm{d}\boldsymbol{x} \tag{3.85}$$

with $\nu_1 = O(1)$ a given constant. Numerical experiments show that this artificial viscosity form is rather local and does not behave well on locally refined grids. We therefore propose to augment the left-hand side of (3.73), b) the form

$$\Phi_h^2(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}) = \nu_2 \sum_{i \in I} \sum_{j \in s(i)} \frac{1}{2}\big(G^k(i) + G^k(j)\big) \int_{\Gamma_{ij}} [\mathbf{w}_h^{k+1}] \cdot [\boldsymbol{\varphi}] \, \mathrm{d}S, \tag{3.86}$$

where $\nu_2 = O(1)$, which allows to strengthen the influence of neighbouring elements and improves the behaviour of the method in the case, when strongly unstructured and/or anisotropic meshes are used.

Thus, the resulting scheme obtained from (3.73), b) reads:

$$\begin{aligned} &\text{a)} \quad \mathbf{w}_h^{k+1} \in \boldsymbol{S}_h, \\ &\text{b)} \quad \left(\frac{\mathbf{w}_h^{k+1} - \mathbf{w}_h^k}{\tau_k}, \boldsymbol{\varphi}_h \right)_h + b_h(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \Phi_h^1(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) \\ &\qquad + \Phi_h^2(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = 0, \ \forall \, \boldsymbol{\varphi}_h \in \boldsymbol{S}_h, \ k = 0, 1, \ldots, \\ &\text{c)} \quad \mathbf{w}_h^0 = \Pi_h \mathbf{w}^0. \end{aligned} \tag{3.87}$$

This method overcomes problems with the Gibbs phenomenon in the context of the semi-implicit scheme. It is important that $G^k(i)$ vanishes in regions, where the solution is regular. Therefore, the scheme does not produce nonphysical entropy in these regions. For examples see Figures 3.8, 3.9 and 3.20 in the following Section.

3.8 Numerical experiments

In this section we present the solution of some test problems in order to demonstrate the accuracy and robustness of the proposed method. The computational

grids were constructed with the aid of the anisotropic mesh adaptation technique [13]. In all examples quadratic elements ($r = 2$) were applied. Steady state solutions were obtained via time stabilization for "$t \to \infty$". This means that scheme (3.87) was used as an iterative process for "$k \to \infty$". This process was stopped, when the approximation of the time derivative satisfied the condition

$$\left\| \frac{\mathbf{w}_h^{k+1} - \mathbf{w}_h^k}{\tau_k} \right\|_{L^\infty(\Omega)} < 10^{-8}. \tag{3.88}$$

We have tried to collect examples, for which the analytical solution of the flow equations is known. Usually these are derived under certain constraints, namely, the flow is required to be incompressible and in some cases irrotational. In order to compare these exact incompressible solutions with the approximate solution of the compressible Euler equations, we take the Mach number to be small (usually $M = 10^{-4}$), since for $M \to 0$, the compressible Euler equations tend to the incompressible limit. In these cases, the maximum density variation is negligible in comparison with the transonic examples. This means that the computed low Mach number flow behaves as incompressible flow.

3.8.1 Irrotational flow past a Joukowski profile

1) Irrotational flow past a nonsymmetric Joukowski airfoil. First we consider flow past a negatively oriented Joukowski profile given by parameters $\Delta = 0.07, a = 0.5, h = 0.05$ (under the notation from [20], Section 2.2.68) with zero angle of attack. The far field quantities are constant, which implies that the flow is irrotational and homoentropic. Using the complex function method from [20], we can obtain the exact solution of incompressible inviscid irrotational flow satisfying the Kutta–Joukowski trailing condition, provided the velocity circulation around the profile, related to the magnitude of the far field velocity, $\gamma_{\text{ref}} = 0.7158$. We assume that the far field Mach number of compressible flow $M_\infty = 10^{-4}$. The computational domain is of the form of a square with side of the length equal to 10 chords of the profile. The mesh (in the whole computational domain) was formed by 5418 triangular elements and refined towards the profile. Figure 3.2 shows a detail near the profile of the velocity isolines for the exact solution of incompressible flow and for the approximate solution of compressible flow. In Figure 3.3, pressure isolines of incompressible and compressible flow are plotted. Figure 3.4 shows streamlines of the computed compressible flow. We see that the flow past the trailing edge is smooth. Further, in Figures 3.5 and 3.6, the velocity distribution and pressure coefficient distribution, respectively, past the profile is plotted in the direction from the leading edge to the trailing edge ($\circ\circ\circ$ – exact solution of incompressible flow, —— – approximate solution of compressible flow). The pressure coefficient was defined as $10^7 \cdot (p - p_\infty)$, where p_∞ denotes the far field pressure.

In the computed example the maximum density variation is $1.04 \cdot 10^{-8}$, which is in agreement with theoretical results (e. g. [29]), which state that the maximum density variation behaves as $O(M^2)$. The computed velocity circulation related to the magnitude of the far field velocity is $\gamma_{\text{refcomp}} = 0.7205$, which gives the relative error 0.66% with respect to the theoretical value γ_{ref} obtained for incompressible flow. The CFL number from the stability condition (3.62) was during the computational process successively increased from 1 (the start of the compuation) to $6 \cdot 10^6$.

In order to establish the quality of the computed pressure of the low Mach compressible flow in a quantitative way, we introduce the function

$$B = \frac{p}{\rho} + \frac{1}{2}|\boldsymbol{v}|^2, \tag{3.89}$$

which is constant for incompressible, inviscid, irrotational flow, as follows from the Bernoulli equation. In the considered compressible case, the relative variation of the function B, i.e. $(B_{\max} - B_{\min})/B_{\max} = 3.84 \cdot 10^{-6}$. This means that the Bernoulli equation is satisfied with a small error in the case of the compressible low Mach number flow computed by the developed method.

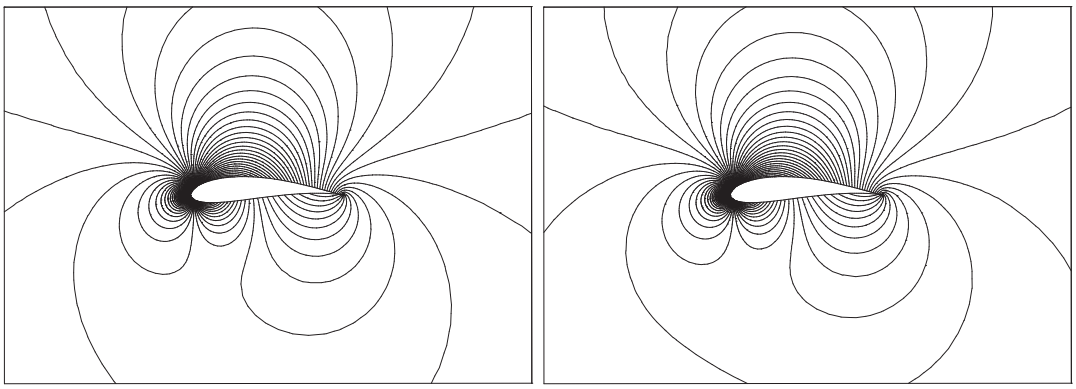


Figure 3.2: Velocity isolines for the exact solution of incompressible flow (left) and approximate solution of compressible flow (right).

2) Irrotational flow past a symmetric Joukowski airfoil We consider flow past a negatively oriented Joukowski profile given by parameters $\Delta = 0.07, a = 0.5, h = 0.0$ (under the notation from [20], Section 2.2.68) with zero angle of attack. This means that the flow, as well as the profile, is symmetric along the horizontal axis. We assume that the far field Mach number of compressible flow $M_\infty = 10^{-4}$. Figure 3.7 shows a detail near the profile of the velocity isolines for the approximate solution of compressible flow and for the exact solution of incompressible flow, respectively. The mesh was formed by 4103 triangular elements.

3) Transonic flow past a nonsymmetric Joukowski airfoil

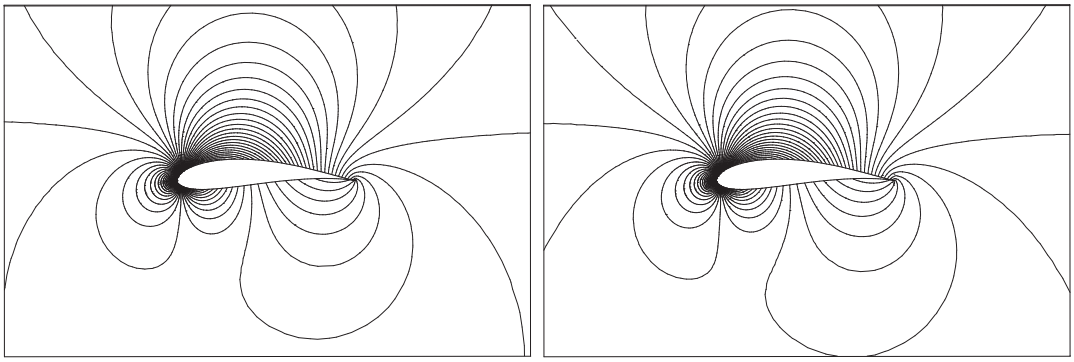


Figure 3.3: Pressure isolines for the exact solution of incompressible flow (left) and approximate solution of compressible flow (right).

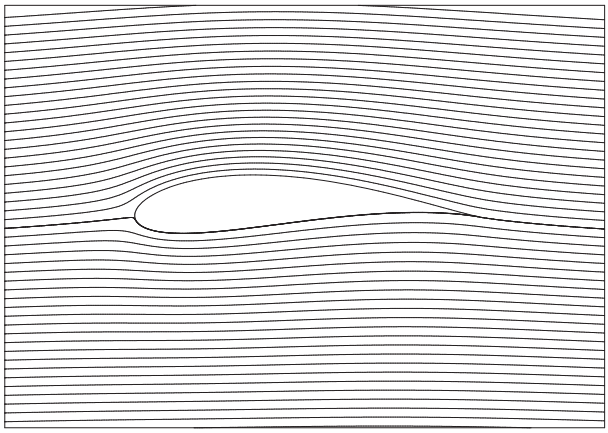


Figure 3.4: Compressible flow past a Joukowski profile, approximate solution, streamlines.

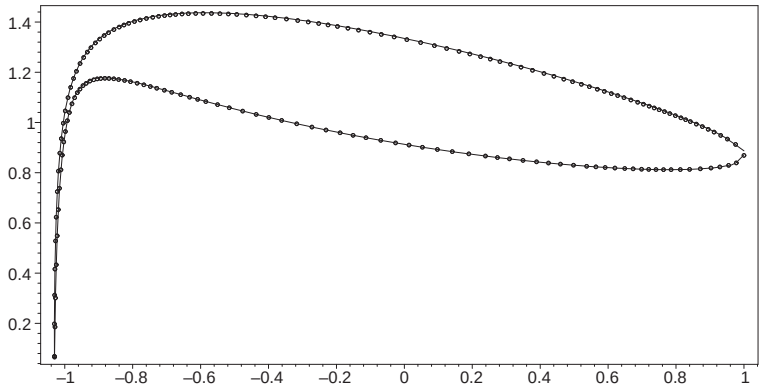


Figure 3.5: Flow past a Joukowski profile, velocity distribution along the profile: ∘ ∘ ∘ – exact solution of incompressible flow, —— – approximate solution of compressible flow.

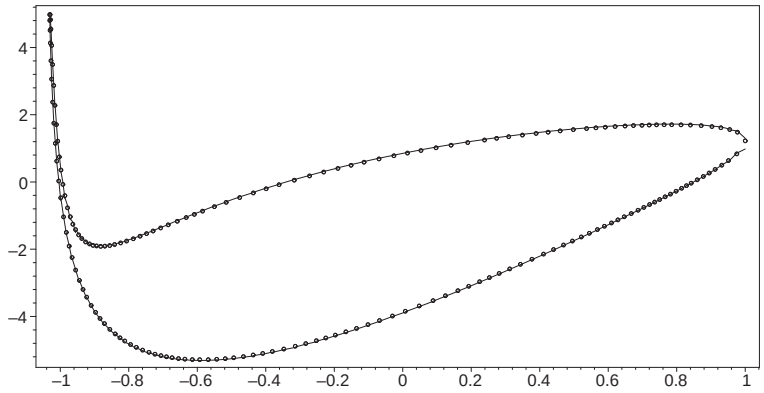


Figure 3.6: Flow past a Joukowski profile, pressure distribution along the profile: ∘ ∘ ∘ – exact solution of incompressible flow, —— – approximate solution of compressible flow.

We shall consider stationary flow past a negatively oriented Joukowski profile given by parameters $\Delta = 0.07, a = 0.5, h = 0.05$ (under the notation from [20], Section 2.2.68) with zero angle of attack.

In order to demonstrate that the presented method allows the solution of high-speed flow with shock waves, we present the results obtained for the transonic ($M_\infty = 0.8$) and hypersonic ($M_\infty = 2.0$) flow with shock waves past the Joukowski profile. In these cases the starting CFL number was chosen 0.08 due to the initial condition, which was constant in the whole computational domain. Then during the computational process the CFL number was successively increased up to 1500 and 2040 in the case of $M_\infty = 0.8$ and $M_\infty = 2.0$, respectively. The computational domain is of the form of a square with side of the length equal to 10 chords of the profile. The numbers of elements were 4451 for $M_\infty = 0.8$ and 4537 for $M_\infty = 2.0$. In both cases the constants from (3.85) and (3.86) had values $\nu_1 = \nu_2 = 0.1$. The maximum density variation was 0.94 and 2.61 in the case $M_\infty = 0.8$ and $M_\infty = 2.0$, respectively.

Figure 3.8 shows Mach number and entropy isolines of transonic flow past the nonsymmetric Joukowski profile for the far field Mach number $M_\infty = 0.8$. Figure 3.8 shows Mach number and entropy isolines of supersonic flow past the nonsymmetric Joukowski profile for the far field Mach number $M_\infty = 2.0$. Since the stabilization proposed in Section 3.7 has a local character, entropy is produced only on shock waves, which is correct from the physical point of view. Figure 3.10 shows elements on which the discrete discontinuity indicator G_i^k, defined by (3.84), is equal to one. By definition, stabilization is applied only on these elements. In figure 3.11 the density distribution along the profile surface is plotted.

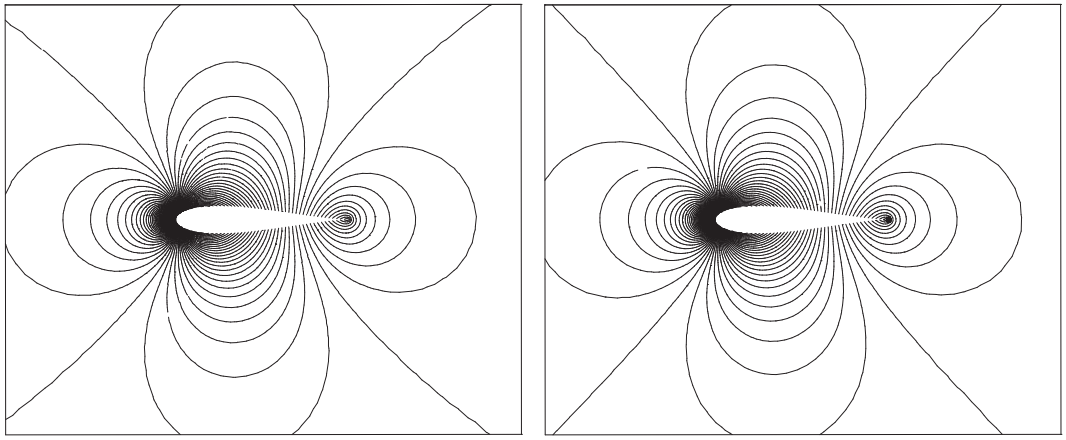


Figure 3.7: Velocity isolines for the approximate solution of compressible flow (left) and for the exact solution of incompressible flow (right).

3.8.2 Irrotational flow past a circular cylinder

Let us consider stationary inviscid flow past a circular cylinder with the far field velocity parallel to the axis x_1 and the Mach number $M_\infty = 10^{-4}$. The problem was solved in a computational domain in the form of a square with sides of the length equal to 20 diameters of the cylinder. We show here details of the flow in the vicinity of the cylinder. Figure 3.12 shows isolines of the absolute value of the velocity for the compressible flow computed by scheme (3.73) with piecewise quadratic elements (i. e. $r = 2$), on a coarse mesh formed by 361 elements and on a fine mesh with 8790 elements, compared with the exact solution of incompressible flow (computed by the method of complex functions – see [20], Section 2.2.35).

In Figure 3.13, the distribution of the absolute value of the velocity along the cylinder, computed on the fine mesh with 8790 elements is shown in dependence on the variable $\vartheta - \pi$, where ϑ is the angle from cylindrical coordinates. We see that the compressible and incompressible velocity distributions are almost identical.

Moreover, Table 3.2 presents the behaviour of the error and experimental order of convergence of the approximate solution $\mathbf{w}_h$ of compressible flow to the exact incompressible solution, measured in $L^\infty(\Omega_h)$-norm. The maximum variation of the density $\rho_{\max} - \rho_{\min} = 2.3 \cdot 10^{-8}$, which corresponds to theoretical results, and $\max_{K \in \mathcal{T}_h} |\nabla \rho_h|_K| < 1.99 \cdot 10^{-6}$. This indicates that the computed flow field behaves as incompressible flow.

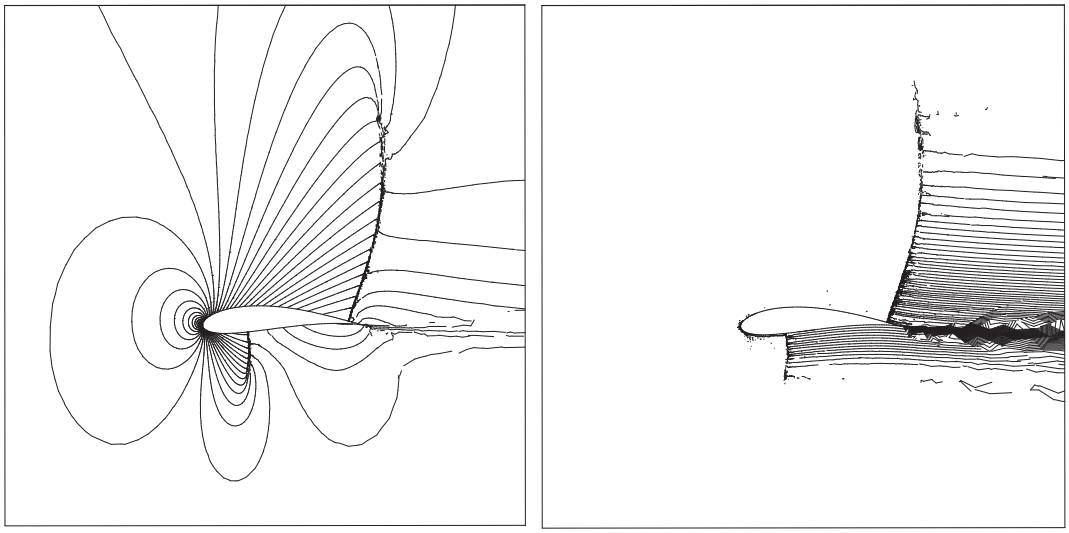


Figure 3.8: Transonic flow past nonsymmetric Joukowski airfoil with $M_\infty = 0.8$, Mach number isolines (left) and entropy isolines (right).

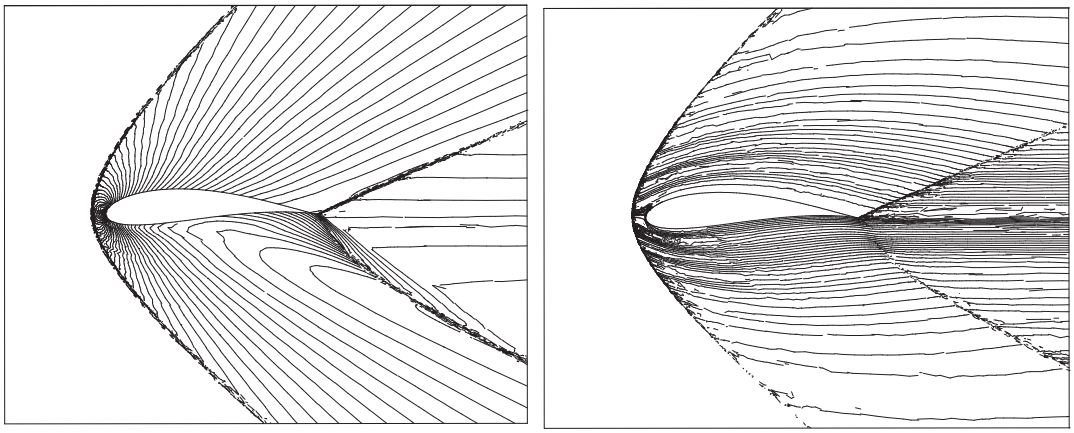


Figure 3.9: Supersonic flow past nonsymmetric Joukowski airfoil with $M_\infty = 2.0$, Mach number isolines (left) and entropy isolines (right).

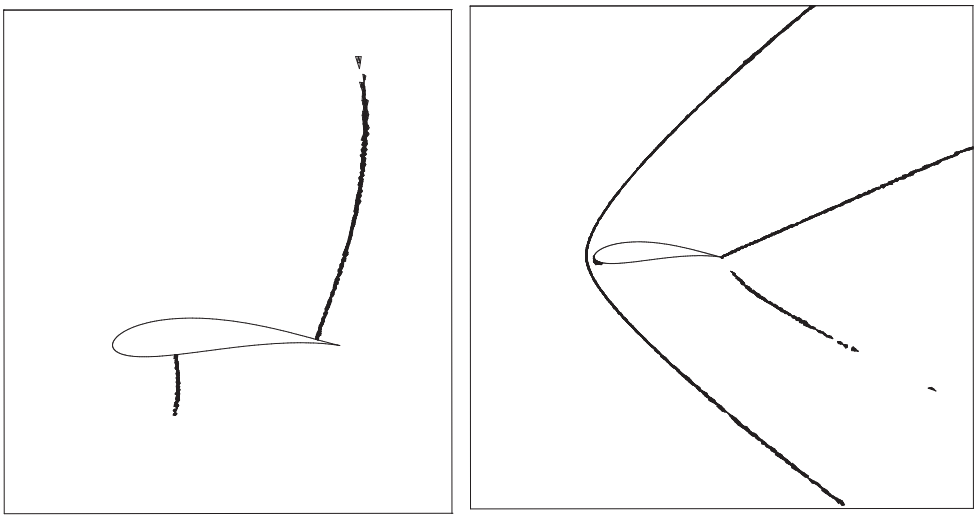


Figure 3.10: Flow past nonsymmetric Joukowski airfoil, elements with active discontinuity indicator, $M_\infty = 0.8$ (left) and $M_\infty = 2.0$ (right).

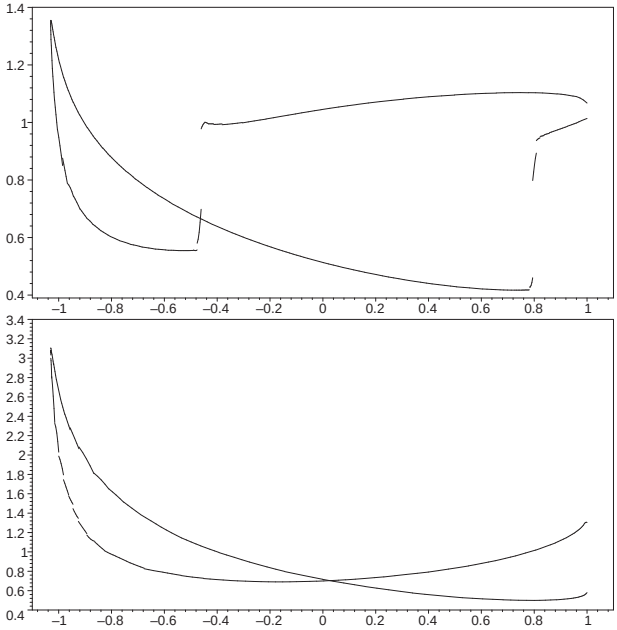


Figure 3.11: Flow past nonsymmetric Joukowski airfoil, density distribution along the profile, $M_\infty = 0.8$ (top) and $M_\infty = 2.0$ (bottom).

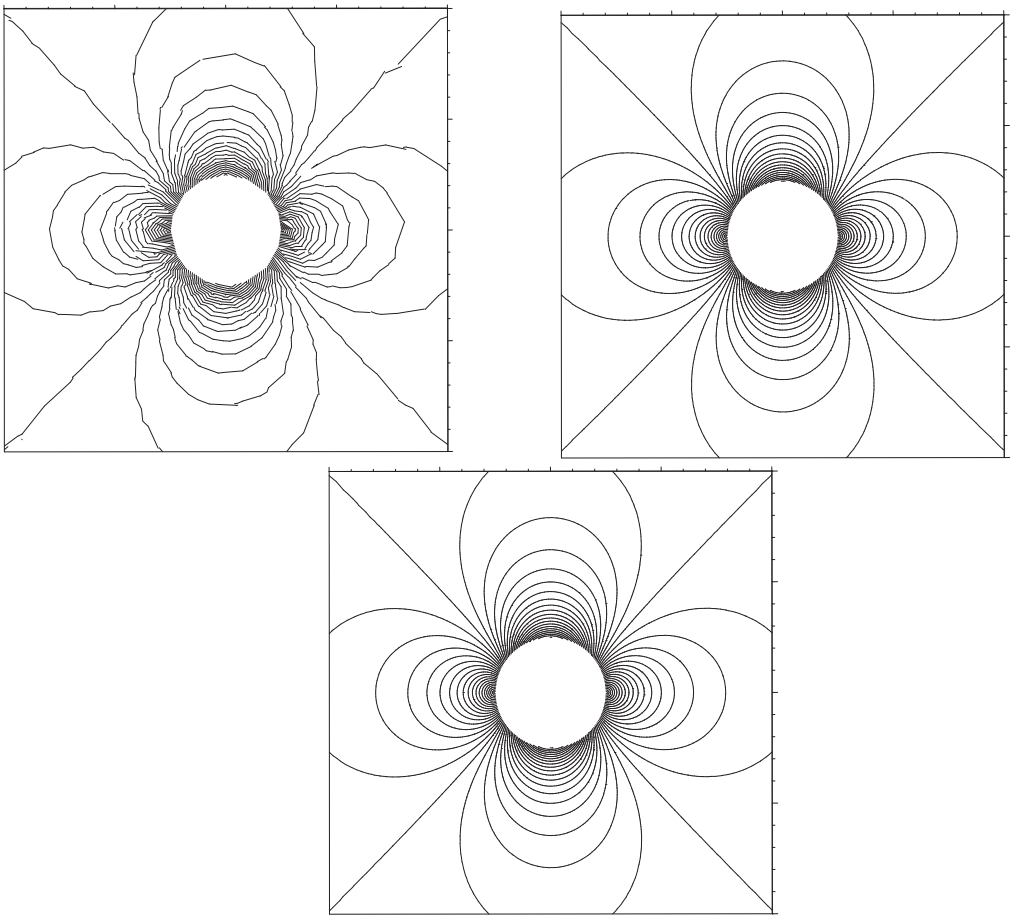


Figure 3.12: Velocity isolines for the approximate solution of compressible flow – coarse mesh (upper left), fine mesh (upper right), compared with the exact solution of incompressible flow (lower)

$\#_h^{\mathrm{T}}$	$\|error\|_{L^\infty(\Omega_h)}$	EOC
1251	5.05E-01	–
1941	4.23E-01	0.406
5031	2.77E-02	2.86
8719	6.68E-03	2.59

Table 3.2: Error in L^∞-norm and corresponding experimental order of convergence for the approximation of incompressible flow by low Mach number compressible flow with respect to $h \to 0$, irrotational flow past a cylinder.

3.8.3 Rotational flow past a circular half-cylinder

In the following example we present the comparison of the exact solution of incompressible inviscid rotational flow past a circular half-cylinder, with center at

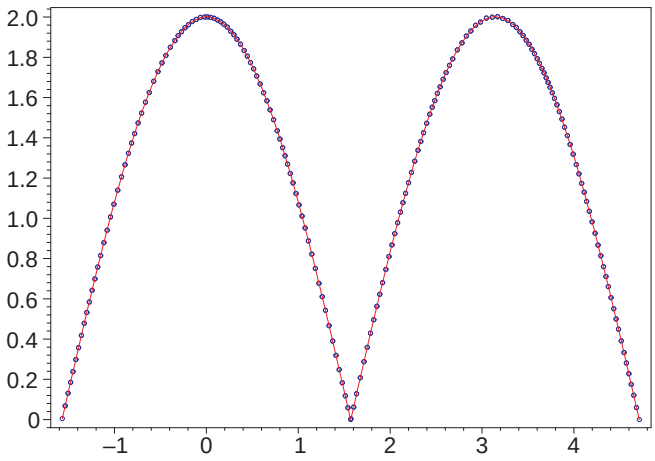


Figure 3.13: Velocity distribution along the cylinder (full line – compressible flow, dotted line – incompressible flow).

the origin and diameter equal to one, with an approximate solution of compressible flow. The analytically exact solution was obtained in [24] by L. Fraenkel under the assumptions of incompressibility and (nonzero) constant vorticity of the flow. Fraenkel notes that this second assumption is not unnatural, since it can be shown that viscous flows bounded by closed streamlines have constant vorticity in the limit of infinite Reynolds number. This flow is interesting for its corner vortices, which develop even though the flow is inviscid.

The far field Mach number is 10^{-4} and the far field velocity has the components $v_1 = x_2, v_2 = 0$. The computational domain was chosen in the form of a rectangle with the length 10 and width 5, from which the half-cylinder was cut off. The mesh was formed by 3541 elements. We present here computational results in the vicinity of the half-cylinder. Figures 3.14 and 3.15 show streamlines of incompressible and compressible flow, respectively. In Figure 3.16 and 3.17 we see the comparison of velocity isolines. Figure 3.18 shows the velocity distribution along the surface of the half-cylinder in dependence on the variable $\vartheta - \pi/2$, where ϑ is the angle from cylindrical coordinates ($\circ\circ\circ$ – exact solution of incompressible flow, —— – approximate solution of compressible flow). The maximum density variation is $3.44 \cdot 10^{-9}$.

3.8.4 Transonic flow through the GAMM channel

The 2D channel used in the following example was proposed by the Gesellschaft für Angewandte Mathematic (GAMM) as a test channel for transonic compressible flow simulation. The channel consists of a 10% circular bump in a rectangular

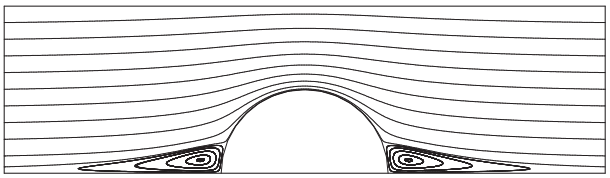


Figure 3.14: Rotational incompressible flow past a half-cylinder, exact solution, streamlines.

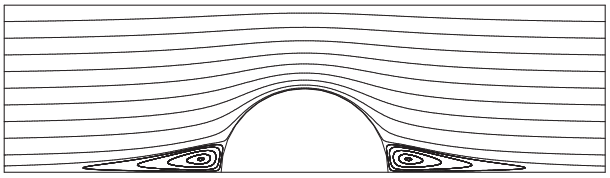


Figure 3.15: Rotational compressible flow past a half-cylinder, approximate solution, streamlines.

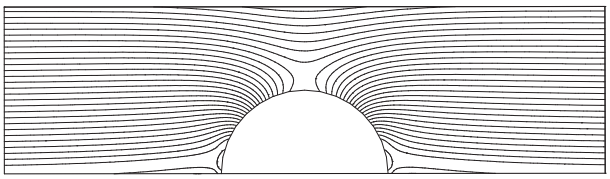


Figure 3.16: Rotational incompressible flow past a half-cylinder, velocity isolines of the exact solution.

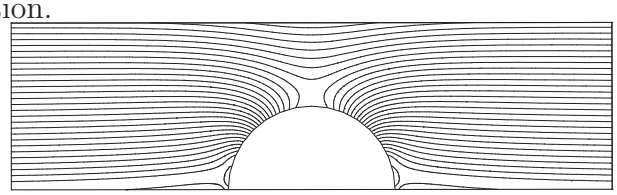


Figure 3.17: Rotational compressible flow past a half-cylinder, velocity isolines of the approximate solution.

domain and the inlet Mach number is taken to be equal to 0.67. In this setup a stable shockwave develops, we therefore apply the shock capturing terms from Section 3.7. Figures 3.19 and 3.20 show Mach number isolines and entropy isolines computed by scheme (3.87). One can see that this scheme yields the entropy production on the shock wave only. In Figure 3.21, the density distribution on the lower wall is plotted. We see that the shock wave is very thin and is ended by the well resolved Zierep singularity (small local maximum). The maximum density variation is 0.693 in this case. In the computational process, the CFL number was successively increased from 30 up to $3 \cdot 10^3$.

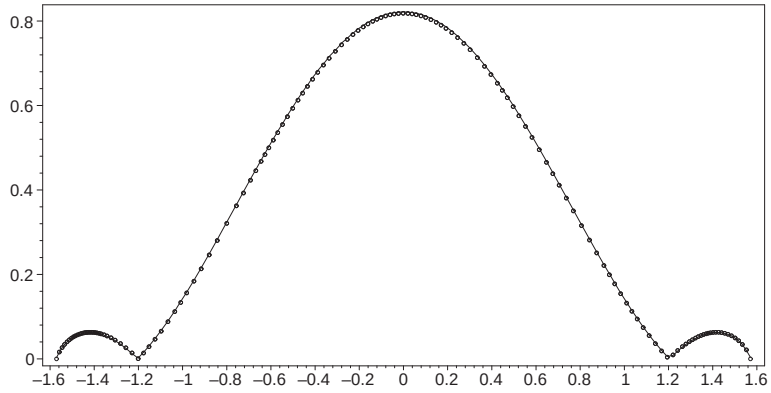


Figure 3.18: Rotational flow past a half-cylinder, velocity distribution on the half-cylinder: ∘ ∘ ∘ – exact solution of incompressible flow, —— – approximate solution of compressible flow.

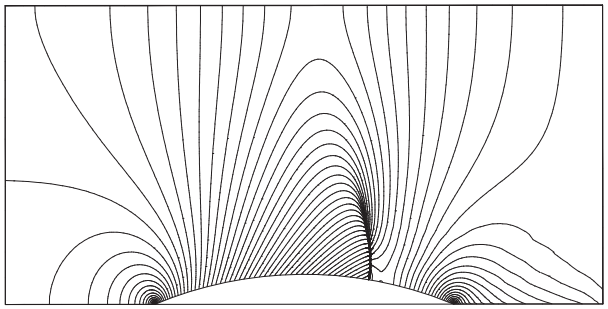


Figure 3.19: Transonic flow through the GAMM channel, Mach number isolines.

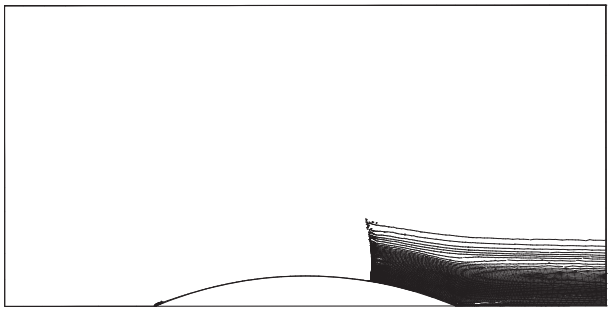


Figure 3.20: Transonic flow through the GAMM channel, entropy isolines.

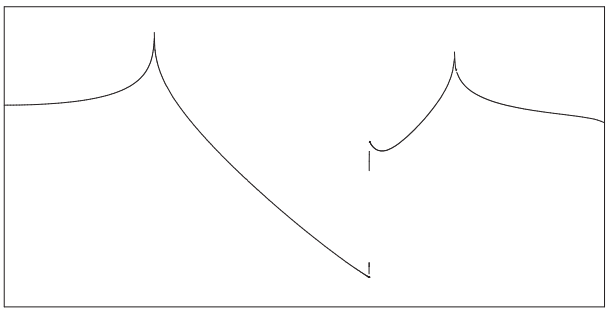


Figure 3.21: Transonic flow through the GAMM channel, density distribution on the lower wall.

Chapter 4

Compressible Navier-Stokes equations

In this chapter we shall apply the discontinuous Galerkin finite element method to compressible viscous flows governed by the compressible Navier-Stokes equations. This system of equations is similar to the nonlinear convection diffusion equation studied in Chapters 1 and 2. However, the treatment of second order terms is not so straightforward for systems of equations as in the scalar case. Again we present three variants how to discretize viscous (diffusion) terms – symmetric, nonsymmetric and incomplete.

4.1 Continuous problem

Let $\Omega \subset \mathbb{R}^2$, $T > 0$, Q_T and the boundary parts $\Gamma_I, \Gamma_O, \Gamma_W$ be the same as in Chapter 3. We want to find a vector-valued function $\mathbf{w} : Q_T \to \mathbb{R}^4$ such that

$$\frac{\partial \mathbf{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\mathbf{w})}{\partial x_s} = \sum_{s=1}^{2} \frac{\partial \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w})}{\partial x_s} + \mathbf{F}(\mathbf{w}) \quad \text{in } Q_T, \tag{4.1}$$

where

$$\begin{aligned}
\mathbf{w} &= (\rho, \rho v_1, \rho v_2, e)^{\mathrm{T}} \in \mathbb{R}^4, \\
\boldsymbol{f}_i(w) &= (f_{i1}(\mathbf{w}), \dots, f_{i4}(\mathbf{w}))^{\mathrm{T}}, \\
&= (\rho v_i, \rho v_1 v_i + \delta_{1i} p, \rho v_2 v_i + \delta_{2i} p, (e+p) v_i)^{\mathrm{T}}, \\
\mathbf{R}_i(\mathbf{w}, \nabla \mathbf{w}) &= (0, \tau_{i1}, \tau_{i2}, \tau_{i1} v_1 + \tau_{i2} v_2 + k \partial \theta / \partial x_i)^{\mathrm{T}}, \\
\tau_{ij} &= \lambda \delta_{ij} \mathrm{div} \mathbf{v} + 2\mu d_{ij}(\mathbf{v}),\ d_{ij}(\mathbf{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right), \\
\mathbf{F}(\mathbf{w}) &= \rho(0, f_1, f_2, q)^{\mathrm{T}}.
\end{aligned} \tag{4.2}$$

To system (4.1) we add the thermodynamical relations

$$p = (\gamma - 1)(e - \rho|\boldsymbol{v}|^2/2), \quad \theta = \left(\frac{e}{\rho} - \frac{1}{2}|\boldsymbol{v}|^2\right)/c_v. \tag{4.3}$$

We use the following notation: θ – absolute temperature, c_v – specific heat at constant volume, μ, λ – viscosity coefficients, k – heat conduction coefficient, $q = q(x,t)$ – density of heat sources. We assume $\mu, k > 0$, $2\mu + 3\lambda \geq 0$. Usually we set $\lambda = -2/\mu 3$. An important quantity in viscous flow is the so-called *Reynolds number*, defined as

$$Re = \frac{U^\star L^\star \rho^\star}{\mu^\star}, \tag{4.4}$$

where $U^\star$ is the characteristic velocity, $L^\star$ is the characteristic length, $\rho^\star$ is the characteristic density and $\mu^\star$ is the characteristic viscosity of the given configuration.

System (4.1) is equipped with the initial condition

$$\mathbf{w}(x,0) = \mathbf{w}^0(x), \quad x \in \Omega, \tag{4.5}$$

and the following set of boundary conditions on appropriate parts of the boundary:

$$\begin{aligned} \text{Case } \Gamma_I: \quad & \text{a) } \rho|_{\Gamma_I \times (0,T)} = \rho_D, \quad \text{b) } \mathbf{v}|_{\Gamma_I \times (0,T)} = \mathbf{v}_D = (v_{D1}, v_{D2})^{\mathrm{T}}, \\ & \text{c) } \sum_{j=1}^{2} \left(\sum_{i=1}^{2} \tau_{ij} n_i \right) v_j + k \frac{\partial \theta}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma_I \times (0,T); \end{aligned} \tag{4.6}$$

$$\text{Case } \Gamma_W: \quad \text{a) } \mathbf{v}_{\Gamma_W \times (0,T)} = 0, \quad \text{b) } \frac{\partial \theta}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma_W \times (0,T); \tag{4.7}$$

$$\text{Case } \Gamma_O: \quad \text{a)} \sum_{i=1}^{2} \tau_{ij} n_i = 0, \; j = 1,2, \quad \text{b) } \frac{\partial \theta}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma_O \times (0,T); \tag{4.8}$$

The viscous fluxes $\mathbf{R}_i(\mathbf{w}, \nabla \mathbf{w})$ have a property similar to the homogeneity of the inviscid fluxes (3.6). The term $\mathbf{R}_i(\mathbf{w}, \nabla \mathbf{w})$ can be expressed in the form

$$\mathbf{R}_i(\mathbf{w}, \nabla \mathbf{w}) = \sum_{j=1}^{2} \mathbb{K}_{ij}(\mathbf{w}) \frac{\partial \mathbf{w}}{\partial x_j}, \tag{4.9}$$

where $\mathbb{K}_{ij}$ are 4×4 matrices dependent on $\mathbf{w}$ and independent of $\nabla \mathbf{w}$. Explicit formulae for $\mathbb{K}_{ij}$ can be found e.g. in [14] or [21]. This property enables us to apply similar concepts as in the semi-implicit linearization of the Euler equations also in the case of viscous flows.

4.2 Discretization

In the DGFE discretization we proceed similarly as in Chapter 1 and 3. The approximate solution $\mathbf{w}_h$ as well as test functions $\boldsymbol{\varphi}_h$ are elements of the finite dimensional space of vector-valued functions $\mathbf{S}_h = [S_h]^4$. By $\gamma_D(i)$ we denote the subset of indices from $j \in \gamma(i)$ such that for at least one component w_r of the solution $\mathbf{w}$ the Dirichlet condition is prescribed on the edge $\Gamma_{ij} \subset \partial\Omega$.

As usual we assume that $\mathbf{w}$ is a sufficiently regular classical solution of the Navier-Stokes equations (4.1), which we multiply by a test function $\boldsymbol{\varphi} \in H^2(\Omega, \mathcal{T}_h)^4$, integrate over $K_i \in \mathcal{T}_h$ and using Green's theorem arrive at the identity

$$\begin{aligned}
&\int_\Omega \frac{\partial \mathbf{w}}{\partial t} \cdot \boldsymbol{\varphi}\, dx + \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{4} \boldsymbol{f}_s(\mathbf{w}) n_{ij}^{(s)} \cdot \boldsymbol{\varphi}\, dS \\
&\quad - \sum_{i\in I} \int_{K_i} \sum_{s=1}^{2} \boldsymbol{f}_s(\mathbf{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s}\, dx + \sum_{i\in I} \int_{K_i} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s}\, dx \\
&\quad - \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \langle \mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w})\rangle n_{ij}^{(s)} \cdot [\boldsymbol{\varphi}]\, dS \qquad (4.10) \\
&\quad - \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w}) n_{ij}^{(s)} \cdot \boldsymbol{\varphi}\, dS \\
&= \int_\Omega \mathbf{F}(\mathbf{w}) \cdot \boldsymbol{\varphi}\, dx.
\end{aligned}$$

In the boundary convective terms, we use the approximation

$$\int_{\Gamma_{ij}} \sum_{s=1}^{2} \boldsymbol{f}_s(\mathbf{w}) n_s^{ij} \cdot \boldsymbol{\varphi}\, dS \approx \int_{\Gamma_{ij}} \mathbf{H}(\mathbf{w}|_{\Gamma_{ij}}, \mathbf{w}|_{\Gamma_{ji}}, \mathbf{n}_{ij}) \cdot \boldsymbol{\varphi}\, dS, \qquad (4.11)$$

incorporating a numerical flux $\mathbf{H}$ as in the case of the Euler equations. Again we use the numerical fluxes discussed in Section 3.3.

If the viscous fluxes $\mathbf{R}_s$ were linear, we could proceed similarly as in Chapter 1. Since this is not the case, we cannot simply exchange the roles of $\mathbf{w}$ and $\boldsymbol{\varphi}$, since the resulting terms would not be linear with respect to $\boldsymbol{\varphi}$. This problem is treated in two ways in [21] and [14]. We briefly describe these methods and derive a different approach. The first method uses the fact that $\mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w})$, $s = 1, 2$, are linear with respect to $\nabla\mathbf{w}$. This leads to the idea of adding the following

terms to the left-hand side of (4.10):

$$\begin{aligned}&\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\big\langle \mathbf{R}_s(\mathbf{w},\nabla\boldsymbol{\varphi})\big\rangle n_{ij}^{(s)}\cdot[\mathbf{w}]\,dS\\ &+\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\boldsymbol{\varphi})n_{ij}^{(s)}\cdot\mathbf{w}\,dS.\end{aligned}\tag{4.12}$$

In the second term we use the zero natural Neumann boundary conditions (4.6), c), (4.7), b) and (4.8). This approach seems natural, however, in [21] it is stated that the scheme based on this linearization does not give satisfactory results. The presumed reason is the fact that the continuity equation is perturbed by other than stabilizing terms. This fact led to development of the second method, a partial linearization with respect to ∇w_i for $i = 2, 3, 4$.

This linearization is obtained by the differentiation inside the definition of $\mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w})$. In the 2D case, when $\lambda = -2\mu/3$ we obtain from (4.2):

$$\begin{aligned}&\mathbf{R}_1(\mathbf{w},\nabla\mathbf{w})\\ &=\begin{pmatrix}0\\ \frac{2}{3}\frac{\mu}{w_1}\big[2(\frac{\partial w_2}{\partial x_1}-\frac{w_2}{w_1}\frac{\partial w_1}{\partial x_1})-(\frac{\partial w_3}{\partial x_2}-\frac{w_3}{w_1}\frac{\partial w_1}{\partial x_2})\big]\\ \frac{\mu}{w_1}\big[(\frac{\partial w_3}{\partial x_1}-\frac{w_3}{w_1}\frac{\partial w_1}{\partial x_1})+(\frac{\partial w_2}{\partial x_2}-\frac{w_2}{w_1}\frac{\partial w_1}{\partial x_2})\big]\\ \frac{w_2}{w_1}\mathbf{R}_1^{(2)}+\frac{w_3}{w_1}\mathbf{R}_1^{(3)}+\frac{k}{c_v w_1}\big[\frac{\partial w_4}{\partial x_1}-\frac{w_4}{w_1}\frac{\partial w_1}{\partial x_1}-\frac{1}{w_1}(w_2\frac{\partial w_2}{\partial x_1}+w_3\frac{\partial w_3}{\partial x_1})\\ \qquad+\frac{1}{w_1^2}(w_2^2+w_3^2)\frac{\partial w_1}{\partial x_1}\big]\end{pmatrix},\end{aligned}\tag{4.13}$$

$$\begin{aligned}&\mathbf{R}_2(\mathbf{w},\nabla\mathbf{w})\\ &=\begin{pmatrix}0\\ \frac{\mu}{w_1}\big[(\frac{\partial w_3}{\partial x_1}-\frac{w_3}{w_1}\frac{\partial w_1}{\partial x_1})+(\frac{\partial w_2}{\partial x_2}-\frac{w_2}{w_1}\frac{\partial w_1}{\partial x_2})\big]\\ \frac{2}{3}\frac{\mu}{w_1}\big[2(\frac{\partial w_3}{\partial x_2}-\frac{w_3}{w_1}\frac{\partial w_1}{\partial x_2})-(\frac{\partial w_2}{\partial x_1}-\frac{w_2}{w_1}\frac{\partial w_1}{\partial x_1})\big]\\ \frac{w_2}{w_1}\mathbf{R}_2^{(2)}+\frac{w_3}{w_1}\mathbf{R}_2^{(3)}+\frac{k}{c_v w_1}\big[\frac{\partial w_4}{\partial x_2}-\frac{w_4}{w_1}\frac{\partial w_1}{\partial x_2}-\frac{1}{w_1}(w_2\frac{\partial w_2}{\partial x_2}+w_3\frac{\partial w_3}{\partial x_2})\\ \qquad+\frac{1}{w_1^2}(w_2^2+w_3^2)\frac{\partial w_1}{\partial x_2}\big]\end{pmatrix},\end{aligned}\tag{4.14}$$

where $\mathbf{R}_s^{(r)} = \mathbf{R}_s^{(r)}(\mathbf{w}, \nabla\mathbf{w})$ denotes the r-th component of $\mathbf{R}_s$ ($s = 1, 2$, $r = 2, 3$). Now for $\mathbf{w}$ and $\boldsymbol{\varphi}$ we define the vector-valued functions

$$\begin{aligned}&\mathbf{D}_1(\mathbf{w},\nabla\mathbf{w},\boldsymbol{\varphi},\nabla\boldsymbol{\varphi})\\ &=\begin{pmatrix}0\\ \frac{2}{3}\frac{\mu}{w_1}\big[2(\frac{\partial \varphi_2}{\partial x_1}-\frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_1})-(\frac{\partial \varphi_3}{\partial x_2}-\frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_2})\big]\\ \frac{\mu}{w_1}\big[(\frac{\partial \varphi_3}{\partial x_1}-\frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_1})+(\frac{\partial \varphi_2}{\partial x_2}-\frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_2})\big]\\ \frac{w_2}{w_1}\mathbf{D}_1^{(2)}+\frac{w_3}{w_1}\mathbf{D}_1^{(3)}+\frac{k}{c_v w_1}\big[\frac{\partial \varphi_4}{\partial x_1}-\frac{\varphi_4}{w_1}\frac{\partial w_1}{\partial x_1}-\frac{1}{w_1}(w_2\frac{\partial \varphi_2}{\partial x_1}+w_3\frac{\partial \varphi_3}{\partial x_1})\\ \qquad+\frac{1}{w_1^2}(w_2\varphi_2+w_3\varphi_3)\frac{\partial w_1}{\partial x_1}\big]\end{pmatrix},\end{aligned}\tag{4.15}$$

$$\mathbf{D}_2(\mathbf{w}, \nabla\mathbf{w}, \boldsymbol{\varphi}, \nabla\boldsymbol{\varphi})$$
$$= \begin{pmatrix} 0 \\ \frac{\mu}{w_1}\left[\left(\frac{\partial\varphi_3}{\partial x_1} - \frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_1}\right) + \left(\frac{\partial\varphi_2}{\partial x_2} - \frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_2}\right)\right] \\ \frac{2}{3}\frac{\mu}{w_1}\left[2\left(\frac{\partial\varphi_3}{\partial x_2} - \frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_2}\right) - \left(\frac{\partial\varphi_2}{\partial x_1} - \frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_1}\right)\right] \\ \frac{w_2}{w_1}\mathbf{D}_2^{(2)} + \frac{w_3}{w_1}\mathbf{D}_2^{(3)} + \frac{k}{c_v w_1}\left[\frac{\partial\varphi_4}{\partial x_2} - \frac{\varphi_4}{w_1}\frac{\partial w_1}{\partial x_2} - \frac{1}{w_1}\left(w_2\frac{\partial\varphi_2}{\partial x_2} + w_3\frac{\partial\varphi_3}{\partial x_2}\right)\right. \\ \left. + \frac{1}{w_1^2}(w_2\varphi_2 + w_3\varphi_3)\frac{\partial w_1}{\partial x_2}\right] \end{pmatrix}, \tag{4.16}$$

where $\mathbf{D}_s^{(r)}$ denotes the r-th component of $\mathbf{D}_s$ ($s = 1, 2$, $r = 2, 3$). Obviously $\mathbf{D}_1$ and $\mathbf{D}_2$ are linear with respect to $\boldsymbol{\varphi}, \nabla\boldsymbol{\varphi}$ and furthermore

$$\mathbf{D}_s(\mathbf{w}, \nabla\mathbf{w}, \mathbf{w}, \nabla\mathbf{w}) = \mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w}), \quad s = 1, 2. \tag{4.17}$$

Now we add the following terms to the left-hand side of (4.10):

$$\begin{aligned} &\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\left\langle\mathbf{D}_s(\mathbf{w}, \nabla\mathbf{w}, \boldsymbol{\varphi}, \nabla\boldsymbol{\varphi})\right\rangle n_{ij}^{(s)}\cdot[\mathbf{w}]\,dS \\ &+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\mathbf{D}_s(\mathbf{w}, \nabla\mathbf{w}, \boldsymbol{\varphi}, \nabla\boldsymbol{\varphi})n_{ij}^{(s)}\cdot\mathbf{w}\,dS \end{aligned} \tag{4.18}$$

(these terms represent a modification of terms (4.12) with appropriate Neumann boundary conditions taken into account). To balance the second term and incorporate the Dirichlet boundary condition we need additional terms on the right-hand side of (4.10). Moreover in analogy with Chapter 1 we add the vanishing *interior penalty* terms

$$\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sigma[\mathbf{w}]\cdot[\boldsymbol{\varphi}]\,dS \tag{4.19}$$

with $\sigma|_{\Gamma_{ij}} = C_W(Re.d(\Gamma_{ij}))^{-1}$ and boundary penalty terms balanced by additional right-hand side terms containing the Dirichlet boundary data.

We can see that the extension from the scalar case to systems is not straightforward. In the next section, we present a different possibility how to discretize the viscous terms.

4.2.1 Discretization of viscous terms

In [2], a framework is provided for a unified treatment of discontinuous Galerkin formulations for elliptic problems. The authors use this approach to derive several discontinuous Galerkin discretizations of the Poisson equation, and as special cases they derive the SIPG and NIPG schemes defined in Chapter 1. Applying

this methodology in the case of the SIPG and NIPG methods gives another possibility how to discretize elliptic terms in the case of systems. The situation in the IIPG case is simpler, since we use directly equation (4.10).

The methodology used in [2] is based on introducing an auxiliary variable, which approximates the gradient of the sought solution in a weak sense. This equation is coupled with the weak formulation for the Poisson equation, which results in a system of the first order equations. After discretizing this system with the discontinuous Galerkin method with a special choice of the numerical flux, one can eliminate the auxiliary variable to obtain the so called *primal formulation*. For an appropriate choice of the numerical flux for the auxiliary equation, one obtains e.g. the NIPG or SIPG methods. In the following, we apply this method to systems of second order equations to obtain a possible generalization of these schemes to systems.

In this section we are interested mainly in the discretization of viscous terms, since convective terms are discussed in Chapter 3. We therefore treat a simplified equation consisting only of the viscous terms contained in the Navier-Stokes equations equipped with a homogeneous Dirichlet boundary condition:

$$\begin{aligned} -\sum_{s=1}^{2} \frac{\partial \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w})}{\partial x_s} = 0 \quad \text{in } \Omega, \\ \mathbf{w} = 0 \quad \text{on } \partial\Omega \end{aligned} \tag{4.20}$$

In order to derive a discontinuous Galerkin formulation, we shall introduce an auxiliary variable $\boldsymbol{\sigma} \approx \nabla \mathbf{w}$, under the notation $\boldsymbol{\sigma}^{(k)} \approx \frac{\partial \mathbf{w}}{\partial x_k}$, for $k = 1, 2$. We assume that $\mathbf{w}$ is a sufficiently regular classical solution of (4.20) and write an equivalent formulation of (4.20) for unknowns $\mathbf{w}, \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2$:

$$\begin{aligned} -\sum_{s=1}^{2} \frac{\partial \mathbf{R}_s(\mathbf{w}, \boldsymbol{\sigma})}{\partial x_s} = 0 \quad \text{in } \Omega, \\ \boldsymbol{\sigma}^{(k)} = \frac{\partial \mathbf{w}}{\partial x_k} \quad \text{for } k = 1, 2. \end{aligned} \tag{4.21}$$

To derive a suitable weak formulation, we multiply the first equation by a test function $\boldsymbol{\varphi} \in [S_h]^4$ and the second equation by the test function $\boldsymbol{\tau} \in \Sigma_h$, where Σ_h is an appropriate function space, which we shall discuss later. We integrate over an element $K_i \in \mathcal{T}_h$, apply Green's theorem and sum over all elements:

$$\begin{aligned} \sum_{i \in I} \int_{K_i} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}, \boldsymbol{\sigma}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s} \, dx - \sum_{i \in I} \int_{\partial K_i} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}, \boldsymbol{\sigma}) n_{K_i}^{(s)} \cdot \boldsymbol{\varphi} \, dS \\ = 0, \quad \forall \boldsymbol{\varphi} \in [S_h]^4, \\ \sum_{i \in I} \int_{K_i} \boldsymbol{\sigma}^{(k)} \cdot \boldsymbol{\tau} \, \mathrm{d}x = -\sum_{i \in I} \int_{K_i} \mathbf{w} \cdot \frac{\partial \boldsymbol{\tau}}{\partial x_k} \, \mathrm{d}x + \sum_{i \in I} \int_{\partial K_i} \mathbf{w} n_{K_i}^{(k)} \cdot \boldsymbol{\tau} \, \mathrm{d}S, \quad \forall \boldsymbol{\tau} \in \Sigma_h. \end{aligned} \tag{4.22}$$

To derive a discontinuous Galerkin formulation of (4.22), we proceed as in the case of convective terms. We introduce numerical fluxes H_{ij}^{w} and H_{ij}^{σ} into the boundary integrals in each equation:

$$\begin{aligned}\int_{\Gamma_{ij}} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w},\boldsymbol{\sigma}) n_{ij}^{(s)} \cdot \boldsymbol{\varphi}\, dS &\approx \int_{\Gamma_{ij}} H_{ij}^{\mathbf{w}} \cdot \boldsymbol{\varphi}\, dS \\ \int_{\Gamma_{ij}} \mathbf{w} n_{ij}^{(k)} \cdot \boldsymbol{\tau}\, \mathrm{d}S &\approx \int_{\Gamma_{ij}} H_{ij}^{\sigma} n_{ij}^{(k)} \cdot \boldsymbol{\tau}\, \mathrm{d}S.\end{aligned} \tag{4.23}$$

The choice of different numerical fluxes H_{ij}^{w} and H_{ij}^{σ} leads to different numerical schemes. For instance, according to [2], for the Poisson equation, the following choices lead to the *nonsymmetric* variant of Chapter 1:

$$\begin{aligned} H_{ij}^{\mathbf{w}} &:= \begin{cases} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}_{ij}, \nabla \mathbf{w}_{ij}) n_{ij}^{(s)} & \text{for } \Gamma_{ij} \subset \partial\Omega, \\ \sum_{s=1}^{2} \big\langle \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w}) \big\rangle n_{ij}^{(s)} & \text{otherwise.} \end{cases} \\ H_{ij}^{\sigma} &:= \begin{cases} 2\mathbf{w}_{ij} & \text{for } \Gamma_{ij} \subset \partial\Omega, \\ \langle \mathbf{w} \rangle + [\mathbf{w}] & \text{otherwise.} \end{cases} \end{aligned} \tag{4.24}$$

To obtain the *symmetric* variant, we use a different numerical flux definition. Namely, $H_{ij}^{\mathbf{w}}$ is the same as in (4.24) and H_{ij}^{σ} is defined as

$$H_{ij}^{\sigma} := \begin{cases} 0 & \text{for } \Gamma_{ij} \subset \partial\Omega, \\ \langle \mathbf{w} \rangle & \text{otherwise.} \end{cases} \tag{4.25}$$

By replacing boundary terms in (4.23) by the numerical fluxes defined in (4.24), we obtain after some manipulation the following discrete formulation of system (4.21), which leads to the NIPG scheme:

$$\begin{aligned} &\sum_{i\in I} \int_{K_i} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w},\boldsymbol{\sigma}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s}\, dx - \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \langle \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w}) \rangle n_{ij}^{(s)} \cdot [\boldsymbol{\varphi}]\, dS \\ &\qquad - \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}_{ij}, \nabla \mathbf{w}_{ij}) n_{ij}^{(s)} \cdot \boldsymbol{\varphi}_{ij}\, dS = 0, \quad \forall \boldsymbol{\varphi} \in [S_h]^4, \\ &\sum_{i\in I} \int_{K_i} \boldsymbol{\sigma}^{(k)} \cdot \boldsymbol{\tau}\, \mathrm{d}x = -\sum_{i\in I} \int_{K_i} \mathbf{w} \cdot \frac{\partial \boldsymbol{\tau}}{\partial x_k}\, \mathrm{d}x + \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \big(\langle \mathbf{w} \rangle + [\mathbf{w}]\big) n_{ij}^{(k)} \cdot [\boldsymbol{\tau}]\, \mathrm{d}S \\ &\qquad + \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} 2\mathbf{w}_{ij} n_{ij}^{(k)} \cdot \boldsymbol{\tau}_{ij}\, \mathrm{d}S, \quad k = 1,2,\ \forall \boldsymbol{\tau} \in \Sigma_h, \end{aligned} \tag{4.26}$$

Similarly as in [2] for the Poisson equation, we shall eliminate the auxiliary variable $\boldsymbol{\sigma}$ from system (4.26) due to the choice of numerical fluxes (4.24). This procedure then leads to the so-called *primal formulation*, which will, in our case, give a generalization of the NIPG and SIPG schemes from the scalar case derived in Chapter 1 to systems of equations.

To eliminate the variable $\boldsymbol{\sigma}$ from the first equation in (4.26), we notice that this variable figures only in the first term. Using property (4.9), we write

$$\begin{aligned}
&\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\boldsymbol{\sigma})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx\\
&\quad=\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\left(\sum_{k=1}^{2}\mathbb{K}_{sk}(\mathbf{w})\boldsymbol{\sigma}^{(k)}\right)\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx\\
&\quad=\sum_{i\in I}\int_{K_i}\sum_{k=1}^{2}\boldsymbol{\sigma}^{(k)}\cdot\left(\sum_{s=1}^{2}\mathbb{K}_{sk}^{T}(\mathbf{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\right)dx.
\end{aligned}\tag{4.27}$$

Now we take the second equation in (4.26) and apply Green's theorem to its first right-hand side term:

$$\begin{aligned}
&\sum_{i\in I}\int_{K_i}\boldsymbol{\sigma}^{(k)}\cdot\boldsymbol{\tau}\,\mathrm{d}x\\
&\quad=-\sum_{i\in I}\int_{K_i}\mathbf{w}\cdot\frac{\partial\boldsymbol{\tau}}{\partial x_k}\,\mathrm{d}x+\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\big(\langle\mathbf{w}\rangle+[\mathbf{w}]\big)n_{ij}^{(k)}\cdot[\boldsymbol{\tau}]\,\mathrm{d}S\\
&\qquad+\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}2\mathbf{w}_{ij}n_{ij}^{(k)}\cdot\boldsymbol{\tau}_{ij}\,\mathrm{d}S\\
&\quad=\sum_{i\in I}\int_{K_i}\frac{\partial\mathbf{w}}{\partial x_k}\cdot\boldsymbol{\tau}\,\mathrm{d}x+\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}[\mathbf{w}]n_{ij}^{(k)}\cdot\langle\boldsymbol{\tau}\rangle\,\mathrm{d}S\\
&\qquad+\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}\mathbf{w}_{ij}n_{ij}^{(k)}\cdot\boldsymbol{\tau}_{ij}\,\mathrm{d}S,\quad\forall\boldsymbol{\tau}\in\Sigma_h.
\end{aligned}\tag{4.28}$$

If we set set $\boldsymbol{\tau}:=\sum_{s=1}^{2}\mathbb{K}_{sk}^{T}(\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}$ in (4.28), we can express (4.27) without the use of $\boldsymbol{\sigma}$. We proceed in the following way. We introduce the notation similar to (4.9):

$$\widetilde{\mathbf{R}}_k(\mathbf{w},\nabla\boldsymbol{\varphi}):=\sum_{s=1}^{2}\mathbb{K}_{sk}^{T}(\mathbf{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_s}.\tag{4.29}$$

Now, the use of (4.28) in (4.27) gives us

$$
\begin{aligned}
&\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\boldsymbol{\sigma})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx\\
&\quad=\sum_{i\in I}\int_{K_i}\sum_{k=1}^{2}\boldsymbol{\sigma}^{(k)}\cdot\Big(\sum_{s=1}^{2}\mathbb{K}_{sk}^{T}(\mathbf{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\Big)\,dx\\
&\quad=\sum_{i\in I}\int_{K_i}\sum_{k=1}^{2}\frac{\partial\mathbf{w}}{\partial x_k}\cdot\Big(\sum_{s=1}^{2}\mathbb{K}_{sk}^{T}(\mathbf{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\Big)\,\mathrm{d}x\\
&\qquad+\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{k=1}^{2}[\mathbf{w}]n_{ij}^{(k)}\cdot\Big\langle\sum_{s=1}^{2}\mathbb{K}_{sk}^{T}(\mathbf{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\Big\rangle\,\mathrm{d}S\\
&\qquad+\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}\sum_{k=1}^{2}\mathbf{w}_{ij}n_{ij}^{(k)}\cdot\Big(\sum_{s=1}^{2}\mathbb{K}_{sk}^{T}(\mathbf{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\Big)\,\mathrm{d}S\\
&\quad=\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,\mathrm{d}x\\
&\qquad+\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{k=1}^{2}[\mathbf{w}]n_{ij}^{(k)}\cdot\big\langle\widetilde{\mathbf{R}}_k(\mathbf{w},\nabla\boldsymbol{\varphi})\big\rangle\,\mathrm{d}S\\
&\qquad+\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}\sum_{k=1}^{2}\mathbf{w}n_{ij}^{(k)}\cdot\widetilde{\mathbf{R}}_k(\mathbf{w},\nabla\boldsymbol{\varphi})\,\mathrm{d}S.
\end{aligned}
\tag{4.30}
$$

Finally, if we use this expression in the first equation of (4.26), we obtain the following discontinuous Galerkin formulation of problem (4.20):

$$
\begin{aligned}
&\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,\mathrm{d}x\\
&-\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\langle\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\rangle n_{ij}^{(s)}\cdot[\boldsymbol{\varphi}]\,dS-\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})n_{ij}^{(s)}\cdot\boldsymbol{\varphi}\,dS\\
&+\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\big\langle\widetilde{\mathbf{R}}_s(\mathbf{w},\nabla\boldsymbol{\varphi})\big\rangle n_{ij}^{(s)}\cdot[\mathbf{w}]\,\mathrm{d}S+\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\widetilde{\mathbf{R}}_s(\mathbf{w},\nabla\boldsymbol{\varphi})n_{ij}^{(s)}\cdot\mathbf{w}\,\mathrm{d}S\\
&=0,\quad\forall\boldsymbol{\varphi}\in[S_h]^4.
\end{aligned}
\tag{4.31}
$$

As in Chapter 1, we need to add an interior and boundary penalty term $J_h(\mathbf{w}, \boldsymbol{\varphi})$ to the left-hand side of (4.31):

$$J_h(\mathbf{w}, \boldsymbol{\varphi}) = \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j<i}} \int_{\Gamma_{ij}} \sigma[\mathbf{w}] \cdot [\boldsymbol{\varphi}] \, dS + \sum_{i \in I} \sum_{j \in \gamma_D(i)} \int_{\Gamma_{ij}} \sigma \mathbf{w} \cdot \boldsymbol{\varphi} \, dS, \tag{4.32}$$

where σ is an appropriate parameter defined by

$$\sigma|_{\Gamma_{ij}} = \frac{C_W}{Re.d(\Gamma_{ij})}, \tag{4.33}$$

Where $C_W > 0$. This procedure leads to the *nonsymmetric* formulation, the *symmetric* can be derived analogously, and the *incomplete* variant is the simplest, since it does not use artificially added terms. As in Section 2.2, the constant $C_W > 0$ must be chosen large enough in the symmetric and incomplete variants to ensure *coercivity* of the resulting diffusion form.

4.2.2 Discrete problem

On the basis of the preceding section, we introduce the following forms defining the discrete formulation of problem (4.1) equipped with boundary conditions (4.6)-(4.8). Here $\widetilde{\mathbf{R}}_s(\mathbf{w}, \nabla\boldsymbol{\varphi})$ is defined in (4.29).
Nonsymmetric variant of the diffusion form:

$$\begin{aligned} a_h^N(\mathbf{w}, \boldsymbol{\varphi}) = & \sum_{i \in I} \int_{K_i} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s} \, dx \\ & - \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j<i}} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \langle \mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w}) \rangle n_{ij}^{(s)} \cdot [\boldsymbol{\varphi}] \, dS - \sum_{i \in I} \sum_{j \in \gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \mathbf{R}_s(\mathbf{w}, \nabla\mathbf{w}) n_{ij}^{(s)} \cdot \boldsymbol{\varphi} \, dS \\ & + \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j<i}} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \langle \widetilde{\mathbf{R}}_s(\mathbf{w}, \nabla\boldsymbol{\varphi}) \rangle n_{ij}^{(s)} \cdot [\mathbf{w}] \, dS + \sum_{i \in I} \sum_{j \in \gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \widetilde{\mathbf{R}}_s(\mathbf{w}, \nabla\boldsymbol{\varphi}) n_{ij}^{(s)} \cdot \mathbf{w} \, dS, \end{aligned} \tag{4.34}$$

symmetric variant of the diffusion form:

$$
\begin{aligned}
a_h^S(\mathbf{w},\boldsymbol{\varphi}) &= \sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx \\
&- \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\langle\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\rangle n_{ij}^{(s)}\cdot[\boldsymbol{\varphi}]\,dS - \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})n_{ij}^{(s)}\cdot\boldsymbol{\varphi}\,dS \\
&- \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\langle\widetilde{\mathbf{R}}_s(\mathbf{w},\nabla\boldsymbol{\varphi})\rangle n_{ij}^{(s)}\cdot[\mathbf{w}]\,dS - \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\widetilde{\mathbf{R}}_s(\mathbf{w},\nabla\boldsymbol{\varphi})n_{ij}^{(s)}\cdot\mathbf{w}\,dS,
\end{aligned}
\tag{4.35}
$$

incomplete variant of the diffusion form:

$$
\begin{aligned}
a_h^I(\mathbf{w},\boldsymbol{\varphi}) &= \sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx \\
&- \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\langle\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\rangle n_{ij}^{(s)}\cdot[\boldsymbol{\varphi}]\,dS - \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})n_{ij}^{(s)}\cdot\boldsymbol{\varphi}\,dS,
\end{aligned}
\tag{4.36}
$$

interior and boundary penalty jump terms:

$$
J_h(\mathbf{w},\boldsymbol{\varphi}) = \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sigma[\mathbf{w}]\cdot[\boldsymbol{\varphi}]\,dS + \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sigma\mathbf{w}\cdot\boldsymbol{\varphi}\,dS \tag{4.37}
$$

Further we define the *nonsymmetric* right-hand side form:

$$
\begin{aligned}
l_h^N(\mathbf{w},\boldsymbol{\varphi}) &= \int_\Omega\mathbf{F}(\mathbf{w})\cdot\boldsymbol{\varphi}\,dx + \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\sigma\mathbf{w}_B\cdot\boldsymbol{\varphi}\,dS \\
&+ \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\widetilde{\mathbf{R}}_s(\mathbf{w},\nabla\boldsymbol{\varphi})n_{ij}^{(s)}\cdot\mathbf{w}_B\,dS,
\end{aligned}
\tag{4.38}
$$

symmetric right-hand side form:

$$
\begin{aligned}
l_h^S(\mathbf{w},\boldsymbol{\varphi}) &= \int_\Omega\mathbf{F}(\mathbf{w})\cdot\boldsymbol{\varphi}\,dx + \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\sigma\mathbf{w}_B\cdot\boldsymbol{\varphi}\,dS \\
&- \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\widetilde{\mathbf{R}}_s(\mathbf{w},\nabla\boldsymbol{\varphi})n_{ij}^{(s)}\cdot\mathbf{w}_B\,dS,
\end{aligned}
\tag{4.39}
$$

incomplete right-hand side form:

$$l_h^I(\mathbf{w},\boldsymbol{\varphi}) = \int_\Omega \mathbf{F}(\mathbf{w})\cdot\boldsymbol{\varphi}\,dx + \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\sigma\mathbf{w}_B\cdot\boldsymbol{\varphi}\,dS. \tag{4.40}$$

Finally, we define the convective terms:

$$\begin{aligned} b_h(\mathbf{w},\boldsymbol{\varphi}) = &-\sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\boldsymbol{f}_s(\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx \\ &+\sum_{i\in I}\sum_{j\in S(i)}\int_{\Gamma_{ij}}\mathbf{H}(\mathbf{w}|_{\Gamma_{ij}},\mathbf{w}|_{\Gamma_{ji}},\mathbf{n}_{ij})\cdot\boldsymbol{\varphi}\,dS, \end{aligned} \tag{4.41}$$

here $\Gamma_{ij}\subset\partial\Omega$ than $\mathbf{w}|_{\Gamma_{ji}}$ is determined on the basis of inviscid boundary conditions as presented in Section 3.4.

The boundary state $\mathbf{w}_B = (w_{B1},\dots,w_{B4})^{\mathrm{T}}$ is defined as follows: we set

$$w_{Br}|_{\Gamma_{ij}} = w^*_{Br}|_{\Gamma_{ij}}, \tag{4.42}$$

if the r-th component w_r of $\mathbf{w}$ is prescribed on Γ_{ij} in the boundary conditions (4.6) – (4.8). By $\mathbf{w}^*$ we mean the function satisfying the Dirichlet boundary conditions. Otherwise, we set

$$w_{Br}|_{\Gamma_{ij}} = w_r|_{\Gamma_{ij}}, \tag{4.43}$$

i.e. we use "extrapolation" of w_r onto Γ_{ij} from $K_i\in\mathcal{T}_h$. In particular, we have

$$\begin{aligned} &\mathbf{w}_B = (\rho_{ij},0,0,c_v\rho_{ij}\theta_{ij}) \quad \text{on } \Gamma_W,\\ &\mathbf{w}_B = \left(\rho_D,\rho_D v_{D1},\rho_D v_{D2},c_v\rho_D\theta_D+\frac{1}{2}\rho_D|\mathbf{v}_D|^2\right) \quad \text{on } \Gamma_I,\\ &\mathbf{w}_B = \mathbf{w}_{ij} \quad \text{on } \Gamma_O. \end{aligned} \tag{4.44}$$

As for the Neumann-type boundary conditions (4.6),c), (4.7),b) and (4.8),a),b), they are incorporated into the definition of boundary terms in the definition diffusion forms. Finally, in the convective terms (4.41) we incorporate the inviscid characteristic-based boundary conditions as defined in Section 3.4.3.

For simplicity of notation, we omit the superscripts *N, S* and *I* in the following definition and use the generic notation for the diffusion and right-hand side forms $a_h(u_h,\varphi)$ and $l_h(\varphi_h)$. The *symmetric, nonsymmetric* and *incomplete* variants can be obtained by taking in turn $a_h := a_h^S$, $l_h := l_h^S$ and so on. Now we can define the discrete DGFE Navier-Stokes problem:

Definition 4.2.1 *We say that* $\mathbf{w}_h$ *is a DGFE solution of the compressible Navier-Stokes equations (4.1) - (4.2), if*

$$\begin{aligned}
&a)\ \mathbf{w}_h \in C^1([0,T];\mathbf{S}_h),\\
&b)\ \frac{d}{dt}(\mathbf{w}_h(t),\boldsymbol{\varphi}_h) + b_h(\mathbf{w}_h(t),\boldsymbol{\varphi}_h) + J_h(\mathbf{w}_h(t),\boldsymbol{\varphi}_h) + a_h(\mathbf{w}_h(t),\boldsymbol{\varphi}_h)\\
&\qquad = l_h(\mathbf{w}_h,\boldsymbol{\varphi}_h)(t), \quad \forall \boldsymbol{\varphi}_h \in \mathbf{S}_h,\ \forall t \in (0,T),\\
&c)\ \mathbf{w}_h(0) = \mathbf{w}_h^0,
\end{aligned} \tag{4.45}$$

where $\mathbf{w}_h^0$ *is an* $\mathbf{S}_h$ *approximation of the initial condition* $\mathbf{w}^0$.

4.3 Time discretization

The semi-implicit scheme presented in Section 3.6.2 is a discretization of the Euler equations. Here we incorporate the viscous terms into the semi-implicit scheme.

This method is based on relation (4.9), which gives a possibility to linearize the viscous terms similarly as in the semi-implicit linearization of the Euler equations. We linearize only the nonsymmetric variant, the symmetric and incomplete variants are analogous. For simplicity of notation we again omit the superscript N and write only a_h^{SI}.

Let us define the following linearized forms based on definition (4.34) and the definition of $\widetilde{\mathbf{R}}_k(\mathbf{w},\nabla\boldsymbol{\varphi})$ given in (4.29):

$$\begin{aligned}
a_h^{SI}(\mathbf{w}_h^k,\mathbf{w}_h^{k+1},\boldsymbol{\varphi}_h) &= \sum_{i\in I}\int_{K_i}\sum_{s=1}^{2}\sum_{t=1}^{2}\mathbb{K}_{st}(\mathbf{w}_h^k)\frac{\partial \mathbf{w}_h^{k+1}}{\partial x_t}\cdot\frac{\partial\boldsymbol{\varphi}_h}{\partial x_s}\,dx\\
&- \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\left\langle\sum_{t=1}^{2}\mathbb{K}_{st}(\mathbf{w}_h^k)\frac{\partial \mathbf{w}_h^{k+1}}{\partial x_t}\right\rangle n_{ij}^{(s)}\cdot[\boldsymbol{\varphi}_h]\,dS\\
&- \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\sum_{t=1}^{2}\mathbb{K}_{st}(\mathbf{w}_h^k)\frac{\partial \mathbf{w}_h^{k+1}}{\partial x_t}n_{ij}^{(s)}\cdot\boldsymbol{\varphi}_h\,dS\\
&+ \sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\left\langle\sum_{t=1}^{2}\mathbb{K}_{ts}^T(\mathbf{w}_h^k)\frac{\partial\boldsymbol{\varphi}_h}{\partial x_t}\right\rangle n_{ij}^{(s)}\cdot[\mathbf{w}_h^{k+1}]\,dS\\
&+ \sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{2}\sum_{t=1}^{2}\mathbb{K}_{ts}^T(\mathbf{w}_h^k)\frac{\partial\boldsymbol{\varphi}_h}{\partial x_t}n_{ij}^{(s)}\cdot\mathbf{w}_h^{k+1}\,dS,
\end{aligned} \tag{4.46}$$

(nonsymmetric linearized diffusion form),

$$J_h^{SI}(\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sigma[\mathbf{w}_h^{k+1}] \cdot [\boldsymbol{\varphi}_h]\, dS + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma \mathbf{w}_h^{k+1} \cdot \boldsymbol{\varphi}_h\, dS \tag{4.47}$$

(linearized interior and boundary penalty jump terms),

$$\begin{aligned} l_h^{SI}(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = &\int_\Omega \mathbf{F}(\mathbf{w}_h^k) \cdot \boldsymbol{\varphi}_h\, dx \\ &+ \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \sum_{t=1}^{2} \mathbb{K}_{ts}^T(\mathbf{w}_h^k) \frac{\partial \boldsymbol{\varphi}_h}{\partial x_t} n_{ij}^{(s)} \cdot \mathbf{w}_B\, dS \\ &+ \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \sigma \mathbf{w}_B \cdot \boldsymbol{\varphi}_h\, dS \end{aligned} \tag{4.48}$$

(right-hand side form).

With the aid of the linearized inviscid form b_h^{SI} defined in Section 3.6.2, we define the semi-implicit linearized scheme:

Definition 4.3.1 *For each $k = 0, 1, \ldots$ find $\mathbf{w}_h^{k+1}$ such that*

$$\begin{aligned} &a)\ \mathbf{w}_h^{k+1} \in \mathbf{S}_h, \\ &b)\ (\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \tau_k b_h^{SI}(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \tau_k a_h^{SI}(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) \\ &\qquad + \tau_k J_h^{SI}(\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = \tau_k l_h^{SI}(\mathbf{w}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + (\mathbf{w}_h^k, \boldsymbol{\varphi}_h), \\ &\qquad \forall \boldsymbol{\varphi}_h \in \mathbf{S}_h,\ k = 0, 1, \ldots, \\ &c)\ \mathbf{w}_h^0 = \tilde{\mathbf{w}}_h^0. \end{aligned} \tag{4.49}$$

where $\tilde{\mathbf{w}}_h^0$ is an S_h approximation of the initial condition $\mathbf{w}^0$.

This scheme is linear with respect to $\mathbf{w}_h^{k+1}$ and can be implemented with the aid of an efficient linear solver as in the inviscid case.

4.4 Numerical experiments

In this section we present the solution of some test problems in order to demonstrate the performance of the proposed discontinuous Galerkin formulation of the compressible Navier-Stokes equations. As in Section 3.8, the computational grids were constructed with the aid of the anisotropic mesh adaptation technique [13] and quadratic elements ($r = 2$) were applied in the following examples. Steady

state solutions were obtained using scheme the incomplete variant of scheme (4.49) for "$t_k \to \infty$".

In the viscous case, the matrix sparsity structure of the resulting system of equations is more complicated than in the case of inviscid flows, when Lagrange basis functions are used. This caused the poor performance of the UMFPACK direct solver, which was not able to handle larger systems with the memory available. Therefore block Jacobi preconditioned GMRES was used, however this iterative solver proved to be insufficient in the case of low-Mach flows. Therefore in following examples the exact incompressible solutions are compared with numerical solutions of the compressible equations with Mach number $M = 0.1$, which is much larger than in Section 3.8, where $M = 10^{-4}$ is commonly attainable with the direct linear solver.

4.4.1 Viscous boundary layer

In the first example we consider a simple configuration consisting of a boundary layer near an impermeable wall. In our case, the wall (or flat plate) is represented by the set $\Gamma_W = \{(x_1, x_2) \in I\!R^2 : x_1 \in [0, +\infty), x_2 = 0\}$ and the far field velocity has components $\mathbf{v}_\infty = (v_\infty^{(1)}, 0)$. In the numerical computation we choose the freestream Mach number $M = 0.1$ and Reynolds number $Re = 10^3$. The computational mesh has 2479 elements and was adaptively refined along the flat plate. If we can neglect the compressibility of the flow, the computed stationary solution can be compared to the Blasius solution for a laminar incompressible flat-plate boundary layer derived by H. Blasius in [6]. Specifically, we are interested in the distribution of the skin-friction coefficient c_f along the wall. This quantity c_f is defined as

$$c_f = \frac{\tau_W}{\frac{1}{2}\rho_\infty |\mathbf{v}_\infty|^2}, \tag{4.50}$$

The wall shear stress τ_W is defined in our case as

$$\tau_W = \mu \mathbf{t}.(\boldsymbol{\tau}\mathbf{n}), \tag{4.51}$$

where $\mathbf{t}$ and $\mathbf{n}$ is the tangent and normal to the wall surface , respectively, and $\boldsymbol{\tau}$ is the stress tensor with components τ_{ij} given in (4.2). For the Blasius solution, the distribution of the skin friction coefficient is given by

$$c_f^{exact} = 0.664 Re_x^{-1/2}, \text{ where } Re_x = \frac{|\mathbf{v}_\infty| x_1}{\mu}. \tag{4.52}$$

Figure 4.1 shows a detail of the velocity isolines of the computed numerical solution, while in Figure 4.2, the theoretical ($\circ\circ\circ$) and computed (——) skin friction coefficients are compared.

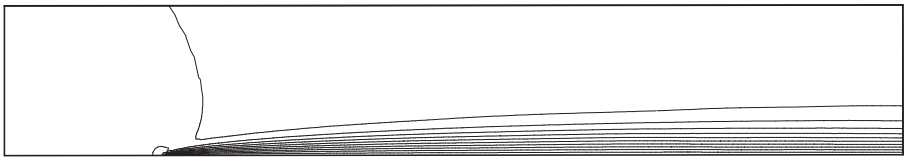


Figure 4.1: Laminar flat-plate boundary layer, velocity isolines.

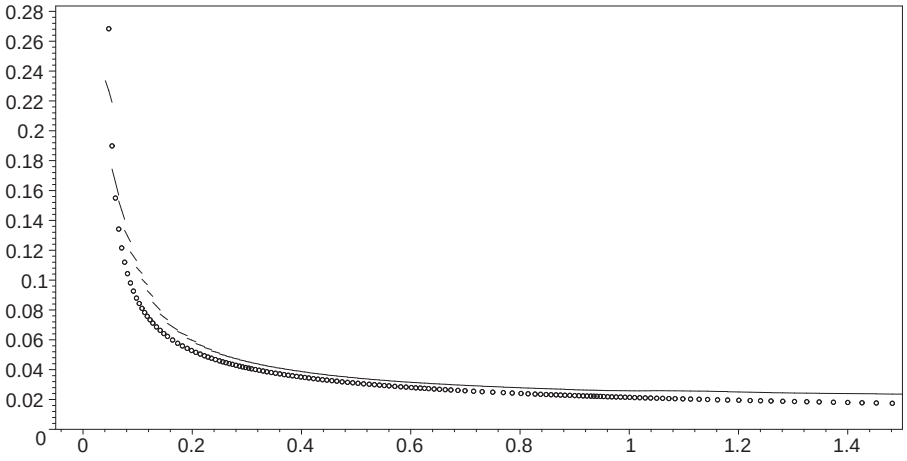


Figure 4.2: Laminar flat-plate boundary layer, distribution of the skin friction coefficient along the wall surface, ooo – exact Blasius solution, —— – approximate solution.

4.4.2 Channel flow

In this numerical example, we shall be concerned with the stationary flow through a straight narrow channel. From the theory of laminar incompressible viscous flow, the so-called *Poiseuille flow* described in [30], is known as the stationary solution of flow through an infinitely long channel $\Omega = \{(x_1, x_2) \in I\!R^2 : x_1 \in I\!R,\ 0 < x_2 < 1\}$ with walls at $\Gamma_W = \{(x_1, x_2) \in I\!R^2 : x_2 = 0 \ \vee \ x_2 = 1\}$. If we assume that the flow is stationary and derivatives of quantities are zero in the x_1-direction, we obtain as an exact solution the well known parabolic profile of the velocity and linearly descending pressure. The boundary layer does not affect the pressure in the sense that the derivative of the pressure in the x_2-direction is zero.

The computational domain is $\Omega_h = \{(x_1, x_2) \in I\!R^2 : -5 < x_1 < 5,\ 0 < x_2 < 1\}$, the mesh is nonuniform and is formed by 2131 triangular elements. At the inlet we prescribe a constant state, the freestream Mach number is $M = 0.1$ and the Reynolds number related to the thickness of the channel is $Re = 100$. Figure 4.3 shows velocity isolines (top) and pressure isolines (bottom) of the computed solution. The effect of the constant inlet boundary condition is visible at the inlet (left side) and at the outlet a small perturbation of the pressure is also

visible. In Figure 4.4, we see a section of the channel near the outlet along the line $x_1 = 4$. The parabolic profile of the velocity is clearly visible (solid line) and for comparison an exact parabolic profile fitted to the numerical solution is also plotted (circles).

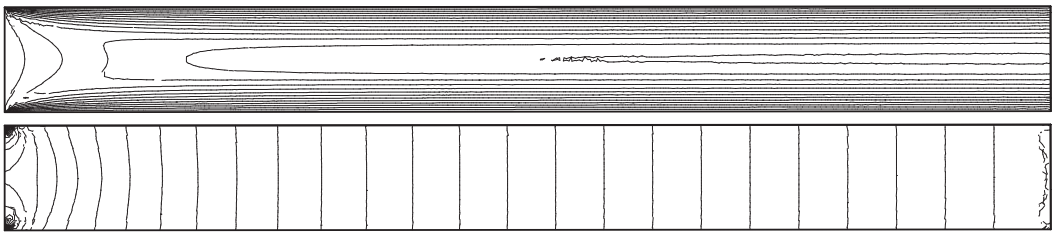


Figure 4.3: Poiseuille flow in channel, velocity isolines (top) and pressure isolines (bottom).

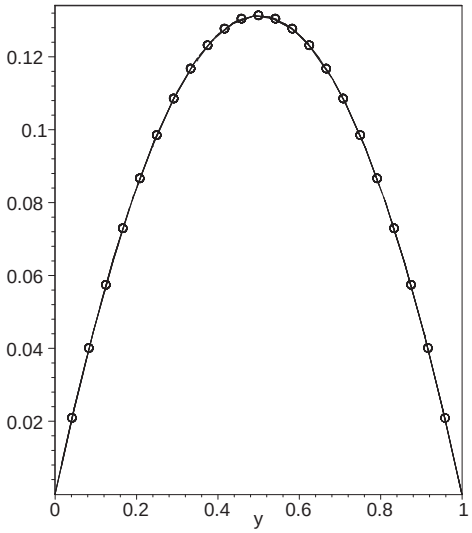


Figure 4.4: Poiseuille flow in channel, cross section near outlet, —— – values of velocity, ∘ ∘ ∘ – exact parabolic profile fitted to the numerical solution.

4.4.3 Flow past a NACA0012 profile

1) Stationary flow. Let us consider a viscous subsonic flow past the NACA0012 airfoil with small angle of attack (2°). In this case the flow is stationary, obtained by time stabilization with "$t_k \to \infty$". Here the upper and lower surfaces of the airfoil geometry are given by the functions $f^{\pm}$, respectively, where

$$f^{\pm}(\vartheta) = \pm 0.6 * (0.2969\sqrt{\vartheta} - 0.126\vartheta - 0.3516\vartheta^2 + 0.2843\vartheta^3 - 0.1015\vartheta^4)^+, \quad (4.53)$$

where $(.)^+$ denotes the positive part. This function is re-scaled in order to yield a chord of unit length. The far-field flow has Mach number $M = 0.5$, angle of attack $\alpha = 2^\circ$ and Reynolds number $Re = 5000$. The computational mesh has 2367 elements and is adaptively refined near the profile. Figure 4.5 – left – shows Mach number isolines, the boundary layer and wake behind the airfoil are visible. Figure 4.5 – right – shows pressure isolines. As in Section 4.4.1 in the case of a single impermeable wall, the boundary layer and wake should not be visible in the pressure distribution. Finally in Figure 4.6 the entropy is plotted. This should be produced only in the boundary layer and convected by the flow field.

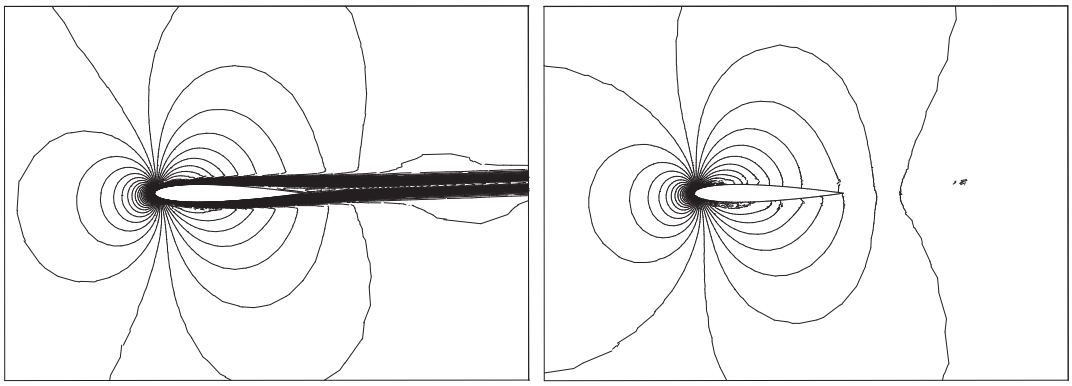


Figure 4.5: NACA0012 $\alpha = 2^\circ$ viscous flow, Mach number isolines (left), pressure isolines (right).

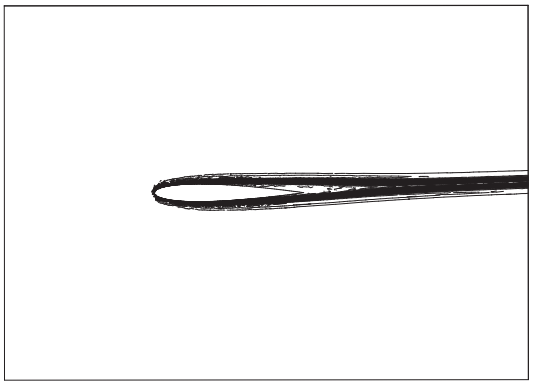


Figure 4.6: NACA0012 $\alpha = 2^\circ$ viscous flow, entropy isolines.

2) Nonstationary flow. We treat the compressible flow around a NACA0012 profile with a large angle of attack (25°). Unlike the preceding example, the flow is nonstationary with vortex formation and shedding at the upper wall of the profile. The far-field flow has Mach number $M = 0.5$, angle of attack $\alpha = 25^\circ$ and Reynolds number $Re = 5000$. The computational mesh has 2898 elements and is adaptively refined near the profile. Due to the nonstationary character

of the flow, the following figures illustrate the flow situation at time $t = 8.5$. Figure 4.7 shows a detail of the Mach number isolines, the boundary layer and complicated flow structure behind the airfoil are visible. In Figure 4.8 a detail of streamlines with the vortex structure at $t = 8.5$ is shown. Finally in Figure 4.9 we plot the entropy, which should be produced only in the boundary layer and convected by the flow field as in the previous stationary case.

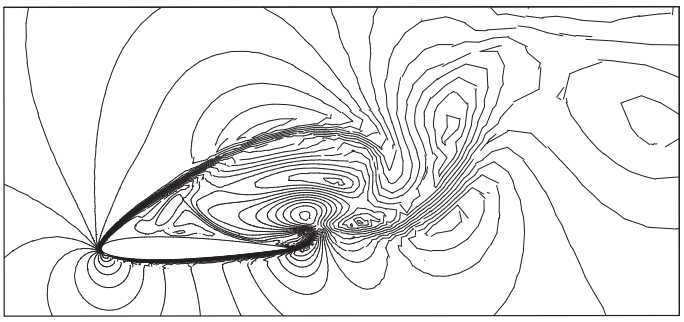


Figure 4.7: NACA0012 $\alpha = 25^\circ$ viscous flow, Mach number isolines.

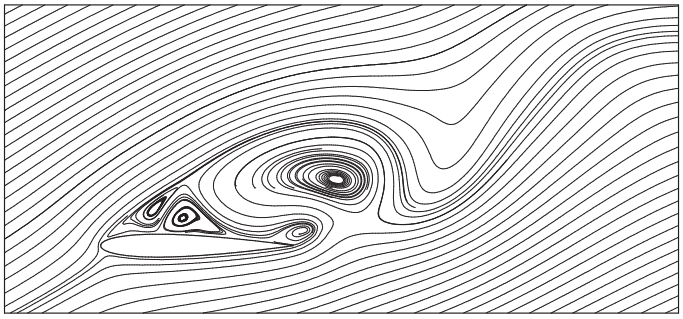


Figure 4.8: NACA0012 $\alpha = 25^\circ$ viscous flow, streamlines.

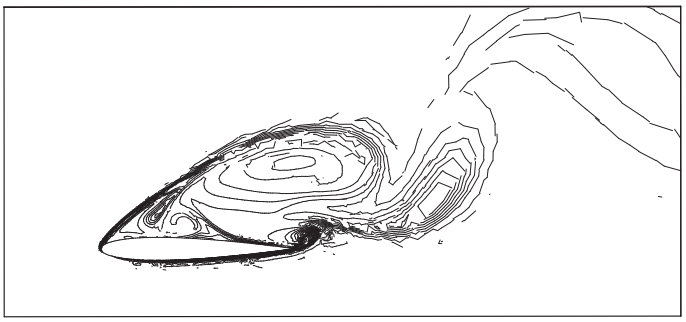


Figure 4.9: NACA0012 $\alpha = 25^\circ$ viscous flow, entropy isolines.

Conclusions

In the first two chapters of this thesis, we have formulated and theoretically analyzed the discontinuous Galerkin finite element method for a scalar convection-diffusion equation with nonlinear convection and diffusion. Error estimates in the $L^2(H^1)$- and $L^\infty(L^2)$-norms are derived, however these are suboptimal with respect to the latter norm. In Section 2.3 optimal error estimates in the $L^\infty(L^2)$-norm for the symmetric variant (SIPG) of the DGFEM are derived, but due to the use of specific techniques there are several limitations in this result. Namely, we assume that Ω is convex, we do not permit a Neumann boundary condition and the diffusion must be linear. Further work will be aimed at removing these additional assumptions. As for the nonsymmetric (NIPG) and incomplete (IIPG) variants, numerical experiments indicate that these methods may have $L^\infty(L^2)$-optimal convergence when the approximation order p is odd. To the authors knowledge, this phenomenon has not yet been theoretically analyzed.

The discontinuous Galerkin method combined with semi-implicit linearization applied to the Euler equations yields a accurate and robust method capable of solving a wide range of problems. In the computationally challenging case of small Mach number flows, the performance of the presented method relies mainly on the use of an efficient linear solver. In these cases we have mainly used the direct solver UMFPACK, which limits the size of the problems solved due to high memory consumption. Therefore further work must be invested in the development of an efficient preconditioner to be used in an iterative solver, e.g. GMRES. The shock capturing technique presented in Section 3.7 must be further refined to obtain a more robust method, especially with respect to the size of the time step and choice of the constants involved.

In the last chapter, we have applied the methods theoretically justified in chapters 1 and 2 to the compressible Navier-Stokes equations. Several extensions from the scalar case were discussed, including a new approach based on a unified methodology of [2]. Numerical experiments were performed using the incomplete interior penalty (IIPG) method, because of its simplicity and robustness with respect to the choice of the penalty parameter σ. Since the resulting matrices have a more complicated structure than those arising from the Euler equations, the direct linear solver was unable to compute solutions of larger problems, therefore an iterative solver was used, which has insufficient performance in the case of

small Mach numbers. Again the need for effective preconditioning arises.

We believe that the discontinuous Galerkin finite element method yields an effective higher order scheme for the solution of conservation laws and singularly perturbed problem due to its local character. Especially of interest in applications is the capability to solve compressible flows which are near the so-called incompressible limit as well as transonic and supersonic flows.

Bibliography

[1] Adjerid S., Devine D., Flaherty J. E., Krivodonova L.: *A posteriori error estimation for discontinuous Galerkin solutions of hyperbolic problems,* Comput. Methods Appl. Mech. Eng., 191 (2002), pp. 1097-1112.

[2] Arnold D. N., Brezzi F., Cockburn B., Marini L. D.: *Unified analysis of discontinuous galerkin methods for elliptic problems*, Siam J. Numer. Anal., Vol. 39, No. 5 (2002), pp. 17491779.

[3] Bassi F., Rebay S.: *A high-order accurate discontinuous finite element method for the numerical solution of compressible Navier-Stokes equations*, J. Comput. Phys 131 (1997): pp. 267-279.

[4] Bassi F., Rebay S.: *High-order accurate discontinuous finite element solution of the 2D Euler equations*, Journal of Computational Physics, v.138 n.2 (1997), pp. 251-285.

[5] Bejček M., Dolejší V., Feistauer M.: *On discontinuous Galerkin methods for numerical solution of conservation laws and convection-diffusion problems,* Proceedings of the XIVth Summer School Software and Algorithms of Numerical Mathematics, pp. 7-32. University of West Bohemia, Pilsen, 2002.

[6] Blasius, H.:, *Grenzschichten in Flüssigkeiten mit kleiner Reibung*, Z. Angew. Math. Phys. 56, 1908, pp. 1-37 (English translation in NACA Technical memo. 1256).

[7] Ciarlet, P. G.: *The Finite Elements Method for Elliptic Problems*, North-Holland, Amsterdam, New York, Oxford, 1979.

[8] Cockburn B., Shu C. W.: *TVB Runge-Kutta local projection discontinuous Galerkin finite element for scalar conservation laws II: General framework*, In *Math. Comp.,* 52 (1989): pp. 411-435.

[9] Cockburn B., Karniadakis G. E., Shu C. W. (ed.): *Discontinuous Galerkin Methods,* Number 11 in Lecture Notes in Computational Science and Engineering. Springer, Berlin, 2000.

[10] Cockburn B.: *Discontinuous Galerkin methods for convection dominated problems,* In *High-Order Methods for Computational Physics* (ed. T. J. Barth and H. Deconinck), Number 9 in Lecture Notes in Computational Science and Engineering, pp. 69-224. Springer, Berlin, 1999.

[11] Davis, T. A.: *A column pre-ordering strategy for the unsymmetric-pattern multifrontal method*, ACM Trans. Math. Software, vol 30. no 2, 2004, pp. 165-195.

[12] Davis, T. A.: *Algorithm 832: UMFPACK, an unsymmetric-pattern multifrontal method*, same issue, ACM Trans. Math. Software, vol 30. no 2, 2004, pp. 196-199

[13] Dolejší V., *Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes*, Comput. Vis. Sci. 1(3) (1998), pp. 165–178.

[14] Dolejší V., Feistauer M.: *Discontinuous Galerkin finite element method for convection-diffusion problems and the compressible Navier-Stokes equations*, The Preprint Series of the School of Mathematics MATH-KNM-2003/3, Charles University, Prague.

[15] Dolejší V., FeistauerM.: *Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems*, Numer. Funct. Anal. Optimiz., 26 (2005) pp. 349–383.

[16] Dolejší V., Feistauer M.: *A Semi-Implicit Discontinuous Galerkin Finite Element Method for the Numerical Solution of Inviscid Compressible Flow*, J. Comput. Phys. 198 (2004) pp. 727746.

[17] Dolejší V., Feistauer M., Kučera V., Sobotíková V.: *An optimal $L^\infty(L^2)$-error estimate for the discontinuous Galerkin approximation of a nonlinear non-stationary convectiondiffusion problem*, IMA Journal of Numerical Analysis (to appear).

[18] Dolejší V., Feistauer M., Schwab C.: *On some aspects of the discontinuous Galerkin finite element method for conservation laws*, Math. Comput. Simul. 61 (2003) pp. 333–346.

[19] Dolejší V., Feistauer M., Sobotíková V.: *Analysis of the discontinuous Galerkin method for nonlinear convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg. 194 no. 25-26 (2005), pp. 2709-2733.

[20] Feistauer, M.: *Mathematical Methods in Fluid Dynamics*, Longman Scientific & Technical, Harlow, 1993.

[21] Feistauer M., Felcman J., Straškraba I.: *Mathematical and Computational Methods for Compressible Flow*, Oxford University Press, Oxford, (2003).

[22] Feistauer, M.; Švadlenka, K.: *Discontinuous Galerkin method of lines for solving nonstationary singularly perturbed linear problems*, J. Numer. Math., 12 (2004), pp. 97-118.

[23] Flaherty J. E., Loy R. M., Shephart M. S., Teresco J. D.: *Software for the parallel adaptive solution of conservation laws by discontinuous Galerkin methods.* In *Discontinuous Galerkin Methods: Theory, Computation and Applications* (ed. B. Cockburn, G. E. Karniadakis and G. W. Shu), pp. 113-124. Springer, Berlin, 2000.

[24] Fraenkel L.: *On corner eddies in plane inviscid shear flow*, J. Fluid Mech 11 (1961), pp. 400–406.

[25] Grisvard, P.: *Singularities in Boundary Value Problems.* Springer, Berlin, 1992.

[26] Jaffre J., Johnson C., Szepessy A.: *Convergence of the discontinuous Galerkin finite elements method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci 5 (1995), pp. 367–386.

[27] Kučera V.: *Řešení stlačitelného proudění s malými Machovými čísly (Solution of Compressible Flow with low Mach Numbers)*, Diploma thesis, Faculty of Math. and Phys., Charles University, Prague 2004.

[28] A. Kufner, O. John, and S. Fučík: *Function Spaces*, Academia, Prague, 1977.

[29] Loitsianskii L. G.: *Mechanics of Fluids and Gases.* Nauka, Moscow, 1973 (in Russian).

[30] Poiseuille, J. L. M.: *Recherches expérimentelles sur le mouvement des liquids dans les tubes de trés petits diamétres*, Comptes Rendus, 11 (1840), pp. 961-967 and 1041-1048.

[31] Roe P. L.: *Approximate Riemann solvers, parameter vectors, and difference schemes*, J. Comp. Phys., 43 (1981), pp. 357–372.

[32] Schwab, C.: *p- and hp-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*, Clarendon Press, Oxford, 1998.

[33] Sobotíková V., Feistauer M.: *On the effect of numerical integration in the DGFEM for nonlinear convection-diffusion problems*, Numer Methods Partial Differential Eq (accepted), (2007).

[34] Šolín P., Segeth K., Doležel I.: *Higher order finite element methods*, Chapman & Hall, 2004.