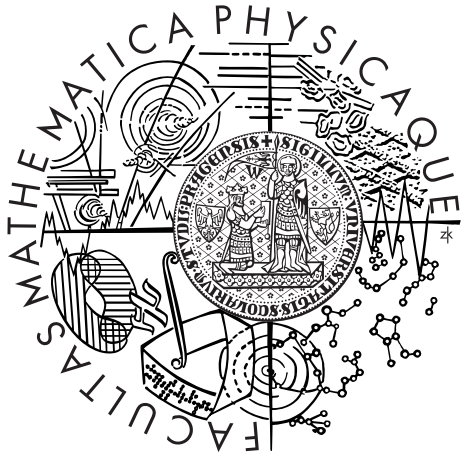
Charles University in Prague
Faculty of Mathematics and Physics

HABILITATION THESIS



Discontinuous Galerkin method: theory and applications

Václav Kučera

Mathematics
Mathematical modelling and numerical mathematics

Prague, August 2015

Contents

Preface

During the past two decades, the discontinuous Galerkin finite element method has become increasingly popular as a robust and high order numerical method for the solution of nonlinear partial differential equations of convective or convection-dominated character. Many advancements have been made in the theoretical analysis and practical application of the discontinuous Galerkin method to real world problems establishing the method as a competitive alternative to other approaches, especially the finite element and finite volume methods.

This thesis aims to present results I obtained within the last six years of my research dealing with the discontinuous Galerkin method. The main part of the thesis consists of five papers, three of which deal with theoretical analysis of the discontinuous Galerkin method, namely the derivation of a priori error estimates for various problems. These are the papers [33], [23] and [34], in chronological order. The remaining two papers deal with the numerical simulation of compressible fluid flow in time-dependent domains and its interaction with rigid or elastic structures. These are the papers [10] and [26]. All the papers upon which this thesis is based were published in foreign impacted journals in the years 2010–2014.

The thesis itself consists of an introductory Chapter 1, where the discontinuous Galerkin method is briefly introduced and the main ideas and contributions of the individual papers are outlined. The main part of the thesis consists of the papers themselves in the following order:

- Chapter 2: A priori optimal order $L^\infty(L^2)$-error estimates are derived for the discontinuous Galerkin method applied to a nonlinear convection-diffusion problem with nonlinear convection as well as diffusion using a nonlinear version of the Aubin-Nitsche technique, [33].

- Chapter 3: A priori error estimates for the space-time discontinuous Galerkin method applied to a convection-diffusion problem with linear diffusion are derived, [23].

- Chapter 4: A priori error estimates are derived for the discontinuous Galerkin method applied to a convection-diffusion problem that are uniform with respect to the diffusion parameter $\varepsilon \to 0$ and valid even in the purely convective case $\varepsilon = 0$, [34].

- Chapter 5: The arbitrary Lagrangian-Eulerian method is used to numerically solve compressible flow problems in time dependent domains using a semi-implicit discontinuous Galerkin method, [10].

- Chapter 6: The arbitrary Lagrangian-Eulerian formulation of flow problems is coupled with the equations of linear elasticity in order to simulate full fluid structure interaction, specifically a model problem for the simulation of voice formation in human vocal folds, [26].

The presented papers are included in this thesis as they were published, with only the text style being united to conform to the style of the thesis. Therefore each of the chapters must be, to some extent, viewed as an individual self-contained entity, otherwise various minor collisions of notation may occur due to the diversity of concepts and techniques covered in this thesis.

I would like to thank all my collaborators and colleagues, especially Miloslav Feistauer, Jaroslava Hasnedlová née Prokopová, Karel Najzar, Adam Kosík, Jan Česenek,

Jaromír Horáček and others for their support, stimulating discussions and work on joint projects.

Finally, I want to thank my family and friends for their support and encouragement, especially my wife Monika.

Prague, August 2015

Václav Kučera

1. Introduction

In many areas of applied mathematics in science and engineering, one encounters the need to solve partial differential equations of convective or convective-diffusive nature. Perhaps the most prominent of these fields is computational fluid dynamics, which plays an extremely important role in practical applications. Much work has been devoted to research in this area due to the mathematically and computationally challenging nature of the equations of fluid dynamics problems.

The area of compressible fluid dynamics is especially of interest due to the complicated phenomena encountered in the governing equations and their solutions. From the computational point of view, one of the main problems associated with compressible fluid dynamics is the rise of discontinuities (shock waves and contact discontinuities) or very steep gradients (internal or boundary layers) in the solutions, cf. [20], [36]. This behavior is typical of the wider class of (nonlinear) first order hyperbolic partial differential equations and their singular perturbations by diffusion.

The presence of discontinuities in the sought solutions is problematic from the point of view of numerical mathematics. High order numerical methods, such as the *finite element method* [12] in general suffer from the so-called Gibbs phenomenon manifested by the presence of spurious oscillations, overshoots and undershoots in the numerical solution. In the finite element method, which usually uses globally continuous (conforming) piecewise polynomial approximations, these problems are typically overcome by the use of suitable stabilization techniques (Streamline upwind Petrov Galerkin, Galerkin least squares, etc.) or layer adapted meshes, cf. [20], [46]. Another approach is the *finite volume method*, cf. [36], [8], where piecewise constant approximations are used. Such functions are naturally globally discontinuous on the given partition and are therefore more suitable for the approximation of discontinuous functions. The drawback of the finite volume method is that it is only first order accurate at most and the extension to orders higher than quadratic using for example reconstruction operators is problematic, [31], [36], [8].

The *discontinuous Galerkin* finite element method first developed for a neutron transport equation in [45], can be viewed as a generalization of the finite element and finite volume methods. The method uses higher order piecewise polynomial approximations that are globally discontinuous with respect to the given partition of the computational domain. The discontinuity of the discrete function space is taken into account by the use of so-called numerical fluxes to approximate the physical fluxes through interelement boundaries, similarly as in the finite volume method. This means that arbitrarily high orders of accuracy can be obtained, while the discontinuous nature of the discrete solution helps to alleviate the Gibbs phenomenon compared to the conforming finite element method. Unlike the finite element method, where the Gibbs phenomenon eventually pollutes the entire computational domain, in the discontinuous Galerkin method the oscillations remain localized in the vicinity of the discontinuity. They can then be effectively removed using e.g. limiting techniques (ENO, WENO, etc., cf. [25], [37], [36], [31]) or artificial diffusion and shock capturing techniques, cf. [21], [42], [29]. The drawback of the discontinuous Galerkin method is the increased number of degrees of freedom as compared to the finite element method.

The discontinuous Galerkin method was first analyzed in the papers [35], [30], [2] and in the papers [3], [4], [6] the method is analyzed for elliptic problems. The first three papers included in the main part of this thesis, i.e. Chapters 2, 3 and 4, deal with the theoretical analysis of the discontinuous Galerkin method for nonlinear convection-diffusion equations, namely a priori error estimates for smooth solutions are derived. Many papers have been written on this subject, however the three presented papers

build on and naturally extend the results of [14], [15] and [16]. In these papers, a priori error estimates of optimal orders in the $L^\infty(L^2)$- and $L^2(H^1)$-norms are derived for a scalar nonstationary nonlinear convection-diffusion problem with linear diffusion. In the paper [22], suboptimal $L^\infty(L^2)$ error estimates are derived for convection-diffusion problem with nonlinear diffusion. In the presented thesis, these results are generalized to optimal order $L^\infty(L^2)$ error estimates for the nonlinear diffusion case (Chapter 2) and the analysis of a space-time discontinuous Galerkin method for the considered problem (Chapter 3). Finally, using the ideas of [50], in Chapter 4, error estimates for the convection-diffusion with nonlinear convection and linear diffusion are derived that are uniform with respect to the diffusion coefficient $\varepsilon \geq 0$.

The last two chapters of this thesis deal with the practical application of the discontinuous Galerkin method to the problem of simulation of compressible flow interaction with rigid and elastic bodies. The discontinuous Galerkin method is becoming increasingly popular in the computational compressible fluid dynamics community with many groups working on the development and implementation of efficient and accurate algorithms. Here we only mention, as an example, the papers [7], [9] and [48]. Chapters 5 and 6 use the semi-implicit discontinuous Galerkin scheme of [21] along with the arbitrary Lagrangian-Eulerian method, [40], in order to simulate compressible fluid flow in time-dependent domains. In Chapter 5, the movement of the computational domain is either prescribed (model of air flow through the human vocal folds) or described by a simple system of ordinary differential equations (interaction of air flow and an elastically supported aerodynamic profile). In the final Chapter 6, these results are generalized to include the interaction of air flow and an isotropic elastic body described by the equations of dynamic elasticity and generalized Hooke's law with the aim of simulating true fluid-structure interaction in a simplified model of the human glottis.

The five papers contained in this thesis demonstrate that the discontinuous Galerkin method is a robust and accurate numerical method for the solution of partial differential equations of convection-diffusion character, which has solid mathematical foundations.

1 Discontinuous Galerkin method

In this short section, we shall briefly introduce the basic concepts of the discontinuous Galerkin method and its formulation for the type of problems considered in the main part of this thesis. We present only the necessary minimum of notation and concepts in order to formulate the method – many technical subtleties and notions will be omitted for brevity, since they can be found in the subsequent chapters of the main part of the thesis, which will be refered to frequently.

The discontinuous Galerkin method is best suited for the numerical solution of advective or convective problems in conservative form. If, for simplicity, we consider the scalar case, such a problem reads: Let $\Omega \subset \mathbb{R}^d, d \in \mathbb{N}$, be a bounded open polygonal (polyhedral) domain with Lipschitz-continuous boundary $\partial\Omega$. Find $u : \Omega \times (0,T) \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) = g \quad \text{in } \Omega \times (0,T). \tag{1}$$

Since problem (1) is an evolutionary partial differential equation, it must be equipped with an initial condition and appropriate boundary conditions, cf. [20], [36]. From the theoretical and practical point of view, the more interesting case is when $\mathbf{f} : \mathbb{R} \to \mathbb{R}^d$, representing the convective or advective terms is nonlinear. The papers included in this thesis are concerned with this case. Problem (1) represents a general conservation law for the conserved quantity u and depending on the specific form of f, can describe such

phenomena as fluid flows, city traffic, electrons in semiconductors or elastic waves in solids, cf. [36]. The function g, usually equal to zero, is a prescribed right-hand side.

In many applications, diffusion enters the process, which leads to the convection-diffusion problem

$$\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) - \varepsilon \Delta u = g \quad \text{in } \Omega \times (0,T), \tag{2}$$

where $\varepsilon > 0$ represents the constant diffusion parameter. Typically in the discontinuous Galerkin method, one is interested in the convection-dominated case $\varepsilon \ll 1$. From the practical point of view, in Chapters 5 and 6 we shall be concerned with the compressible Navier-Stokes equations, for which (2) represents a simplified model with linear diffusion terms.

Discrete space

The discontinuous Galerkin method is similar to the finite element method in that it uses a suitable weak form of (1), but also the finite volume method since it uses piecewise polynomial approximations. As in both of these methods, we use a triangulation $\mathcal{T}_h$ of Ω, i.e. a partition into closed simplexes with mutually disjoint interiors. Here

$$h = \max_{K \in \mathcal{T}_h} \operatorname{diam} K \tag{3}$$

is the parameter used to measure the convergence rate of the method, as in the finite volume or finite element methods, cf. [12].

Using the partition $\mathcal{T}_h$, the discontinuous Galerkin method seeks a suitable approximation of u from the space of globally discontinuous, piecewise polynomial functions:

Definition 1. *We define the discrete space of discontinuous, piecewise polynomial functions*

$$S_h = \{v;\, v|_K \in P^p(K), \forall K \in \mathcal{T}_h\},$$

where $P^p(K)$ is the space of all polynomials on K of degree less than or equal to p.

The finite element method uses a similar function space, however with the added assumption of continuity:

$$V_h = \{v \in C(\overline{\Omega});\, v|_K \in P^p(K), \forall K \in \mathcal{T}_h\}, \tag{4}$$

cf. [12]. The discontinuity of functions from S_h is the main advantage of the discontinuous Galerkin method applied to (1) when approximating discontinuities or steep gradients. However, one needs to use a more complicated weak form of the governing equation than in the finite element method. This is true especially for the diffusion terms in (2), cf. [4].

Because we deal with discontinuous approximations, suitable notation is required. Since this chapter is only a short introduction, we refer to Chapters 2, 3 and 4 for the full details and technicalities of the derivation and notation. By $\mathcal{F}_h$, we denote the set of all faces (edges in 2D) of $\mathcal{T}_h$. For each $\Gamma \in \mathcal{F}_h$, we define an arbitrary but fixed unit normal vector $\mathbf{n}$. For $v \in S_h$, $v^{(L)}$ and $v^{(R)}$ represent left and right traces of the discontinuous function $v \in S_h$ on a given face $\Gamma \in \mathcal{F}_h$ with respect to the orientation of the normal $\mathbf{n}$. Finally, $[v] = v^{(L)} - v^{(R)}$ is the *jump* of v on $\Gamma \in \mathcal{F}_h$ and $\langle v \rangle = \frac{1}{2}(v^{(L)} + v^{(R)})$ is the *average* of v on Γ.

1.1 Discrete formulation of convective problems

Similarly as in the finite element method, the discrete formulation of problem (1) is obtained by multiplying the equation by a test function φ, integrating over an element $K \in \mathcal{T}_h$ and applying Green's theorem. After summing over all elements, we obtain the following, cf. [20], [21].

Definition 2. *We say that $u_h \in C^1([0,T]; S_h)$ is a discontinuous Galerkin solution of problem (1) if for all $t \in (0,T)$ and all $\varphi_h \in S_h$*

$$\frac{d}{dt}\big(u_h(t), \varphi_h\big) + b_h\big(u_h(t), \varphi_h\big) = \big(g(t), \varphi_h\big), \tag{5}$$

where the convective form $b_h(\cdot,\cdot)$ is defined by

$$b_h(v_h, \varphi_h) = -\sum_{K \in \mathcal{T}_h} \int_K \mathbf{f}(v_h) \cdot \nabla \varphi_h \,\mathrm{d}x + \int_{\mathcal{F}_h} H(v_h^{(L)}, v_h^{(R)}, \mathbf{n})[\varphi_h] \,\mathrm{d}S. \tag{6}$$

In the definition of b, in the second term the integration is performed over all interelement faces. The function H being integrated is the so-called *numerical flux*, which approximates the physical flux $\mathbf{f}(u)\cdot\mathbf{n}$ through each edge Γ using the two traces $u_h^{(L)}, u_h^{(R)}$ of the discrete solution. This term arises by the element-wise application of Green's theorem due to the discontinuity of $u_h(t), \varphi_h \in S_h$. We note that the second term in (6) is not present in the finite element method, since in this case $u_h(t), \varphi_h \in V_h$ are continuous functions and therefore $[\varphi_h] = 0$. Since S_h is finite-dimensional, (5) represents a system of ordinary differential equations, which must be equipped with an appropriate initial condition $u_h^0 \in S_h$.

Numerical flux

The numerical flux H is an important ingredient of the discontinuous Galerkin method (5), (6). The concept is well known and studied in the finite volume method, therefore one can use one of the many available constructions of numerical fluxes known from the finite volume literature, cf. [36].

In the scalar case, $H(u,v,\mathbf{n})$ is defined in $\mathbb{R}^2 \times B_1$, where $B_1 = \{\mathbf{n} \in \mathbb{R}^d; |\mathbf{n}| = 1\}$. From the analytic point of view, natural assumptions on H are the following, cf. Chapters 2 and 3:

(H1) $H(u,v,\mathbf{n})$ is *Lipschitz-continuous* with respect to u, v whenever $\mathbf{f}$ is Lipschitz-continuous:

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

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

$$H(u,u,\mathbf{n}) = \mathbf{f}(u)\cdot\mathbf{n}, \quad u \in \mathbb{R},\ \mathbf{n} \in B_1.$$

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

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

Assumptions (H1), (H2) and (H3) are essential to any analysis of the discontinuous Galerkin method and are used in Chapters 2 and 3. In Chapter 4, a more subtle analysis of the convective terms is performed and the *E-flux* property is additionally assumed. This will be discussed in more detail in Section 4 of this introduction.

1.2 Discrete formulation of convection-diffusion problems

In Section 1.1, the discontinuous Galerkin formulation of (1) is briefly outlined. However, except for Chapter 4, the bulk of this thesis deals with the convection-diffusion case (2) either theoretically or practically. The discontinuous Galerkin discretization of the diffusion term $-\varepsilon\Delta u$ is rather technical and lengthy in full detail, therefore here we shall only introduce the final form and refer to Chapters 2, 3 and 4 for details, cf. also the fundamental paper [4].

Definition 3. *We say that $u_h \in C^1([0,T]; S_h)$ is a discontinuous Galerkin solution of problem (1) if for all $t \in (0,T)$ and all $\varphi_h \in S_h$*

$$\frac{d}{dt}\big(u_h(t),\varphi_h\big) + b_h\big(u_h(t),\varphi_h\big) + \varepsilon J_h\big(u_h(t),\varphi_h\big) + \varepsilon a_h\big(u_h(t),\varphi_h\big) = l_h\big(\varphi_h\big)(t), \tag{7}$$

where the diffusion form $a_h(\cdot,\cdot)$ is defined by

$$\begin{aligned} a_h(v_h,\varphi_h) = \sum_{K\in\mathcal{T}_h}\int_K \nabla v_h\cdot\nabla\varphi_h\,\mathrm{d}x - \int_{\mathcal{F}_h^I}\langle\nabla v_h\rangle\cdot\mathbf{n}[\varphi_h]\,\mathrm{d}S - \Theta\int_{\mathcal{F}_h^I}\langle\nabla\varphi_h\rangle\cdot\mathbf{n}[v_h]\,\mathrm{d}S \\ - \int_{\mathcal{F}_h^D}\nabla v_h\cdot\mathbf{n}\varphi_h\,\mathrm{d}S - \Theta\int_{\mathcal{F}_h^D}\nabla\varphi_h\cdot\mathbf{n}v_h\,\mathrm{d}S, \end{aligned} \tag{8}$$

the interior and boundary penalty jump terms are defined by

$$J_h(v_h,\varphi_h) = \int_{\mathcal{F}_h^I}\sigma[v_h][\varphi_h]\,\mathrm{d}S + \int_{\mathcal{F}_h^D}\sigma v_h\varphi_h\,\mathrm{d}S \tag{9}$$

and the right-hand side form is

$$l_h(\varphi_h)(t) = \int_\Omega g(t)\varphi_h\,\mathrm{d}x + \int_{\mathcal{F}_h^N} g_N(t)\varphi_h\,\mathrm{d}S + \varepsilon\int_{\mathcal{F}_h^D}\sigma u_D(t)\varphi_h - \Theta\nabla\varphi_h\cdot\mathbf{n}u_D(t)\,\mathrm{d}S. \tag{10}$$

Without going into full detail, in (8)–(10), $\mathcal{F}_h^I, \mathcal{F}_h^D$ and $\mathcal{F}_h^N$ denote the sets of edges lying in the interior of Ω, on the part of the boundary $\partial\Omega$ corresponding to Dirichlet boundary conditions and to Neumann boundary conditions, respectively. By u_D and g_N, we denote the corresponding Dirichlet and Neumann boundary data, respectively. The parameter σ in (9) and (10) is constant on every edge and defined by

$$\sigma|_\Gamma = \frac{C_W}{|\Gamma|}, \quad \forall\ \Gamma\in\mathcal{F}_h, \tag{11}$$

where $C_W > 0$ is a suitably chosen constant. Finally, the parameter Θ is typically taken as the $\Theta = 1, 0, -1$, leading to the *symmetric, incomplete* and *nonsymmetric interior penalty* variants of the discontinuous Galerkin method, respectively.

As is the case with the convective form b_h, if we instead consider the classical finite element case, i.e. $u_h(t), \varphi_h \in V_h$, then again $[u_h(t)] = [\varphi_h] = 0$ on each Γ due to continuity and the Dirichlet boundary condition is exactly satisfied. Therefore a_h reduces only to its first term, J_h is identically zero and l_h reduces to its first two terms, therefore we obtain the classical weak formulation of (2), cf. [12].

As in the analysis of the finite element method, the diffusion terms $A_h(v,w) := a_h(v,w) + J_h(v,w)$ are shown to be elliptic and bounded in an appropriate "energy" norm, in our case the so-called *DG norm*

$$\|w\|_{DG} = \Big(\sum_{K\in\mathcal{T}_h}|w|^2_{H^1(K)} + J_h(w,w)\Big)^{1/2}. \tag{12}$$

The ellipticity and boundedness of A_h holds provided the constant C_W in (11) is large enough, cf. Chapter 3, Section 4.2 and Chapter 2, Lemma 6 for the nonlinear version of the diffusion term.

2 Overview of Chapter 2: Optimal $L^\infty(L^2)$-error estimates for nonlinear convection-diffusion problems.

Chapter 2 consists of the paper *Optimal $L^\infty(L^2)$-error Estimates for the DG Method Applied to Nonlinear Convection-Diffusion Problems with Nonlinear Diffusion*, published in the journal Numerical Functional Analysis and Optimization in 2010, [33]. This paper deals with the analysis of a generalized version of (2) where the diffusion term is nonlinear:

$$\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) - \operatorname{div}\big(\beta(u)\nabla u\big) = g. \tag{13}$$

Here $\beta(u)$ represents the diffusion coefficient dependent on the solution u. If $\beta(u) := \varepsilon$, a constant, we obtain problem (2) as a special case with linear diffusion. The analysis of (13) is challenging because both the convective and diffusive terms are nonlinear. For the analysis presented in Chapter 2, we need the following assumptions on the nonlinearity β:

$$\begin{aligned} \beta : \mathbb{R} \to [\beta_0, \beta_1], \quad 0 < \beta_0 < \beta_1 < \infty, \\ |\beta(u_1) - \beta(u_2)| \le L|u_1 - u_2|, \quad \forall u_1, u_2 \in \mathbb{R}, \end{aligned} \tag{14}$$

where the first assumption is used to obtain monotonicity-type estimates, while the other induces Lipschitz continuity of the resulting forms.

Chapter 2 is concerned with the derivation of a priori error estimates for the discontinuous Galerkin method applied to problem (13). The numerical scheme analyzed is therefore (7) with the diffusion form $a(\cdot,\cdot)$ modified to discretize the nonlinear diffusion term, cf. Chapter 2 for details. If we denote the error of the method as $e_h := u - u_h$, in a priori error analysis of evolutionary problems one is typically interested in estimates of the type

$$\|e_h\|_{L^\infty(0,T;L^2(\Omega))} = \sup_{t\in(0,T)} \|e_h(t)\|_{L^2(\Omega)} \le Ch^\mu, \tag{15}$$

where μ is typically p or $p+1$ and C is a constant independent of h. Estimate (15) therefore gives us the convergence rate of the method with respect to $h \to 0$.

2.1 Aubin-Nitsche technique and the A_h-projection

Problem (13) was already analyzed in the author's doctoral thesis [32], cf. also [22]. Error estimates of the suboptimal order $\mu = p$ were obtained in the general nonlinear case (13). Furthermore, using the so-called Aubin-Nitsche technique ([5], [39]), optimal error estimates of order $\mu = p+1$ were obtained for problem (2), i.e. with linear diffusion. The key problem in the application of the Aubin-Nitsche technique is that it requires the use of a dual problem corresponding to the diffusion terms. Formulating such a dual problem in the nonlinear case is not straightforward, since in this case, test functions in the weak formulation will formally end up inside the nonlinearity.

In the case of linear diffusion, the dual problem which enables the analysis of the scheme (7) is: Given $z \in L^2(\Omega)$ find $\psi \in H_0^1(\Omega)$ such that

$$-\Delta\psi = z \tag{16}$$

in the weak sense. One then uses the assumption that for Ω convex, $\psi \in H^2(\Omega)$ and uses this regularity to obtain the missing order of h in estimate (15). For nonlinear diffusion, the situation is more complicated, since even if some formal form of the nonlinear dual problem was used, results on the $H^2(\Omega)$ regularity of ψ are readily available only in the linear case, cf. [24]. In order to avoid these obstacles, in Chapter 2 a linearized version

of the dual problem is used: Given $z \in L^2(\Omega)$ and $t \in (0,T)$, find $\psi(t) \in H_0^1(\Omega)$ such that

$$-\mathrm{div}\big(\beta(u(t))\nabla\psi(t)\big) = z. \tag{17}$$

This problem is linear in $\psi(t)$ and if sufficient regularity of u is assumed, one can prove that $\psi(t) \in H^2(\Omega)$ using results for the Poisson problem (16), cf. Chapter 2, Lemma 9. If the "dual" problem (17) is used, one also needs the regularity of the derivative $\frac{\partial\psi(t)}{\partial t} \in H^2(\Omega)$ with respect to the parameter t. This purely technical, yet essential result is proved in Chapter 2, Lemma 12, using difference approximations to the derivative $\frac{\partial\psi(t)}{\partial t}$.

Another important ingredient in the Aubin-Nitsche technique is the use of a suitable Ritz or A_h-projection of $u(t)$ onto the space S_h. In the case of linear diffusion, we seek $u^\star(t) \in S_h$ such that

$$A_h\big(u^*(t),\varphi_h\big) = A_h\big(u(t),\varphi_h\big) \quad \forall\,\varphi_h \in S_h, \tag{18}$$

where $A_h(v,w) := a_h(v,w) + J_h(v,w)$ represents all terms related to the discretization of the diffusion term. One then uses the dual problem to show that the error of the A_h-projection $\chi(t) = u(t) - u^\star(t)$ is of the order $O(h^{p+1})$ in the $L^2(\Omega)$ norm.

Without going into technical details, in Chapter 2 the A_h-projection is again constructed using a linearized version of the discrete form a_h, where the linearization is carried out by replacing all arguments in the function β by the exact solution $u(t)$, similarly as in (17), cf. Chapter 2, Section 5.1. Finally, using the linearized "dual" problem (17), $O(h^{p+1})$ approximation properties for this linearized A_h-projection are proved in Chapter 2, Lemmas 11 and 13.

After further technical estimates of the convection and diffusion forms b_h and a_h, the final result of Chapter 2 is proved in Theorem 18, the optimal-order error estimate

$$\|e_h\|_{L^\infty(0,T;L^2(\Omega))} \le Ch^{p+1}. \tag{19}$$

We note that the final result is derived under standard regularity assumptions $\frac{\partial u}{\partial t} \in L^2\big(0,T;H^{p+1}(\Omega)\big)$, along with additional regularity assumptions required throughout the analysis. Namely $\nabla u(t)$, $\frac{\partial u}{\partial t}(t)$ and $\nabla\frac{\partial u}{\partial t}(t)$ are assumed bounded in $L^\infty(\Omega)$ for a.a. $t \in (0,T)$. These additional regularity assumptions are not needed in the analysis of the linear diffusion case.

3 Overview of Chapter 3: Analysis of space-time discontinuous Galerkin method.

In Chapter 3, the paper *Analysis of space-time discontinuous Galerkin method for nonlinear convection-diffusion problems* is presented, which was published in the journal Numerische Mathematik in 2011, [23]. The goal of the paper is the analysis of the space-time discontinuous Galerkin method applied to problem (2). Up to now, we have only considered the spatial discretization of our problem. This so-called space semidiscretization leads to the system of ordinary differential equations (7). One can then apply one of the many numerical methods to discretize this system with respect to time, for example in Chapter 4, the implicit or backward Euler method is considered and analyzed. Another possibility is to view (2) as an equation in the entire space-time domain $Q_T := \Omega \times (0,T)$ and to discretize using the discontinuous Galerkin method with respect to space and time simultaneously. This leads to the *space-time discontinuous Galerkin method*. The analysis builds on and generalizes ideas from the works [1], [19], [47].

3.1 Space-time discontinuous Galerkin method

As in the previous section, we shall only outline the main points of the discrete space-time formulation and refer to Chapter 3, Section 3 for details. We consider a time partition of $(0,T)$ into disjoint intervals: $[0,T] = \cup_{i=1}^{M} \overline{I_m}$ where $I_m = (t_{m-1}, t_m)$. We denote $\tau_m = t_m - t_{m-1}$ and $\tau = \max_{m=1,\dots,M} \tau_m$ is the parameter with respect to which the temporal error is measured. For every I_m we consider a triangulation $\mathcal{T}_{h,m}$ of Ω. Each $\mathcal{T}_{h,m}$ generates a different discrete space S_h, cf. Definition 1, which we shall denote $S^p_{h,m}$. The approximate solution will then be sought in the space

$$S^{p,q}_{h,\tau} = \Big\{ \varphi \in L^2(Q_T); \varphi\big|_{I_m}(t) = \sum_{i=0}^{q} t^i \, \varphi_i, \text{ where } \varphi_i \in S^p_{h,m}, \ m = 1, \dots, M \Big\}. \tag{20}$$

This space is finite-dimensional and consists of piecewise polynomial functions of degree at most p in space and q in time on the space-time partition induced by all I_m and $\mathcal{T}_{h,m}$. Since $\varphi \in S^{p,q}_{h,\tau}$ is discontinuous with respect to time at each t_m, we define the one-sided limits and jump of φ at t_m as

$$\varphi^{\pm}_m = \varphi(t_m \pm) = \lim_{t \to t_m \pm} \varphi(t), \quad \{\varphi\}_m = \varphi(t_m +) - \varphi(t_m -). \tag{21}$$

The discrete forms b_h, a_h, J_h and l_h remain the same as in Definitions 2 and 3, however since each I_m has a different triangulation $\mathcal{T}_{h,m}$ in general, the forms b_h etc. differ on each I_m. This is taken into account by the notation $b_{h,m}$ etc. Again, we define $A_{h,m}(v,w) := a_{h,m}(v,w) + J_{h,m}(v,w)$. Similarly, the DG-norms (12) also depend on the triangulation $\mathcal{T}_{h,m}$ and are thus denoted $\|\cdot\|_{DG,m}$.

Definition 4. *We say $U \in S^{p,q}_{h,\tau}$ is the space-time discontinuous Galerkin solution of problem (2), if for all $\varphi \in S^{p,q}_{h,\tau}$ and $m = 1, \dots, M$*

$$\int_{I_m} \big((U', \varphi) + A_{h,m}(U, \varphi) + b_{h,m}(U, \varphi) \big) \, \mathrm{d}t + \big(\{U\}_{m-1}, \varphi^+_{m-1} \big) = \int_{I_m} \ell_{h,m}(\varphi) \, \mathrm{d}t, \tag{22}$$

where $U^-_0 \in S^p_{h,m}$ is an approximation of the initial condition to problem (2).

The derivation of (22) follows the same lines as the derivation of (5) or (7): We multiply (2) by a test function $\varphi \in S^{p,q}_{h,\tau}$, integrate over a space-time element $K \times I_m$, where $K \in \mathcal{T}_{h,m}$, and apply Green's theorem in space and twice in time. The term $\big(\{U\}_{m-1}, \varphi^+_{m-1} \big)$ plays a similar role as the numerical flux or the interior and boundary penalty terms in the spatial discretization. Since U is discontinuous at t_m, one needs to prescribe in some "weak" sense the "initial condition" U^-_{m-1} for $U|_{I_m}$. The mentioned term does this by so-called penalization, cf. [4].

We note that in the special case $q = 0$, (22) reduces to a variant of the implicit Euler scheme, which is analyzed in Chapter 4.

3.2 Analysis of the space-time discontinuous Galerkin method

The goal of Chapter 3 is to derive a priori error estimates for scheme (22) in the $L^2(0,T;L^2(\Omega))$ and $L^2(0,T;H^1(\Omega))$ norms. As usual in a priori error analysis, we split the error $e = U - u$ into two parts $e = \xi + \eta$, where $\xi = U - \pi u \in S^{p,q}_{h,\tau}$ and $\eta = \pi u - u$, where $\pi : L^2(Q_T) \to S^{p,q}_{h,\tau}$ is a suitably chosen projection operator. For the purposes of Chapter 3, π is constructed so that

$$\begin{aligned} &(1) \quad (\pi v)(t_m -) = \Pi_m v(t_m -), \quad \forall\, m = 1, \dots, M. \\ &(2) \quad \int_{I_m} (\pi v - v, \varphi^*) \, \mathrm{d}t = 0, \quad \forall\, \varphi^* \in S^{p,q-1}_{h,\tau}, \quad \forall\, m = 1, \dots, M, \end{aligned} \tag{23}$$

where Π_m is the $L^2(\Omega)$-projection onto the space $S^p_{h,m}$.

First, an "abstract" estimate of the error e is derived in Chapter 3, Section 4. Equation (22) and the corresponding weak form of (2) are subtracted and tested by $\varphi := \xi$. Individual terms in this error equation are estimated using their ellipticity and boundedness properties, while special care is taken to estimate the evolutionary terms. Specifically, an important ingredient in the analysis is the construction of π by (23). Since then $\int_{I_m}(\eta, \xi')\,\mathrm{d}t = 0$ and $\left(\eta_m^-, \xi_m^-\right) = 0$ where the first term if nonzero would yield suboptimal orders of convergence when estimated directly. Finally, one obtains estimate (47), where $\left\|\xi_m^-\right\|^2$ figures in the left-hand side, while on the right-hand side we have terms containing the interpolation error expressed by η and $\int_{I_m}\|\xi\|^2\,\mathrm{d}t$. This latter term is undesired and must be eliminated.

Estimation of $\int_{I_m}\|\xi\|^2\,\mathrm{d}t$ and bounds on the interpolation error.

The goal of Chapter 3, Section 4.4 is the estimation of $\int_{I_m}\|\xi\|^2\,\mathrm{d}t$ in terms of $\left\|\xi_m^-\right\|^2$ and η (Lemma 5). This is done using a classical procedure, cf. [1], in which the error equation is tested by $\varphi := \tilde{\xi}$, where $\tilde{\xi}$ is the Lagrange interpolation of $(t_m - t_{m-1})\xi(t)/(t - t_{m-1})$ at the right Radau quadrature points on I_m, cf. [1]. Since the right Radau quadrature formulas are exact for polynomials of order up to $2q$, they integrate exactly terms such as $\int_{I_m}\|\xi\|^2\,\mathrm{d}t$, which can therefore be expressed as finite sums and estimated more straightforwardly.

Combining all these estimates gives us Theorem 6, i.e. the abstract estimate (here in simplified form):

$$\max_{m=1,\dots,M}\|e_m^-\|^2 + \frac{\varepsilon}{2}\sum_{m=1}^{M}\int_{I_m}\|e\|^2_{DG,m}\,\mathrm{d}t \le CR(\eta), \tag{24}$$

where C is independent of h, τ and $R(\eta)$ depends only on η.

In Chapter 3, Section 5, the quantity $R(\eta)$ from (24) is estimated in terms of the convergence parameters h and τ. For this purpose, η is written as $\eta = \eta^{(1)} + \eta^{(2)}$, where $\eta^{(1)} = \Pi_m u - u$ and $\eta^{(2)} = \pi u - \Pi_m u = \pi u - \pi(\Pi_m u)$ on each I_m. The estimation of $R(\eta)$ thus reduces to bounding $\eta^{(1)}$ and $\eta^{(2)}$ in various norms.

Estimates for $\eta^{(1)}$ and $\eta^{(2)}$ are obtained using approximation properties of the $L^2(\Omega)$-projection operator Π_m and the interpolation operator π. To this end, π is expressed using a one-dimensional interpolation operator $\tilde{P}_m$, cf. Chapter 3, Lemma 7. Fundamental approximation properties of $\tilde{P}_m$, hence π, are proven in Lemma 8 using the theory of polynomial preserving operators, cf. [12]. For completeness, a self-contained proof of Lemma 8 is provided in the Appendix of Chapter 3.

Finally, combining all the derived estimates gives us the main theorem of Chapter 3, Theorem 12 (here again in concise form):

$$\max_{m=1,\dots,M}\|e_m^-\|^2 + \frac{\varepsilon}{2}\sum_{m=1}^{M}\int_{I_m}\|e\|^2_{DG,m}\,\mathrm{d}t \le C(h^{2p} + \tau^{2q+2}). \tag{25}$$

Since in (25), the $L^2(\Omega)$ estimates are only in the endpoints t_m-, in Section 5.5 an $L^2(Q_T)$-bound is derived:

$$\|e\|^2_{L^2(Q_T)} \le C(h^{2p} + \tau^{2q+2}). \tag{26}$$

The estimates (26), (25) are derived under the assumption $0 < \tau_m \le C^\star\varepsilon$ and $\tau_m \ge Ch_m^2$. The latter condition on τ_m is not necessary if all the triangulations $\mathcal{T}_{h,m}$ are identical, cf. Section 5.4.

4 Overview of Chapter 4: Diffusion-uniform error estimates for singularly perturbed problems

Chapter 4 consists of the paper *On diffusion-uniform error estimates for the DG method applied to singularly perturbed problems*, published in the IMA Journal of Numerical Analysis in 2014, [34]. This paper deals with the singularly perturbed version of problem (2), i.e. the case when the diffusion parameter $\varepsilon \to 0$, or even $\varepsilon = 0$. The purpose of the paper presented in Chapter 4 is to derive a priori error estimates of the type (15) which would be uniform with respect to $\varepsilon \to 0$ and valid also in the purely convective case. This pursuit stems from the fact that error analysis using classical techniques such as those presented in Chapters 2 and 3 lead to estimates where the constant C in (15) blows up exponentially with $\varepsilon \to 0$. The results and techniques if Chapter 4 generalize those of the series of papers by Q. Zhang and C.-W. Shu starting with the paper [50].

4.1 Limitations of the classical parabolic technique

The analysis of Chapters 2 and 3 uses the so-called *parabolic technique*. Problem (2) is treated primarily as a heat equation (i.e. parabolic equation without convection) with an additional convection term. The diffusion term A_h is elliptic in the DG-norm (12). One can therefore use the classical ellipticity-based estimation technique for parabolic problems, [12]. Equations (2) – in the weak form – and (7) are subtracted to obtain an equation for the error $e_h = u - u_h = \eta + \xi$, where $\eta = u - \Pi_h u, \xi = \Pi_h u - u_h \in S_h$ with a suitable projection $\Pi_h u$ of u onto S_h. The error equation is then tested with $\varphi_h := \xi$ and the proved ellipticity estimates for A_h are applied along with estimates of η following from approximation properties of Π_h.

As for the convective terms, they are than estimated straightforwardly as

$$|b_h(u,\xi) - b_h(u_h,\xi)| \le C\|\xi\|_{DG}\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big), \tag{27}$$

cf. Chapter 2, Lemma 17 and Chapter 3, Section 4.2. While the term $\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big)$ in estimate (27) is desirable for the subsequent error analysis, the only possibility how to deal with the term $\|\xi\|_{DG}$ is to "dominate" it by the elliptic terms: we estimate using Young's inequality

$$C\|\xi\|_{DG}\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big) \le \tfrac{\varepsilon}{2}\|\xi\|_{DG}^2 + \tfrac{C}{4\varepsilon}\big(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\big)^2 \tag{28}$$

and the unpleasant term $\frac{\varepsilon}{2}\|\xi\|_{DG}^2$ is subtracted from the left-hand side elliptic term $\varepsilon\|\xi\|_{DG}^2$ stemming from the diffusion terms. The result of this procedure is the constant $\frac{C}{4\varepsilon}$ in the remaining right-hand side terms. After the application of Gronwall's lemma, this results in a constant of the form $\exp(\frac{C}{\varepsilon})$ in the resulting error estimate (15). This constant is unrealistically large for $\varepsilon \ll 1$ and it is not uniform with respect to $\varepsilon \to 0$. Furthermore, the analysis is not valid for $\varepsilon = 0$. We note that in the case of nonlinear diffusion (13), these observations are still valid, with the final constant blowing up exponentially as $\beta_0 \to 0$, cf. (14) and Chapter 2.

4.2 The technique of Zhang and Shu

In the paper [50], the authors have managed to overcome the limitations of the parabolic technique for high order Runge-Kutta Discontinuous Galerkin schemes. The presented analysis is based on a more subtle estimate of the convective terms than the straightforward bound (27). Along with the standard properties (H1)–(H3) of the numerical flux, the addition *E-flux* property is assumed:

(H4) $\big(H(v,w,\mathbf{n}) - \mathbf{f}(q)\cdot\mathbf{n}\big)(v-w) \ge 0, \quad \forall v,\, w \in \mathbb{R},\ \mathbf{n} \in B_1$ and all q between v, w.

This technical assumption is satisfied for all so-called monotone numerical fluxes, a property satisfied by many numerical fluxes used in practice, e.g. Lax-Friedrichs, Godunov, Engquist-Osher and the Roe flux with entropy fix, cf. [41], [8].

Using the E-flux condition, the following estimate can be derived:

$$\big|b_h(u_h,\xi) - b_h(u,\xi)\big| \leq C\Big(1 + \frac{\|e_h(t)\|_\infty^2}{h^2}\Big)\big(h^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2\big), \tag{29}$$

cf. Chapter 4, Lemma 7. The advantage of estimate (29) over (27) is that the term $h^{-2}\|e_h(t)\|_\infty^2$ can be eliminated. For if we knew a priori that the error satisfies $\|e_h(t)\|_\infty = O(h)$, then $h^{-2}\|e_h(t)\|_\infty^2 = O(1)$ and estimate (29) reduces to the term $C(h^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2)$ which is ideal for the application of Gronwall's lemma, leading to the improved estimate $\|e_h(t)\|_{L^2(\Omega)} = O(h^{p+1/2})$. Since the convection terms are estimated independently of the diffusion terms, there is no need to use estimates such as (28) involving ε^{-1}. We therefore obtain estimates which are uniform for $\varepsilon \to 0$ and valid even in the limiting case $\varepsilon = 0$. In [50], the $O(h)$ a priori assumption is eliminated via mathematical induction for an explicit Runge-Kutta time discretization of the discontinuous Galerkin scheme. An artefact of the technique is that the degree of polynomial approximation must satisfy certain conditions, such as $p > (1+d)/2$ in order to carry out the induction steps.

In Chapter 4, the ideas of Zhang and Shu are generalized in the following ways:

- Estimate (29) originally derived in 1D for periodic boundary conditions is generalized to $\mathbb{R}^d$ with mixed Dirichlet-Neumann boundary conditions, cf. Chapter 4, Lemma 7.

- The technique is applied to the space semidiscrete scheme (7), cf. Chapter 4, Section 7, replacing the mathematical induction argument using by continuous mathematical induction, cf. [11].

- The technique is applied to the implicit Euler scheme. To overcome fundamental obstacles with the induction argument in case of an implicit scheme, a suitable continuation of the discrete solution is constructed and again a continuous mathematical induction argument is applied to the continued version of the error, cf. Chapter 4, Lemma 8.

- The error analysis is generalized to the case of only locally Lipschitz continuous $\mathbf{f}$, cf. Chapter 4, Lemma 9.

Due to lack of space in this introductory chapter and the technical nature of the arguments, we only briefly outline the continuation argument from the analysis of the implicit Euler scheme. For the purely convective case this reads: Find $u_h^{n+1} \in S_h$ such that for all $\varphi_h \in S_h$

$$\big(u_h^{n+1} - u_h^n, \varphi_h\big) + \tau_n b_h\big(u_h^{n+1}, \varphi_h\big) = \tau_n l_h\big(\varphi_h\big)(t_{n+1}), \tag{30}$$

where τ_n is the current time step. In Chapter 4, Lemma 14, it is proved that the estimate (29) is insufficient to prove the desired error estimate. Therefore, we construct a continuation of the discrete solution u_h^n: For every $\tau \in [0, \tau_n]$ find $u_\tau \in S_h$ such that for all $\varphi_h \in S_h$

$$\big(u_\tau - u_h^n, \varphi_h\big) + \tau b_h\big(u_\tau, \varphi_h\big) = \tau l_h\big(\varphi_h\big)(t_{n+1}). \tag{31}$$

Formally, if $\tau = 0$, we have $u_\tau = u_h^n$ while for $\tau = \tau_n$, we obtain $u_\tau = u_h^{n+1}$, the solution of (30). It can be proved (Chapter 4, Lemma 16) that between these two values u_τ depends continuously on the parameter τ. Therefore, we can go continuously from u_h^n

to u_h^{n+1}, which allows us to carry out the necessary induction steps in the Zhang-Shu technique. The result is an error estimate of the form

$$\max_{n\in\{0,\cdots,N+1\}} \|e_h^n\|^2 \leq C_T^2\big(h^{2p+1} + \varepsilon h^{2p} + \tau^2\big), \tag{32}$$

where the constant C_T is independent of ε, h, τ, cf. Chapter 4, Theorem 19.

5 Overview of Chapter 5: Simulation of compressible viscous flow in time-dependent domains

Chapter 5 consists of the paper *Simulation of compressible viscous flow in time-dependent domains*, published in the journal Applied Mathematics and Computation in 2013, [10]. Unlike the previous three chapters, Chapters 5 and 6 deal with practical applications of the discontinuous Galerkin method. Both these chapters deal with the numerical solution of the compressible Navier-Stokes equations in time-dependent domains with the aid of the *ALE*, or *Arbitrary Lagrangian-Eulerian* method, cf. e.g. [40]. The resulting equations are discretized by a semi-implicit discontinuous Galerkin scheme. In Chapter 5, the movement of the domain will be either prescribed (air flow through a channel with moving walls) or governed by a simple system of ordinary differential equations (flow induced airfoil vibrations).

The equations which are solved are the compressible Navier-Stokes equations written in conservative form:

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \mathbf{f}_s(\boldsymbol{w})}{\partial x_s} = \sum_{s=1}^{2} \frac{\partial \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_s}, \tag{33}$$

where $\boldsymbol{w} : Q_T \to \mathbb{R}^4$ represents the vector of conserved variables and $\mathbf{f}_s, \boldsymbol{R}_s$ are the so-called Euler and viscous fluxes, respectively, cf. Chapter 5, Section 2. Equation (33) has a similar form as the scalar equation (13) and can be analogically discretized by the discontinuous Galerkin method.

Unlike equation (13), system (33) along with suitable initial and boundary conditions is considered on a spatial domain Ω_t depending on time $t \in [0, T]$. This is taken into account using the ALE method.

5.1 Arbitrary Lagrangian-Eulerian method

In the ALE method, the movement of Ω_t is considered with respect to a reference domain Ω_0. Typically Ω_0 is the computational domain Ω_t taken at the initial time $t = 0$. The domain Ω_t is described using a one-to-one mapping

$$\mathcal{A}_t : \overline{\Omega}_0 \longrightarrow \overline{\Omega}_t \tag{34}$$

which maps points $\boldsymbol{X} \in \overline{\Omega}_0$ to points $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{X}, t) = \mathcal{A}_t(\boldsymbol{X}) \in \overline{\Omega}_t$. Since the domain is time-dependent, it is useful to define the *domain velocity*

$$\boldsymbol{z}(\boldsymbol{x}, t) = \Big(\frac{\partial}{\partial t}\mathcal{A}_t(\boldsymbol{X})\Big)\Big|_{\boldsymbol{X}=\mathcal{A}_t^{-1}(\boldsymbol{x})} \tag{35}$$

for $t \in [0, T]$ and $\boldsymbol{x} \in \Omega_t$. The basis of the ALE method lies in replacing the time derivative in the governing equation (33) with the *ALE derivative*: For a given function $f = f(\boldsymbol{x}, t)$ defined for $\boldsymbol{x} \in \Omega_t$ and $t \in [0, T]$:

$$\frac{D^A}{Dt} f(\boldsymbol{x}, t) = \frac{\partial \tilde{f}}{\partial t}(\boldsymbol{X}, t), \tag{36}$$

where

$$\tilde{f}(\boldsymbol{X},t) = f(\mathcal{A}_t(\boldsymbol{X}),t),\ \boldsymbol{X}\in\Omega_0,\ \boldsymbol{x}=\mathcal{A}_t(\boldsymbol{X}).$$

By the application of the chain rule, one obtains the fundamental relation

$$\frac{D^A f}{Dt} = \frac{\partial f}{\partial t} + \operatorname{div}(\boldsymbol{z} f) - f\operatorname{div}\boldsymbol{z}, \tag{37}$$

which allows us to rewrite the compressible Navier-Stokes equations in the ALE form

$$\frac{D^A \boldsymbol{w}}{Dt} + \sum_{s=1}^{2}\frac{\partial \boldsymbol{g}_s(\boldsymbol{w})}{\partial x_s} + \boldsymbol{w}\operatorname{div}\boldsymbol{z} = \sum_{s=1}^{2}\frac{\partial \boldsymbol{R}_s(\boldsymbol{w},\nabla\boldsymbol{w})}{\partial x_s}, \tag{38}$$

where the ALE modified inviscid fluxes are defined by

$$\boldsymbol{g}_s(\boldsymbol{w}) := \mathbf{f}_s(\boldsymbol{w}) - z_s\boldsymbol{w}, \quad s=1,2. \tag{39}$$

5.2 Discontinuous Galerkin discretization

System (38) is discretized by the discontinuous Galerkin method in space similarly as in Section 1. At time t the domain Ω_t is partitioned into a triangulation $\mathcal{T}_{ht}$. The discrete solution is sought in the discontinuous Galerkin space

$$\boldsymbol{S}_{ht} = [S_{ht}]^4, \quad \text{where}\ \ S_{ht} = \{v; v|_K \in P^p(K)\ \forall\, K\in\mathcal{T}_{ht}\}. \tag{40}$$

Using these discrete spaces, we can discretize system (38) in the following way.

Definition 5. *We say $\boldsymbol{w}_h \in C^1([0,T];\boldsymbol{S}_{ht})$ is a discontinuous Galerkin solution of (38) if for all $t\in(0,T)$ and $\boldsymbol{\varphi}_h\in\boldsymbol{S}_{ht}$*

$$\begin{aligned}\sum_{K\in\mathcal{T}_{ht}}\int_K \frac{D^A\boldsymbol{w}_h(t)}{Dt}\cdot\boldsymbol{\varphi}_h\,dx + b_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) + a_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h)\\ + J_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) + d_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) = \ell_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h).\end{aligned} \tag{41}$$

In (41), the forms b_h, a_h, J_h and l_h are vector analogies of the forms (6) and (8) – (10). Only the convective form b_h is based on the modified fluxes $\boldsymbol{g}_s$, (39) instead of the original fluxes $\mathbf{f}_s$ of (33). The new *reaction form* arises due to the ALE "reaction" term $\boldsymbol{w}\operatorname{div}\boldsymbol{z}$ and is defined as

$$d_h(\boldsymbol{w},\boldsymbol{\varphi}_h) = \sum_{K\in\mathcal{T}_{ht}}\int_K (\boldsymbol{w}\cdot\boldsymbol{\varphi}_h)\operatorname{div}\boldsymbol{z}\,dx. \tag{42}$$

The system of ordinary differential equations (41) is discretized with respect to time using the *semi-implicit* approach of [21]. In the case of a stationary domain Ω, this is essentially the implicit Euler scheme, where the nonlinear terms are linearised with respect to the unknown solution on the next time level $\boldsymbol{w}_h^{k+1}$ using suitable properties of the individual convective and diffusive terms. For example, in the convective terms, the nonlinearities of the form $\boldsymbol{g}_s(\boldsymbol{w}_h^{k+1})$ are approximated as

$$\boldsymbol{g}_s(\boldsymbol{w}_h^{k+1}) = (\mathbb{A}_s(\boldsymbol{w}_h^{k+1}) - z_s^{k+1}\mathbb{I})\boldsymbol{w}_h^{k+1} \approx (\mathbb{A}_s(\overline{\boldsymbol{w}}_h^{k+1}) - z_s^{k+1}\mathbb{I})\boldsymbol{w}_h^{k+1}, \tag{43}$$

where $\mathbb{A}_s(\boldsymbol{w})$ is the Jacobi matrix of $\mathbf{f}_s(\boldsymbol{w})$, cf. [21] and $\overline{\boldsymbol{w}}_h^{k+1}$ is a state vector extrapolated from $\boldsymbol{w}_h^k$ and $\boldsymbol{w}_h^{k-1}$. The first equation in (43) follows from the first order homogeneity of $\mathbf{f}_s$. Similar linearizations can be performed for the diffusion terms and the numerical flux, if it is suitably chosen. In Chapter 5 the Vijayasundaram numerical

flux is used, due to its appropriate form for linearization similar to (43), cf. [21] and [49]. Finally, the ALE derivative is approximated by a second order backward difference formula applied to the time derivative in its definition (36). The advantage of the chosen semi-implicit scheme is its practically unconditional stability obtained with the solution of only one linear algebraic system of equations per time level as demonstrated in [21]. The resulting linear systems are solved using the *Generalized Minimal Residual* (GMRES) method with block-diagonal preconditioning, [21].

In the second numerical experiment from Chapter 5, Section 4.2.1, we also deal with transonic flows. Therefore it is necessary to treat the Gibbs phenomenon occurring in the vicinity of discontinuities and steep gradients. For this purpose, we add local element-wise and interelement artificial viscosity terms to the resulting semi-implicit formulation, cf. Chapter 5, Section 3.3. These artificial viscosity terms are based on the discontinuity indicator which measures the interelement jumps of density, cf. [17] and [21].

An important ingredient in the proposed numerical method is the treatment of boundary conditions. On artificial boundaries (inlet and outlet), transparent, non-reflecting boundary conditions based on local linearizations of the Euler equations are applied, cf. Chapter 5, Section 3.4 and [21]. On moving solid impermeable walls, the no slip boundary condition for the fluid velocity $\boldsymbol{v} = \boldsymbol{z}$ is prescribed, where $\boldsymbol{z}$ is interpreted as the velocity of the moving wall.

Two numerical experiments are performed to test the described numerical method. First, in Chapter 5, Section 4.1, air flow through a channel with moving walls is considered. The shape and movement of the channel is inspired by the human glottis and is taken from [43]. Together with more sophisticated simulations using true fluid-structure interaction from Chapter 6, the goal is the simulation of voice formation in human vocal folds. In the case of Chapter 5, the movement of the solid walls is prescribed as a periodic motion with frequency 100 Hz mimicking the opening and closing of the vocal chord aperture. This movement is then simply interpolated into the domain to obtain the ALE mapping $\mathcal{A}_t$ also in the interior of Ω_t. The inlet Reynolds number is $Re = 5227$, the Mach number is $M_{in} = 0.012$, i.e. a relatively low Mach flow is considered. Complicated interacting vortical structures arise downstream from the moving part of the channel. The results are compared to similar simulations performed in [43] obtained by the finite volume method, cf. also [44].

In the second numerical experiment, Chapter 5, Section 4.2, a simple example of fluid-structure interaction is presented. We consider subsonic and transonic flow around a rigid, elastically supported NACA 0012 airfoil. The motion of the profile is governed by a system of two nonlinear ordinary differential equations for the vertical displacement and rotation angle of the airfoil. Similar test problems have been considered e.g. in [18]. A *strong coupling* iterative procedure was applied to solve the coupled system consisting of the compressible air flow and equations describing the movement of the profile. As in the previous test case, the movement of the boundary (profile) is interpolated to the rest of Ω_t to obtain $\mathcal{A}_t$, similarly as in [18]. In the numerical experiment, we were able to capture the rise of the so-called *flutter instability* for inlet flow velocity 40 m/s. For lower velocities the airfoil vibrations are damped.

6 Overview of Chapter 6: Discontinuous Galerkin for the interaction of a compressible fluid and structures

Chapter 6 consists of the paper *DGFEM for dynamical systems describing interaction of compressible fluid and structures*, [26], published in the Journal of Computational and Applied Mathematics in 2013. In Chapter 5 only a simple case of fluid-structure

interaction is considered, where the equations describing the movement of the structure are simple ordinary differential equations. In Chapter 6 these results are extended to a complicated fluid-structure interaction problem, where the compressible Navier-Stokes equations are coupled with equations describing the deformation of an elastic body governed by the generalized Hooke's law.

6.1 Elasticity equations for the body and ALE mapping.

Similarly as in Chapter 5, we shall consider the compressible Navier-Stokes equations in a time dependent domain $\Omega_t \subset \mathbb{R}^2$, where the ALE method will be used to describe the domain using a mapping $\mathcal{A}_t : \overline{\Omega}_0 \longrightarrow \overline{\Omega}_t$ from a reference domain Ω_0. Along with these equations, in Chapter 6 we shall also consider interaction of the fluid flow with an elastic body. The elastic body will be represented by a bounded open domain $\Omega^b \subset \mathbb{R}^2$, which will be assumed to have a common (part of the) boundary with Ω_0 denoted by Γ_W^b. For $\boldsymbol{X} \in \Omega^b$ we denote the displacement of point $\boldsymbol{X}$ of the elastic body at time t by $\boldsymbol{u}(\boldsymbol{X}, t) = (u_1(\boldsymbol{X}, t), u_2(\boldsymbol{X}, t))$. Therefore at time t, the point $\boldsymbol{X} \in \Omega^b$ will be located at

$$\boldsymbol{x} = \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{X}, t). \tag{44}$$

As the governing equations for the motion of the elastic body we shall take the dynamical equations for the displacement u of an isotropic elastic body

$$\varrho^b \frac{\partial^2 u_i}{\partial t^2} + C\varrho^b \frac{\partial u_i}{\partial t} - \sum_{j=1}^{2} \frac{\partial \tau_{ij}^b}{\partial X_j} = 0 \quad \text{in } \Omega^b \times (0, T), \quad i = 1, 2, \tag{45}$$

where ϱ^b denotes the material density and τ_{ij}^b are the components of the stress tensor defined by the generalized Hooke's law for isotropic bodies, cf. Chapter 6, Section 2.2. We note that $\tau_{ij}^b = \tau_{ij}^b(u)$ depends on the displacement u and its first derivatives via the strain tensor. Therefore (45) represents an equation for the single unknown $\boldsymbol{u}$. The term $C\varrho^b \frac{\partial u_i}{\partial t}$, where $C \geq 0$, represents the dissipative structural damping of the system, which is natural for real bodies.

System (45) must be equipped with initial and boundary conditions. These are taken in a standard way, cf. Chapter 6, Section 2.2, with the exception of the common interface Γ_W^b between the reference fluid domain Ω_0 and the elastic body Ω^b. On Γ_W^b, system (45) is equipped with the so-called *transmission condition*

$$\sum_{j=1}^{2} \tau_{ij}^b(\boldsymbol{X}) n_j(\boldsymbol{X}) = -\sum_{j=1}^{2} \tau_{ij}^f(\boldsymbol{x}) n_j(\boldsymbol{X}), \quad i = 1, 2, \tag{46}$$

where τ_{ij}^f are the components of the stress tensor of the fluid, cf. Chapter 6, Section 2.2. Condition (46) prescribes the normal component of the stress tensor τ^b and expresses the force balance between the aerodynamic forces and the forces on the structure surface.

As for the Navier-Stokes equations, on the common part of the boundary $\tilde{\Gamma_{W_t}}$ corresponding to Γ_W^b by the mapping (44) at time t, the *second transmission condition* is prescribed:

$$\boldsymbol{v}(\boldsymbol{x}, t) = \frac{\partial \boldsymbol{u}(\boldsymbol{X}, t)}{\partial t}, \tag{47}$$

which corresponds to the no-slip moving wall boundary condition $\boldsymbol{v} = \boldsymbol{z}$ of Chapter 5.

System (45) is used not only to describe the deformation of the elastic body, but its stationary version is used to construct the ALE mapping $\mathcal{A}_t$. We seek $\mathcal{A}_t : \overline{\Omega}_0 \to \overline{\Omega}_t$ expressed using a displacement vector field $\boldsymbol{d}$:

$$\mathcal{A}_t(\boldsymbol{X}) = \boldsymbol{X} + \boldsymbol{d}(\boldsymbol{X}, t), \quad \boldsymbol{X} \in \overline{\Omega}_0, \tag{48}$$

in analogy to (44). The unknown $\boldsymbol{d} : \bar{\Omega}_0 \to \mathbb{R}^2$ will be sought as the solution of the artificial static elasticity problem

$$\sum_{j=1}^{2} \frac{\partial \tau_{ij}^a}{\partial x_j} = 0 \quad \text{in } \Omega_0, \quad i = 1, 2, \tag{49}$$

where $\tau_{ij}^a = \tau_{ij}^a(\boldsymbol{d})$ are the components of the artificial stress tensor defined using the generalized Hooke's law similarly as for the elastic problem on Ω^b. Equation (49) is therefore a second order linear elliptic partial differential equation for the unknown $\boldsymbol{d}$.

On Γ_W^b, we equip (49) with the boundary condition

$$\boldsymbol{d}(\boldsymbol{X}, t) = \boldsymbol{u}(\boldsymbol{X}, t). \tag{50}$$

The philosophy behind this approach is that we want to "interpolate" the movement of the boundary $\partial\Omega_t$, which we know from the elastic problem (45) into the whole fluid domain Ω_t. This cannot be done e.g. by straightforward linear interpolation as in Chapter 5, where the domain movement is simple and/or prescribed. For this purpose we view Ω_t as an elastic body with prescribed deformation of (part of) its boundary. For small enough deformations, we can expect that similarly as for elastic bodies, the artificial problem (49) will give us a one-to-one mapping of the reference domain (configuration) Ω_0 onto Ω_t.

6.2 Discretization

The fluid problem, i.e. the compressible Navier-Stokes equations in ALE form are solved using the same semi-implicit discontinuous Galerkin method with backward difference formula time discretization of order 2 as in Chapter 5. Using this method, the new solution w_H^{k+1} can be found whenever the mapping $A_{t_{n+1}}$ is known.

The elasticity problem (45) is discretized in space using the standard piecewise linear conforming finite element method, i.e. a weak form of (45) is taken on the space V_h of (4) with $p = 1$, cf. [12]. With respect to time, the resulting second order system of ordinary differential equations is discretized using the *Newmark method*, [13], which is essentially an implicit scheme. The resulting system of linear algebraic equations is symmetric positive definite, therefore it is solved using the *conjugate gradient method*. The elasticity problem can be solved whenever the data from the boundary condition (46) is known from the fluid simulation.

Similarly, the artificial elasticity problem (49) used for the construction of the ALE mapping $\mathcal{A}_t$ is solved using the piecewise linear conforming finite element method. Equation (49) is stationary and the resulting symmetric positive definite linear algebraic system can be solved by the conjugate gradient method similarly as the systems of linear equations arising in the Newmark method used to solve the nonstationary model (45). The artificial elasticity problem can be solved once the data from boundary condition (50) is available from the solution of the elasticity problem.

Since the solutions of the three individual subproblems depend on each other via the transfer boundary conditions, some strategy is needed to solve the entire coupled problem using the solution procedures for the three described subproblems. In Chapter 6, Section 4.2, the *strong* and *weak coupling* procedures are described. In brief form, these can be described by the iterative procedure performed on each time level:

1. Solve elasticity problem (45) using the aerodynamic forces (46) from the previous iteration.

2. Solve the artificial elasticity problem (49) using the displacement $\boldsymbol{u}$ from point 1.

3. Solve the flow problem (38) using the mapping $\mathcal{A}_t$ obtained from point 2.

The *weak coupling* procedure consists of performing steps 1.–3. only once per time level. On the other hand, in the *strong coupling* procedure, on each time level steps 1.–3. are repeated iteratively until some form of convergence is obtained. After that we move on to the next time level. The advantage of the strong coupling procedure is its higher stability and robustness, while weak coupling requires less CPU time. Our numerical experiments show that only a few inner iterations in the strong coupling procedure are needed to obtain convergence of the computed displacements $\boldsymbol{u}$ in point 1. of the algorithm above.

The developed numerical method is tested on a model problem of air flow through a channel interacting with two elastic bumps representing a simplified model of human vocal folds. The channel is 160 mm long with the narrowest aperture 1.6 mm. The flow parameters are the same as in the vocal fold numerical experiment of Chapter 5. Results are compared on three successively refined meshes and for weak and strong coupling. The results indicate that in the considered test case the differences between weak and strong coupling are not large. To analyze the character of the resulting vocal fold vibrations, several "sensor" points are monitored on the surface of the elastic bumps as well as pressure in a fluid sensor point monitoring the air pressure. A Fourier analysis of the resulting signals is performed demonstrating the presence of a dominant frequency in these signals (approximately 439 Hz for vertical displacement of the sensors and 113 Hz for horizontal displacement). The obtained numerical results are compared to numerical simulations using other methods, e.g. [44], [38], and wind tunnel experiments, cf. [27] and [28].

References

[1] Akrivis, G., Makridakis, C.: Galerkin time-stepping methods for nonlinear parabolic equations. *ESAIM: Math. Modelling and Numer. Anal.*, **38**, 261–289 (2004).

[2] Arnold, D. N.: An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, **19**, 742–760 (1982).

[3] Arnold, D.N., Brezzi, F., Cockburn, B., Marini, D.: Discontinuos Galerkin methods for elliptic problems. In: Discontinuous Galerkin methods. *Theory, Computation and Applications. Lecture Notes in Computational Science and Engineering 11*, Springer, Berlin, 89–101 (2000).

[4] Arnold, D., Brezzi, F., Cockburn, B., Marini, L.D.: Unified Analysis of Discontinuous Galerkin Methods for Elliptic Problems. *SIAM J. Numer. Anal* **39**(5), 1749–1779 (2001).

[5] Aubin, J.P.: Behavior of the Error of the Approximate Solutions of Boundary-Value Problems for Linear Elliptic Operators by Galerkin's Method and Finite Differences. *Annali della Scuola Normale di Pisa*, **3**(21), 599–637 (1967).

[6] Babuška, I., Baumann, C.E., Oden, J.T.: A discontinuous *hp* finite element method for diffusion problems, 1-D analysis. *Comput. Math. Appl.*, **37**, 103–122 (1999).

[7] Bassi, F., Rebay, S.: High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, **138**, 251–285 (1997).

[8] Barth, T., Ohlberger, M.: *Finite Volume Methods: Foundation and Analysis, Encyclopedia of Computational Mechanics, volume 1.* John Wiley & Sons, Chichester, New York, Brisbane, 439–474 (2004).

[9] Baumann, C. E., Oden, J. T.: A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations. *Int. J. Numer. Methods Fluids*, **31**, 79–95 (1999).

[10] Česenek, J., Feistauer, M., Horáček, J., Kučera, V., Prokopová, J.: Simulation of compressible viscous flow in time-dependent domains. *Appl. Math. Comput.* **219**(13), 7139–7150 (2013).

[11] Chao, Y.R.: A note on "Continuous mathematical induction". *Bull. Amer. Math. Soc.*, **26** (1), 17–18 (1919).

[12] Ciarlet, P.G. *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam (1979).

[13] Curnier, A.: *Computational Methods in Solid Mechanics.* Kluwer Academic Publishing Group, Dodrecht (1994).

[14] Dolejší, V., Feistauer, M.: Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems. *Numer. Funct. Anal. Optimiz.,* **26**, 349–383 (2005).

[15] 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 convection–diffusion problem. *IMA J. Numer. Anal.,* **28**, 496–521 (2008).

[16] Dolejší, V., Feistauer, M., Sobotíková, V.: Analysis of the discontinuous Galerkin method for nonlinear convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.,* **194**, 2709-2733 (2005).

[17] Dolejší, V., Feistauer, M., Schwab, C.: On some aspects of the discontinuous Galerkin finite element method for conservation laws. *Math. Comput. Simul.*, **61**, 333–346 (2003).

[18] Dubcová, L., Feistauer, M., Horáček, J., Sváček, P.: Numerical simulation of interaction between turbulent flow and a vibrating airfoil. *Comput. Visual. Sci.*, **12**, 207–225 (2009).

[19] Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: *Computational Differential Equations*. Cambridge University Press, Cambridge (1996).

[20] Feistauer, M., Felcman, J., Straškraba, I. *Mathematical and Computational Methods for Compressible Flow*. Clarendon Press, Oxford (2003).

[21] Feistauer, M., Kučera, V.: On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.* **224**, 208-221 (2007).

[22] Feistauer, M., Kučera, V.: Analysis of the DGFEM for nonlinear convection-diffusion problems, *Electr. Trans. Num. Anal*, **32**, 33–48 (2008).

[23] Feistauer, M., Kučera, V., Najzar, K., Prokopová, J.: Analysis of Space-Time Discontinuos Galerkin Method for Nonlinear Convection-Diffusion Problems. *Numer. Math.* **117**, 251–288 (2011).

[24] Grisvard, P.: *Singularities in Boundary Value Problems*. Springer, Berlin (1992).

[25] Harten, A., Engquist, B., Osher, S., Chakravarthy, S.: Uniformly high order essentially non-oscillatory schemes, III. *J. Comp. Phys.*, **71**, 231–303 (1987).

[26] J. Hasnedlová-Prokopová, J., Feistauer, M., Horáček, J., Kosík, A., Kučera, V.: DGFEM for Dynamical Systems Describing Interaction of Compressible Fluid and Structures. *J. Comput. Appl. Math.* **254**, 17-–30 (2013).

[27] Horáček, J., Šidlof, P., Uruba, V., Veselý, J., Radolf, V., Bula, V.: Coherent structures in the flow inside a model of human vocal tract with self-oscillating vocal folds. *Acta Technica*, **55**, 327–343 (2010).

[28] Horáček, J., Uruba, V., Radolf, V., Veselý J., Bula, V.: Airflow visualization in a model of human glottis near the self-oscillating vocal folds model. *Appl. Comput. Mech.*, **5**, 21–28 (2011).

[29] Huerta, A., Casoni, E., Peraire, J.: A Simple Shock-Capturing Technique for High-Order Discontinuous Galerkin Methods. *Int. J. Numer. Meth. Fluids* **69**(10) 1614 – 1632 (2012).

[30] Johnson, C., Pitkäranta, J.: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, **46**, 1–26 (1986).

[31] Kröner, D.: *Numerical Schemes for Conservation Laws*. Wiley und Teubner, 1996.

[32] Kučera, V.: *Higher order methods for the solution of compressible flows*. Doctoral thesis, Charles University in Prague, Czech Republic (2007).

[33] Kučera, V.: Optimal $L^\infty(L^2)$-error Estimates for the DG Method Applied to Nonlinear Convection-Diffusion Problems with Nonlinear Diffusion. *Numer. Func. Anal. Opt.*, **31**(3), 285—312 (2010).

[34] Kučera, V.: On diffusion-uniform error estimates for the DG method applied to singularly perturbed problems. *IMA J. Numer. Anal.* **34** (2): 820-861 (2014).

[35] Le Saint, P., Raviart, P.-A.: On a finite element method for solving the neutron transport equation. *Mathematical Aspects of Finite Elements in Partial Differential Equations*, Academic Press, 89–145 (1974).

[36] LeVeque, R. J.: *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press (2002).

[37] Liu, X.-D., Osher, S., Chan, T.: Weighted essentially non-oscillatory schemes, *J. Comput. Phys.*, **115**, 200-–212 (1994).

[38] Luo, H., Mittal, R., Zheng, X., Bielamowicz, S.A., Walsh, R.J., Hahn, J.K.: An immersed-boundary method for flow-structure interaction in biological systems with application to phonation. *J. Comput. Phys.*, **227**, 9303–9332 (2008).

[39] Nitsche, J.A.: Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. *Numer. Math.* **11**, 346-348 (1963).

[40] T. Nomura, T., Hughes, T.J.R.: An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. *Comput. Methods Appl. Mech. Engrg.*, **95**, 115–138 (1992).

[41] Osher, S.: Riemann solvers, the entropy condition, and difference approximations. *SIAM. J. Numer. Anal.*, **21**, 217–235 (1984).

[42] Persson, P-O., Peraire, J.: Sub-Cell shock capturing for Discontinuous Galerkin methods. *Proc. of the 44th AIAA Aerospace Sciences Meeting and Exhibit* (2006).

[43] Punčochářová, P., Fürst, J., Kozel, K., Horáček, J.: Numerical solution of compressible flow with low Mach number through oscillating glottis. *Proceedings of the 9th International Conference On Flow-Induced Vibration (FIV 2008)*, Institute of Thermomechanics AS CR, Prague, 135–140 (2008).

[44] Punčochářová-Pořízková, P., Kozel, K., Horáček, J.: Simulation of unsteady compressible flow in a channel with vibrating walls - influence of the frequency. *Computers and Fluids*, **46**, 404–410 (2011).

[45] Reed, W. H., Hill, T. R.: Triangular mesh methods for the neutron transport equation. *Technical Report LA–UR–73–479*, Los Alamos Scientific Laboratory (1973).

[46] Roos, H. G., Stynes, M., Tobiska, L.: *Robust Numerical Methods for Singularly Perturbed Differential Equations, Convection Diffusion and Flow Problems*. Springer-Verlag, Berlin, Heidelberg, (2008).

[47] Thomée, V.: *Galerkin Finite Element Methods for Parabolic Problems*. Springer, Berlin (2006).

[48] Van der Vegt, J. J. W., van der Ven, H.: Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flow. *J. Comput. Phys.*, **182**, 546–585 (2002).

[49] Vijayasundaram, G.: Transonic flow simulation using upstream centered scheme of Godunov type in finite elements. *J. Comput. Phys.*, **63**, 416–433 (1986).

[50] Zhang, Q. & Shu, C.-W.: Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM J. Numer. Anal.*, **42**(2), 641–666 (2004).

Optimal $L^\infty(L^2)$-error estimates for the DG method applied to nonlinear convection-diffusion problems with nonlinear diffusion.

Václav Kučera

Faculty of Mathematics and Physics, Charles University in Prague
Sokolovská 83, Praha 8, 186 75 , Czech Republic

Published in April 2010,
Numerical Functional Analysis and Optimization

Abstract

This paper is concerned with the analysis of the discontinuous Galerkin finite element method (DGFEM) applied to the space semidiscretization of a nonstationary convection-diffusion problem with nonlinear convection and nonlinear diffusion. Optimal estimates in the $L^\infty(L^2)$-norm are derived for the symmetric interior penalty (SIPG) scheme in two dimensions. The error analysis is carried out for nonconforming triangular meshes under the assumption that the exact solution of the problem and the solution of a linearised elliptic dual problem are sufficiently regular.

Keywords: Convection-diffusion equation, nonlinear diffusion, discontinuous Galerkin finite element method, symmetric formulation of diffusion terms, interior and boundary penalty, method of lines, optimal error estimates.

1 Introduction

The numerical solution of nonstationary convection-diffusion problems plays an important role in many areas of applied mathematics ranging from fluid dynamics and heat transfer on one side to image processing on the other side. In the numerical treatment of such problems many difficulties arise due to the occurrence of internal and boundary layers, where steep gradients or discontinuities appear. Many numerical methods have been devised to overcome such difficulties. The finite volume (FV) method, which is often used, is based on piecewise constant approximations. It has good stability properties in the vicinity of discontinuities, however it has a low order of accuracy and its generalization to higher order methods is rather sophisticated. On the other hand, the finite element (FE) method with a high order of accuracy is suitable mainly for elliptic problems and various stabilization techniques (e.g. streamline diffusion or Galerkin least squares methods) must be employed to avoid spurious oscillations in the solution of convection-diffusion problems with dominating convection.

A natural generalization of the FV and FE methods is the discontinuous Galerkin finite element method (DGFEM). This method uses advantages of FV as well as FE methods: it is based on piecewise polynomial but discontinuous approximations, where

boundary fluxes are evaluated with the aid of a numerical flux. The use of discontinuous functions allows a flexible capturing of discontinuities and steep gradients, while the use of higher degree polynomials ensures a higher order of approximation in regions, where the solution is smooth.

Originally, the DGFE method was proposed for the solution of a neutron transport linear equation in [28] and analyzed theoretically in [27] and [24]. As for the numerical solution of elliptic and parabolic problems, discontinuous Galerkin methods are proposed and analysed in the pioneering works [33] and [1] with further theoretical analysis in [4], [2] and [3]. In the following decades, the DGFE method was applied to nonlinear conservation laws ([9], [23]) and the numerical solution of compressible flow ([5], [6], [7], [12], [20], [32], [14], [18]) as well as incompressible viscous flow ([29], [31]), porous media flow ([30]), shallow water flow ([10]), the Hamilton-Jacobi equations ([22]), the Schrödinger equation ([25]) and the Maxwell equations ([21]).

In this paper we are concerned with the analysis of the DGFE method applied to the space semidiscretization of a nonstationary convection-diffusion problem with nonlinear convection and nonlinear diffusion. The motivation to include also nonlinear diffusion along with nonlinear convection comes from the area of numerical treatment of compressible viscous flows governed by the compressible Navier-Stokes equations. This system of equations, when written in conservative form contain nonlinear convective as well as nonlinear viscous (diffusive) terms. Our scalar problem serves as a simplified model of the compressible Navier-Stokes equations.

We extend previous work from [13], [15] and [16], where linear diffusion (and nonlinear convection) is treated. In this case, apriori estimates optimal with respect to the order of convergence are obtained in the $L^2(H^1)-$ and $L^\infty(L^2)-$norms. As for nonlinear diffusion, we extend the work [17], where error estimates suboptimal with respect to the $L^\infty(L^2)-$norm are derived. By using a linearised elliptic dual problem we are able to improve these estimates using the Aubin-Nitsche technique. Optimal estimates in the $L^\infty(L^2)$-norm are derived for the symmetric interior penalty (SIPG) scheme in two dimensions. The error analysis is carried out for nonconforming triangular meshes under the assumption that the exact solution of the problem and the solution of a linearised elliptic dual problem are sufficiently regular.

2 Continuous problem

Let $\Omega \subset \mathbb{R}^2$ be a bounded open convex polygonal domain with Lipschitz-continuous boundary $\partial\Omega$ and $T > 0$. Let $Q_T := \Omega \times (0,T)$. We treat the following nonlinear problem:

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

$$u|_{\partial\Omega\times(0,T)} = u_D, \tag{2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega, \tag{3}$$

where the function $\beta \in C^2(\mathbb{R}) \cap W^{2,\infty}(\mathbb{R})$ is bounded from below and above and Lipschitz continuous:

$$\beta : \mathbb{R} \to [\beta_0, \beta_1], \quad 0 < \beta_0 < \beta_1 < \infty, \tag{4}$$

$$|\beta(u_1) - \beta(u_2)| \le L|u_1 - u_2|, \quad \forall u_1, u_2 \in \mathbb{R}. \tag{5}$$

Condition (5) implies $|\beta'| \le L$. Let $g : Q_T \to \mathbb{R}$, $u_D : \partial\Omega\times(0,T) \to \mathbb{R}$ and $u^0 : \Omega \to \mathbb{R}$ be given functions, and $f_1, f_2 \in C^1(\mathbb{R})$ be prescribed Lipschitz-continuous fluxes. Without loss of generality let $f_1(0) = f_2(0) = 0$.

In the following we shall use standard notation of function spaces. Let $G \subset \mathbb{R}^2$ be a bounded domain with a Lipschitz-continuous boundary ∂G. By $\overline{G}$ we denote the closure of G. Let $k \in \{0, 1, 2, \dots\}$ and $p \in [1, \infty]$. We use the well-known Lebesgue and Sobolev spaces $L^p(G)$, $L^p(\partial G)$, $W^{k,p}(G)$, $H^k(G) = W^{k,2}(G)$, $W^{k,p}(\partial G)$. By $H_0^1(G)$ we denote the space formed by all functions $v \in H^1(G)$ with zero traces on ∂G, i.e. $v|_{\partial G} = 0$. Further, we use the 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. [26]). The symbols $\|\cdot\|_X$ and $|\cdot|_X$ will denote a norm and a seminorm in a space X, respectively. By $(\cdot,\cdot)$ we denote the standard $L^2(\Omega)-$scalar product.

3 Discretization

3.1 Finite element mesh

Let $\mathcal{T}_h$ be a partition of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed triangles with mutually disjoint interiors. We shall call $\mathcal{T}_h$ a triangulation of Ω. We do not require the standard conforming properties of $\mathcal{T}_h$ used in the finite element method. This means that we admit the so-called hanging nodes. We shall use the following notation. By ∂K we denote the boundary of an element $K \in \mathcal{T}_h$ and set $h_K = \mathrm{diam}(K)$, $h = \max_{K\in\mathcal{T}_h} h_K$. By ρ_K we denote the radius of the largest circle inscribed into K and by $|K|$ we denote the area of K.

Let $K, K' \in \mathcal{T}_h$. We say that K and K' are *neighbours*, if the set $\partial K \cap \partial K'$ has positive length. We say that $\Gamma \subset K$ is a *face* (or *edge* in $\mathbb{R}^2$) of K, if it is a maximal connected open subset either of $\partial K \cap \partial K'$, where K' is a neighbour of K, or of $\partial K \cap \partial\Omega$. By $\mathcal{F}_h$ we denote the system of all faces of all elements $K \in \mathcal{T}_h$. Further, we define the set of all inner faces by

$$\mathcal{F}_h^I = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \Omega\}$$

and the set of all boundary faces by

$$\mathcal{F}_h^B = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \partial\Omega\}\,.$$

Obviously, $\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^B$.

For each $\Gamma \in \mathcal{F}_h$ we define a unit normal vector $\mathbf{n}_\Gamma$. We assume that for $\Gamma \in \mathcal{F}_h^B$ the normal $\mathbf{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega$. For each face $\Gamma \in \mathcal{F}_h^I$ the orientation of $\mathbf{n}_\Gamma$ is arbitrary but fixed. Finally, by $d(\Gamma)$ we denote the length of $\Gamma \in \mathcal{F}_h$.

3.2 Spaces of discontinuous functions

Over a triangulation $\mathcal{T}_h$ we define the broken Sobolev spaces

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

equipped with the seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_h)} = \Big(\sum_{K\in\mathcal{T}_h} |v|^2_{H^k(K)}\Big)^{1/2}.$$

For each face $\Gamma \in \mathcal{F}_h^I$ there exist two neighbours $K_\Gamma^{(L)}$, $K_\Gamma^{(R)} \in \mathcal{T}_h$ such that $\Gamma \subset K_\Gamma^{(L)} \cap K_\Gamma^{(R)}$. We use the convention that $\mathbf{n}_\Gamma$ is the outer normal to the element $K_\Gamma^{(L)}$ and the inner normal to the element $K_\Gamma^{(R)}$. For $v \in H^1(\Omega, \mathcal{T}_h)$ and $\Gamma \in \mathcal{F}_h^I$ we introduce the following notation:

$$v|_\Gamma^{(L)} = \text{ the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \qquad v|_\Gamma^{(R)} = \text{ the trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma,$$
$$\langle v \rangle_\Gamma = \tfrac{1}{2}\big(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}\big), \qquad [v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}.$$

The value $[v]_\Gamma$ depends on the orientation of $\mathbf{n}_\Gamma$, but the values $\langle v \rangle_\Gamma$ and $[v]_\Gamma \mathbf{n}_\Gamma$ are independent of this orientation. Now, let $\Gamma \in \mathcal{F}_h^B$ and $K_\Gamma^{(L)} \in \mathcal{T}_h$ be such an element that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial\Omega$. For $v \in H^1(\Omega, \mathcal{T}_h)$ we set

$$v_\Gamma = v|_\Gamma^{(L)} = v|_\Gamma^{(R)} = \text{ the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma,$$

i.e. we define $v|_\Gamma^{(R)}$ by extrapolation.

If $[\cdot]_\Gamma$ and $\langle \cdot \rangle_\Gamma$ appear in an integral of the form $\int_\Gamma \dots \, \mathrm{d}S$, we omit the subscript Γ and write simply $[\cdot]$ and $\langle \cdot \rangle$. For simplicity we shall use the following notation:

$$\int_{\mathcal{F}_h^I} \dots \, \mathrm{d}S = \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma \dots \, \mathrm{d}S$$

and similarly for $\mathcal{F}_h$ and $\mathcal{F}_h^B$.

Let $p \geq 1$ be an integer. The approximate solution will be sought in the space of discontinuous piecewise polynomial functions

$$S_h = \{v;\, v|_K \in P^p(K), \forall K \in \mathcal{T}_h\},$$

where $P^p(K)$ denotes the space of all polynomials on K of degree $\leq p$.

3.3 Discontinuous Galerkin space semidiscretization

We introduce the following forms defined for $u, v, \varphi \in H^2(\Omega, \mathcal{T}_h)$, which yield the SIPG (*Symmetric Interior Penalty Galerkin*) version of the DG approximation.
Symmetric diffusion form:

$$\begin{aligned} a_h(u, v, \varphi) = & \sum_{K \in \mathcal{T}_h} \int_K \beta(u) \nabla v \cdot \nabla \varphi \, dx \\ & - \int_{\mathcal{F}_h^I} \langle \beta(u) \nabla v \rangle \cdot \mathbf{n} [\varphi] \, dS - \int_{\mathcal{F}_h^I} \langle \beta(u) \nabla \varphi \rangle \cdot \mathbf{n} [v] \, dS \\ & - \int_{\mathcal{F}_h^B} \beta(u) \nabla v \cdot \mathbf{n} \varphi \, dS - \int_{\mathcal{F}_h^B} \beta(u) \nabla \varphi \cdot \mathbf{n} v \, dS. \end{aligned}$$

Further we define the *interior and boundary penalty jump terms*:

$$J_h(u, \varphi) = \int_{\mathcal{F}_h^I} \sigma [u][\varphi] \, dS + \int_{\mathcal{F}_h^B} \sigma u \varphi \, dS \tag{6}$$

and the *symmetric right-hand side form*:

$$l_h(u, \varphi)(t) = \int_\Omega g(t) \varphi \, dx - \int_{\mathcal{F}_h^B} \beta(u) \nabla \varphi \cdot \mathbf{n} u_D(t) \, dS + \int_{\mathcal{F}_h^B} \sigma u_D(t) \varphi \, dS. \tag{7}$$

The parameter σ in (6) and (7) is constant on every edge and defined by

$$\sigma|_\Gamma = \frac{C_W}{d(\Gamma)}, \quad \forall\, \Gamma \in \mathcal{F}_h, \tag{8}$$

where $C_W > 0$ is a constant, which must be chosen large enough to ensure coercivity of the diffusion form – cf. Lemma 6.

Finally we define the *convective form*

$$b_h(u,\varphi) = -\sum_{K\in\mathcal{T}_h}\int_K \sum_{s=1}^{2} f_s(u)\frac{\partial\varphi}{\partial x_s}\,dx + \int_{\mathcal{F}_h} H(u^{(L)},u^{(R)},\mathbf{n})[\varphi]\,dS.$$

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

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

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

$$H(u,u,\mathbf{n}) = \sum_{s=1}^{2} f_s(u)\,n_s, \quad \forall u\in\mathbb{R},\ \mathbf{n} = (n_1,n_2)\in B_1.$$

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

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

Definition 1. *We say that u_h is a DGFE solution of the convection-diffusion problem (1) - (3), if*

$$\begin{aligned}
&a)\ u_h \in C^1([0,T];S_h),\\
&b)\ \frac{d}{dt}\big(u_h(t),\varphi_h\big) + b_h\big(u_h(t),\varphi_h\big) + \beta_0 J_h\big(u_h(t),\varphi_h\big) + a_h\big(u_h(t),u_h(t),\varphi_h\big)\\
&\qquad = l_h\big(u_h(t),\varphi_h\big)(t), \quad \forall\varphi_h\in S_h,\ \forall t\in(0,T),\\
&c)\ u_h(0) = u_h^0,
\end{aligned} \tag{9}$$

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

Similarly as in [13] we can show that a sufficiently regular exact solution u of problem (1) satisfies the identity

$$\begin{aligned}
\frac{d}{dt}\big(u(t),\varphi_h\big) + b_h\big(u(t),\varphi_h\big) + \beta_0 J_h\big(u(t),\varphi_h\big) + a_h\big(u(t),u(t),\varphi_h\big)\\
= l_h\big(u(t),\varphi_h\big)(t), \quad \forall\varphi_h\in S_h,\ \forall t\in(0,T),
\end{aligned} \tag{10}$$

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

4 Some necessary results and assumptions

4.1 Regularity of the exact solution

We assume that the weak solution u is sufficiently regular, namely

$$\frac{\partial u}{\partial t} \in L^2\big(0,T;H^{p+1}(\Omega)\big), \tag{11}$$

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

To treat the nonlinear diffusion terms, we need additional regularity assumptions on the solution u: there exists a constant $C_R < \infty$ such that

$$\begin{aligned}
&\|\nabla u(t)\|_{L^\infty(\Omega)} \leq C_R, \quad \text{for all } t \in (0,T),\\
&\|u_t(t)\|_{L^\infty(\Omega)} = \left\|\frac{\partial u}{\partial t}(t)\right\|_{L^\infty(\Omega)} \leq C_R, \quad \text{for a.a. } t \in (0,T),\\
&\|\nabla u_t(t)\|_{L^\infty(\Omega)} \leq C_R, \quad \text{for a.a. } t \in (0,T).
\end{aligned} \tag{12}$$

4.2 Geometry of the mesh

Let us consider a system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$, $h_0 > 0$, of triangulations of the domain Ω with the following properties:

Assumptions (A):

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

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

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

$$h_K \leq C_2\, d(\Gamma), \quad \forall K \in \mathcal{T}_h, \quad \forall \Gamma \subset \partial K,\ \Gamma \in \mathcal{F}_h \quad \forall h \in (0,h_0).$$

(A3) There exists a constant $C_3 > 0$ such that

$$h^p \leq C_3 h_K, \quad \forall K \in \mathcal{T}_h, \quad \forall h \in (0,h_0).$$

Let us note that we do not require the usual conforming properties from the finite element method, particularly, hanging nodes are allowed. In the case of piecewise linear elements (i.e. $p = 1$), condition (A3) reduces to the standard *inverse assumption* of [8] and becomes weaker with growing p. This nonstandard assumption is needed in the proof of Lemmas 15 and 16.

4.3 Some auxiliary results

Throughout this work we denote by C a generic constant independent of h. Now we can state two necessary results needed in the following analysis (cf. [13] and [8]):

Lemma 2 (Multiplicative trace inequality)**.** *There exists a constant $C_M > 0$ independent of h, K such that for all $K \in \mathcal{T}_h$, $v \in H^1(K)$ and $h \in (0,h_0)$*

$$||v||^2_{L^2(\partial K)} \leq C_M\big(||v||_{L^2(K)}|v|_{H^1(K)} + h_K^{-1}||v||^2_{L^2(K)}\big).$$

Lemma 3 (Inverse inequalities)**.** *There exists a constant $C_I > 0$ independent of h, K such that for all $K \in \mathcal{T}_h$ and $v \in P^p(K)$*

$$|v|_{H^1(K)} \le C_I h_K^{-1} ||v||_{L^2(K)},$$
$$\|v\|_{L^\infty(K)} \le C_I h_K^{-1} \|v\|_{L^2(K)}.$$

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

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

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 $\eta(t) = \Pi_h u(t) - u(t) \in H^{p+1}(\Omega, \mathcal{T}_h)$ for $t \in (0, T)$.

Lemma 4. *There exists a constant $C > 0$ independent of h, K such that for all $h \in (0, h_0)$*

$$\begin{aligned}
&a)\ ||\eta||_{L^2(\Omega)} \le C h^{p+1} |u|_{H^{p+1}(\Omega)},\\
&b)\ |\eta|_{H^1(\Omega,\mathcal{T}_h)} \le C h^{p} |u|_{H^{p+1}(\Omega)},\\
&c)\ |\eta|_{H^2(\Omega,\mathcal{T}_h)} \le C h^{p-1} |v|_{H^{p+1}(\Omega)},\\
&d)\ \Big|\Big|\frac{\partial \eta}{\partial t}\Big|\Big|_{L^2(\Omega)} \le C h^{p+1} \Big|\frac{\partial u}{\partial t}\Big|_{H^{p+1}(\Omega)},\\
&e)\ ||\eta||_{L^\infty(\Omega)} \le C h^{p} |u|_{H^{p+1}(\Omega)}.
\end{aligned}$$

Proof. The proof follows from standard approximation results found e.g. in [8]. □

Lemma 5 (Properties of the form J_h)**.** *For all $v, w \in H^1(\Omega, \mathcal{T}_h)$ we have*

$$\begin{aligned}
&a)\ J_h(v, w) \le \big(J_h(v, v)\big)^{1/2} \big(J_h(w, w)\big)^{1/2},\\
&b)\ J_h(\eta, \eta) \le C h^{2p} |u|^2_{H^{p+1}(\Omega)}.
\end{aligned}$$

Proof. The first inequality follows directly from the Cauchy inequality. Statement *b)* follows from the multiplicative trace inequality and approximation results of Lemma 4. □

5 Error analysis

5.1 Properties of the diffusion terms

Throughout the following analysis we shall assume that the constant C_W from (8) satisfies

$$C_W \ge 4\Big(\frac{\beta_1}{\beta_0}\Big)^2 C_M (1 + C_I), \tag{13}$$

where C_M and C_I are constants from Lemma 2 and 3, respectively.

Let us define the form

$$A_h(u, v, w) = a_h(u, v, w) + \beta_0 J_h(v, w), \quad \forall u, v, w \in H^2(\Omega, \mathcal{T}_h),$$

which is linear with respect to the second and third argument and nonlinear with respect to the first argument. Finally, we define the following norm in $H^1(\Omega, \mathcal{T}_h)$:

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

Lemma 6 (Coercivity of A_h)**.** *Let $w : \Omega \to \mathbb{R}$ be an arbitrary measurable function defined almost everywhere in Ω. Under assumption (13) on the constant C_W, we have*

$$\beta_0\|\varphi_h\|^2_{DG} \le A_h(w, \varphi_h, \varphi_h) \tag{14}$$

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

Proof. Since w is measurable, the boundedness and continuity of β imply that $\beta(w)$ is bounded from below by β_0 and $\beta(w) \in L^\infty(\Omega)$. By the definition of the form A_h we get

$$\begin{aligned}
&A_h(w,\varphi,\varphi) = a_h(w,\varphi,\varphi) + \beta_0 J_h(\varphi,\varphi)\\
&\ge \beta_0\|\varphi\|^2_{H^1(\Omega,\mathcal{T}_h)} + \beta_0 J_h(\varphi,\varphi) - 2\int_{\mathcal{F}_h^I}\Big|\big\langle\beta(w)\nabla\varphi\big\rangle\cdot\mathbf{n}[\varphi]\Big|\,dS - 2\int_{\mathcal{F}_h^B}\big|\beta(w)\nabla\varphi\cdot\mathbf{n}\varphi\big|\,dS\\
&\ge 2\beta_0\|\varphi\|^2_{DG} - 2\beta_1\Big(\int_{\mathcal{F}_h^I}\frac{d(\Gamma)}{\Theta}\big\langle|\nabla\varphi|\big\rangle^2\,dS\Big)^{1/2}\Big(\int_{\mathcal{F}_h^I}\frac{\Theta}{d(\Gamma)}[\varphi]^2\,dS\Big)^{1/2}\\
&\quad - 2\beta_1\Big(\int_{\mathcal{F}_h^B}\frac{d(\Gamma)}{\Theta}|\nabla\varphi|^2\,dS\Big)^{1/2}\Big(\int_{\mathcal{F}_h^B}\frac{\Theta}{d(\Gamma)}|\varphi|^2\,dS\Big)^{1/2},
\end{aligned} \tag{15}$$

where $\Theta > 0$ is an arbitrary number. Now using in (15) the fact that for all $\alpha, \beta, \gamma, \delta \in \mathbb{R}$, we have $2(\alpha\gamma + \beta\delta) \le \alpha^2 + \beta^2 + \gamma^2 + \delta^2$, we get

$$A_h(w,\varphi,\varphi) \ge 2\beta_0\|\varphi\|^2_{DG} - \beta_1\omega - \beta_1\frac{\Theta}{C_W}J_h(\varphi,\varphi), \tag{16}$$

where

$$\omega = \int_{\mathcal{F}_h^I}\frac{d(\Gamma)}{\Theta}|\langle\nabla\varphi\rangle|^2\,dS + \int_{\mathcal{F}_h^B}\frac{d(\Gamma)}{\Theta}|\nabla\varphi|^2\,dS.$$

Further using Lemmas 2 and 3, we get

$$\begin{aligned}
\omega &\le \frac{1}{\Theta}\sum_{K\in\mathcal{T}_h} h_K\int_{\partial K}|\nabla\varphi|^2\,dS \le \frac{C_M}{\Theta}\sum_{K\in\mathcal{T}_h} h_K\big(|\varphi|_{H^1(K)}|\nabla\varphi|_{H^1(K)} + h_K^{-1}|\varphi|^2_{H^1(K)}\big)\\
&\le \frac{C_M(1+C_I)}{\Theta}|\varphi|^2_{H^1(\Omega,\mathcal{T}_h)}.
\end{aligned} \tag{17}$$

If we choose

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

and use condition (13), we get from (16) and (17)

$$A_h(w,\varphi,\varphi) \ge 2\beta_0\|\varphi\|^2_{DG} - \frac{\beta_0}{2}\|\varphi\|^2_{H^1(\Omega,\mathcal{T}_h)} - \frac{\beta_0}{2}J_h(\varphi,\varphi) \ge \beta_0\|\varphi\|^2_{DG}. \tag{18}$$

Lemma 7 (Boundedness of A_h)**.** *Let $w : \Omega \to \mathbb{R}$ be an arbitrary measurable function defined almost everywhere in Ω. Then there exists a constant $C > 0$ independent of h such that*

$$A_h(w, v, \varphi_h) \le C\big(\|v\|_{DG} + h|v|_{H^2(\Omega,\mathcal{T}_h)}\big)\|\varphi_h\|_{DG}, \tag{19}$$
$$A_h(w, v_h, \varphi_h) \le C\|v_h\|_{DG}\|\varphi_h\|_{DG}. \tag{20}$$

for all $v \in H^2(\Omega, \mathcal{T}_h)$ and $v_h, \varphi_h \in S_h$.

Proof. We use the fact that $\beta(w) \le \beta_1$ and proceed similarly as in [13]. □

Remark 1. The statement of Lemma 7 is valid also if we replace $\beta(w)$ by another function from $L^\infty(\Omega)$ in the definition of A_h. We shall use this fact in the proof of Lemmas 8 and 13.

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

$$u^*(t) \in S_h, \qquad A_h\big(u(t), u^*(t), \varphi_h\big) = A_h\big(u(t), u(t), \varphi_h\big) \quad \forall\, \varphi_h \in S_h. \tag{21}$$

For simplicity of notation, in what follows, we shall omit the argument t, whenever the role of t is not crucial. The existence of u^* is a consequence of the Lax-Milgram theorem, by the coercivity (Lemma 6) and boundedness (Lemma 7) of the form A_h on the space S_h.

First, we shall derive estimates for 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 8. *There exists a constant $C > 0$ independent of h, such that*

$$\|\chi(t)\|_{DG} \le C\, h^p |u(t)|_{H^{p+1}(\Omega)}, \tag{22}$$
$$\|\chi_t(t)\|_{DG} \le C\, h^p |u_t(t)|_{H^{p+1}(\Omega)} \tag{23}$$

for all $h \in (0, h_0)$ and for a.a. $t \in (0, T)$.

Proof. Let us set $\hat{u} = \Pi_h u$, the L^2-projection of u onto the space S_h. By the coercivity of A_h (14) and the definition of u^*, we obtain

$$\begin{aligned} \beta_0\|\hat{u} - u^*\|_{DG}^2 &\le A_h(u, \hat{u} - u^*, \hat{u} - u^*) \\ &= A_h(u, \hat{u} - u^*, \hat{u} - u^*) + A_h(u, u^* - u, \hat{u} - u^*) \\ &= A_h(u, \hat{u} - u, \hat{u} - u^*). \end{aligned} \tag{24}$$

From Lemmas 7 and 4, b) and c), we obtain

$$\begin{aligned} A_h(u, \hat{u} - u, \hat{u} - u^*) &\le C\big(\|\hat{u} - u\|_{DG} + h|\hat{u} - u|_{H^2(\Omega,\mathcal{T}_h)}\big)\|\hat{u} - u^*\|_{DG} \\ &\le C\, h^p |u|_{H^{p+1}(\Omega)}\|\hat{u} - u^*\|_{DG}. \end{aligned} \tag{25}$$

From (24) and (25) we get

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

Further, in virtue of the regularity of u, Lemmas 4, b) and 5, b), we have

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

Now it is sufficient to use the triangle inequality

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

which implies (22).

Let us deal now with the norm $\|\chi_t\|_{DG}$. By differentiating (21) with respect to time, we get

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} A_h\big(u(t),\chi(t),\varphi_h\big) = \widetilde{a}_h\big(\chi(t),\varphi_h\big) + A_h\left(u(t),\chi_t(t),\varphi_h\right), \quad \forall \varphi_h \in S_h, \tag{26}$$

where

$$\begin{aligned}\widetilde{a}_h\big(v,\varphi\big) = &\sum_{K\in\mathcal{T}_h}\int_K \frac{\partial\beta(u)}{\partial t}\nabla v\cdot\nabla\varphi\,dx\\ &-\int_{\mathcal{F}_h^I}\Big\langle\frac{\partial\beta(u)}{\partial t}\nabla v\Big\rangle\cdot\mathbf{n}\,[\varphi]\,dS - \int_{\mathcal{F}_h^I}\Big\langle\frac{\partial\beta(u)}{\partial t}\nabla\varphi\Big\rangle\cdot\mathbf{n}\,[v]\,dS\\ &-\int_{\mathcal{F}_h^B}\frac{\partial\beta(u)}{\partial t}\nabla v\cdot\mathbf{n}\,\varphi\,dS - \int_{\mathcal{F}_h^B}\frac{\partial\beta(u)}{\partial t}\nabla\varphi\cdot\mathbf{n}\,v\,dS.\end{aligned} \tag{27}$$

Since $\frac{\partial\beta(u)}{\partial t} = \beta'(u)\frac{\partial u}{\partial t} \in L^\infty(\Omega)$ for a.a. $t\in(0,T)$, we have from Remark 1 an estimate for $\widetilde{a}_h$ similar as in Lemma 7:

$$\widetilde{a}_h\big(v,\varphi\big) \le C\big(\|v\|_{DG} + h|v|_{H^2(\Omega,\mathcal{T}_h)}\big)\|\varphi\|_{DG}. \tag{28}$$

Now let $t\in(0,T)$ be fixed. We substitute $\varphi_h := \hat{u}_t(t) - u_t^*(t)$ into (26) and use the coercivity of A_h. Due to (26), we can write (again we omit the argument t)

$$\begin{aligned}\beta_0\|\hat{u}_t-u_t^*\|_{DG}^2 &\le A_h(u,\hat{u}_t-u_t^*,\,\hat{u}_t-u_t^*)\\ &= A_h(u,\hat{u}_t-u_t^*,\,\hat{u}_t-u_t^*) - A_h\left(u,u_t-u_t^*,\hat{u}_t-u_t^*\right) - \widetilde{a}_h\big(u-u^*,\hat{u}_t-u_t^*\big)\\ &= A_h(u,\hat{u}_t-u_t,\,\hat{u}_t-u_t^*) - \widetilde{a}_h\big(u-u^*,\hat{u}_t-u_t^*\big).\end{aligned} \tag{29}$$

From Lemmas 7 and 4, b) and c), we obtain

$$\begin{aligned}A_h(u,\hat{u}_t-u_t,\,\hat{u}_t-u_t^*) &\le C\big(\|\hat{u}_t-u_t\|_{DG} + h|\hat{u}_t-u_t|_{H^2(\Omega,\mathcal{T}_h)}\big)\|\hat{u}_t-u_t^*\|_{DG}\\ &\le C\,h^p|u_t|_{H^{p+1}(\Omega)}\|\hat{u}_t-u_t^*\|_{DG}.\end{aligned} \tag{30}$$

Similarly, due to (28)

$$\begin{aligned}\widetilde{a}_h\big(u-u^*,\hat{u}_t-u_t^*\big) &\le C\big(\|u-u^*\|_{DG} + h|u-u^*|_{H^2(\Omega,\mathcal{T}_h)}\big)\|\hat{u}_t-u_t^*\|_{DG}\\ &\le C\,h^p|u|_{H^{p+1}(\Omega)}\|\hat{u}_t-u_t^*\|_{DG}.\end{aligned} \tag{31}$$

From (29) – (31) we get

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

This and the triangle inequality imply (23). □

Dual problem

In this section we shall derive estimates for the $L^2(\Omega)-$norm of χ and χ_t using the linearised elliptic dual problem: Given $z \in L^2(\Omega)$, for each $t \in (0,T)$ find $\psi(t)$, such that

$$\begin{aligned} -\mathrm{div}\big(\beta(u(t))\nabla\psi(t)\big) &= z \quad \text{in } \Omega, \\ \psi(t)|_{\partial\Omega} &= 0. \end{aligned} \tag{32}$$

The weak formulation of (32) reads: Find $\psi(t) \in H_0^1(\Omega)$ such that

$$\big(\beta(u(t))\nabla\psi(t),\, \nabla v\big) = (z, v), \quad \forall\, v \in H_0^1(\Omega). \tag{33}$$

Lemma 9. *Problem (32) has a unique weak solution $\psi(t)$. Moreover, $\psi(t) \in H^2(\Omega)$ and there exists a constant $C > 0$, independent of z and t, such that for all $t \in (0,T)$*

$$\|\psi(t)\|_{H^2(\Omega)} \le C\|z\|_{L^2(\Omega)}. \tag{34}$$

Proof. Since $\beta(u(t)) \in L^\infty(\Omega)$ for all $t \in (0,T)$, the Lax-Milgram theorem gives us the existence of a unique weak solution $\psi(t) \in H_0^1(\Omega)$ of the dual problem. Let $\varphi \in H_0^1(\Omega)$ be arbitrary. We define the test function $\widetilde{v} := \varphi/\beta(u(t))$. We have

$$\begin{aligned} \|\widetilde{v}\|^2_{H^1(\Omega)} &= \left\| \frac{\varphi}{\beta(u(t))} \right\|^2_{L^2(\Omega)} + \left\| \frac{\nabla\varphi\beta(u(t)) - \varphi\beta'(u(t))\nabla u(t)}{\beta(u(t))^2} \right\|^2_{L^2(\Omega)} \\ &\le \beta_0^{-2}\|\varphi\|^2_{L^2(\Omega)} + \beta_0^{-4}(\beta_1^2 + L^2C_R^2)\|\varphi\|^2_{H^1(\Omega)} \le C\|\varphi\|^2_{H^1(\Omega)}. \end{aligned}$$

Therefore $\widetilde{v} \in H_0^1(\Omega)$ for all $t \in (0,T)$, and we can set $v := \widetilde{v}$ in the weak formulation (33):

$$\Big(\beta(u(t))\nabla\psi(t),\, \frac{\nabla\varphi\beta(u(t)) - \varphi\beta'(u(t))\nabla u(t)}{\beta(u(t))^2}\Big) = \Big(z, \frac{\varphi}{\beta(u(t))}\Big), \quad \forall\, \varphi \in H_0^1(\Omega).$$

Therefore

$$\big(\nabla\psi(t),\, \nabla\varphi\big) = (f, \varphi), \quad \forall\varphi \in H_0^1(\Omega),$$

where

$$f = \frac{1}{\beta(u(t))}\big(z + \beta'(u(t))\nabla u(t) \cdot \nabla\psi(t)\big).$$

This means that $\psi(t) \in H_0^1(\Omega)$ solves, in the weak sense, the problem

$$-\Delta\psi(t) = f, \quad \psi(t)|_{\partial\Omega} = 0.$$

As the domain Ω is convex it follows from [19] that $\psi(t) \in H^2(\Omega)$ for $t \in (0,T)$ and

$$\|\psi(t)\|_{H^2(\Omega)} \le C\|f\|_{L^2(\Omega)} \le C\big(\beta_0^{-1}\|z\|_{L^2(\Omega)} + LC_R|\psi(t)|_{H^1(\Omega)}\big) \le C\|z\|_{L^2(\Omega)}.$$

Here we have used the fact that $|\psi|_{H^1(\Omega)} \le C\|z\|_{L^2(\Omega)}$, which follows from (33) by setting $v := \psi(t)$ and applying the Friedrichs inequality

$$\beta_0^{-1}|\psi|^2_{H^1(\Omega)} \le \big(z, \psi(t)\big) \le \|z\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)} \le C\|z\|_{L^2(\Omega)}|\psi|_{H^1(\Omega)}.$$

□

Let us note that $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$. For convenience, we again omit t in the notation. Let $\psi_h\,(= \psi_h(t))$ be the piecewise linear L^2-projection of the function ψ, i.e. $\psi|_K \in P^1(K)$ and

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

Lemma 10. *There exists a constant independent of h, such that in $(0,T)$*

$$\|\psi - \psi_h\|_{DG} \le C\, h|\psi|_{H^2(\Omega)}.$$

Proof. The proof follows directly from Lemma 2 and approximation results in Lemma 4. □

Now we shall use the dual problem (32) to obtain L^2-optimal error estimates for χ and χ_t.

Lemma 11. *There exists a constant $C > 0$ such that for all $h \in (0, h_0)$ and $t \in (0,T)$*

$$\|\chi\|_{L^2(\Omega)} \quad \le \quad Ch^{p+1}|u|_{H^{p+1}(\Omega)}.$$

Proof. We have

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

The continuity of functions from the space $H^2(\Omega)$ yields

$$[\psi]_\Gamma = 0, \qquad \forall \Gamma \in \mathcal{F}_h^I. \tag{35}$$

By the definition of ψ and (35), for a fixed $z \in L^2(\Omega)$, due to Green's theorem, we have

$$\begin{aligned}
(\chi, z) &= \int_\Omega z\chi \,\mathrm{d}x = -\int_\Omega \operatorname{div}\big(\beta(u)\nabla\psi\big)\chi \,\mathrm{d}x \\
&= \sum_{K\in\mathcal{T}_h} \int_K \beta(u)\nabla\psi\cdot\nabla\chi \,\mathrm{d}x - \int_{\mathcal{F}_h^I} \langle\beta(u)\nabla\psi\rangle\cdot\mathbf{n}\,[\chi]\,\mathrm{d}S - \int_{\mathcal{F}_h^B} \beta(u)\nabla\psi\cdot\mathbf{n}\,\chi\,\mathrm{d}S \\
&= \sum_{K\in\mathcal{T}_h} \int_K \beta(u)\nabla\psi\cdot\nabla\chi \,\mathrm{d}x - \int_{\mathcal{F}_h^I} \langle\beta(u)\nabla\psi\rangle\cdot\mathbf{n}\,[\chi]\,\mathrm{d}S - \int_{\mathcal{F}_h^B} \beta(u)\nabla\psi\cdot\mathbf{n}\,\chi\,\mathrm{d}S \\
&\quad - \int_{\mathcal{F}_h^I} \langle\beta(u)\nabla\chi\rangle\cdot\mathbf{n}\,[\psi]\,\mathrm{d}S - \int_{\mathcal{F}_h^B} \beta(u)\nabla\chi\cdot\mathbf{n}\,\psi\,\mathrm{d}S + \beta_0\int_{\mathcal{F}_h^I} \sigma\,[\psi]\,[\chi]\,\mathrm{d}S + \beta_0\int_{\mathcal{F}_h^B} \sigma\,\psi\,\chi\,\mathrm{d}S,
\end{aligned}$$

i. e.,

$$(\chi, z) = A_h(u, \psi, \chi). \tag{36}$$

Further, the symmetry of A_h and (21) give

$$A_h(u, \psi_h, \chi) = A_h(u, \chi, \psi_h) = A_h(u, u - u^*, \psi_h) = 0. \tag{37}$$

This and Lemmas 7 and 10 imply that for a.a. $t \in (0,T)$

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

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 Lemma 11. □

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 (37). This is the reason, why we are unable to prove optimal error estimates for the nonsymmetric and incomplete variants of the DG scheme (cf. [13]) using the presented technique.

Lemma 12. *Let $\psi(t)$ be the weak solution of (32). Then, under the assumptions on the data of the continuous problem (1), $\psi_t(t) = \frac{\partial \psi(t)}{\partial t} \in H^2(\Omega)$ for a.a. $t \in (0,T)$. Furthermore, there exists a constant $C > 0$ independent of z such that*

$$\|\psi_t(t)\|_{H^2(\Omega)} \leq C\|z\|_{L^2(\Omega)}. \tag{38}$$

Proof. For simplicity, in the following proof we shall use the notation $B(\cdot) := \beta(u(\cdot))$. By formal differentiation of (32) with respect to t, we obtain the identity

$$-\mathrm{div}\big(B(t)\nabla\psi_t(t) + B_t(t)\nabla\psi(t)\big) = 0, \quad \psi_t(t)|_{\partial\Omega} = 0. \tag{39}$$

Since we do not know apriori, whether ψ_t exists, we shall seek a suitable function Ψ such that it satisfies (39), i.e.

$$-\mathrm{div}\big(B(t)\nabla\Psi(t)\big) = \mathrm{div}\big(B_t(t)\nabla\psi(t)\big), \quad \Psi(t)|_{\partial\Omega} = 0. \tag{40}$$

Problem (40) has the same form as the dual problem (32) with a special right-hand side, which, as we shall show, lies in $L^2(\Omega)$. We can therefore apply Lemma 9, which states that there exists a weak solution $\Psi(t)$ of (40), which lies in $H^2(\Omega)$. Finally we shall show that $\Psi(t) = \psi_t(t)$.

First we show that the right-hand side of (40) lies in $L^2(\Omega)$ for a.a. $t \in (0,T)$:

$$\begin{aligned}\big\|\mathrm{div}\big(B_t(t)\nabla\psi(t)\big)\big\|_{L^2(\Omega)} &\leq \big\|\nabla B_t(t)\cdot\nabla\psi(t)\big\|_{L^2(\Omega)} + \big\|B_t(t)\Delta\psi(t)\big\|_{L^2(\Omega)} \\ &\leq \big\|B_t(t)\big\|_{W^{1,\infty}(\Omega)}\big\|\psi(t)\big\|_{H^2(\Omega)} \leq C\big\|B_t(t)\big\|_{W^{1,\infty}(\Omega)}\|z\|_{L^2(\Omega)},\end{aligned} \tag{41}$$

due to (34). Since $B(t) = \beta(u(t))$, we can estimate

$$\begin{aligned}\big\|B_t(t)\big\|_{W^{1,\infty}(\Omega)} &= \|B_t(t)\|_{L^\infty(\Omega)} + \|\nabla B_t(t)\|_{L^\infty(\Omega)} \\ &= \big\|\beta'(u(t))u_t(t)\big\|_{L^\infty(\Omega)} + \big\|\beta''(u(t))u_t(t)\nabla u(t) + \beta'(u(t))\nabla u_t(t)\big\|_{L^\infty(\Omega)} \\ &\leq LC_R + \big\|\beta''\big\|_{L^\infty(\mathbb{R})}C_R^2 + LC_R < \infty, \quad \text{for a.a. } t \in (0,T),\end{aligned} \tag{42}$$

due to assumptions (12) and the properties of β. We note that estimate (42) is independent of t. Hence, $B_t \in L^\infty(0,T;W^{1,\infty}(\Omega))$.

Now we can apply Lemma 9, which states that there exists a solution $\Psi(t) \in H^2(\Omega)$ of (40) and that $\|\Psi(t)\|_{H^2(\Omega)}$ can be estimated by the L^2-norm of the right-hand side, i.e.

$$\|\Psi(t)\|_{H^2(\Omega)} \leq C\big\|\mathrm{div}\big(B_t(t)\nabla\psi(t)\big)\big\|_{L^2(\Omega)} \leq C\|z\|_{L^2(\Omega)}, \text{ for a.a. } t \in (0,T), \tag{43}$$

due to (41), (42). We note that here the constant C is independent of t.

It remains to show that $\Psi(t) = \psi_t(t)$. Let $t \in (0,T)$ and $\delta > 0$ such that $t+\delta \in (0,T)$. For $f : \Omega \times (0,T) \to \mathbb{R}$ we define the difference operator D_δ as

$$D_\delta f(t) = \frac{f(t+\delta) - f(t)}{\delta}.$$

Since $\psi(t) \in H^2(\Omega)$ for all $t \in (0,T)$, we see that $D_\delta\psi(t) \in H^2(\Omega)$ for all $t \in (0,T)$ and $\delta > 0$ sufficiently small. To prove the Lemma, we need to establish the pointwise convergence of $D_\delta\psi(t)$ to $\Psi(t)$ as $\delta \to 0$.

We subtract the dual problem (32) taken at time $t+\delta$ and at time t:

$$-\mathrm{div}\big(B(t+\delta)\nabla\psi(t+\delta) - B(t)\nabla\psi(t)\big) = 0,$$

and to both sides we add the term $\operatorname{div}\big(B(t+\delta)\nabla\psi(t)\big)$ and divide by $\delta > 0$. This yields

$$-\operatorname{div}\big(B(t+\delta)\nabla D_\delta\psi(t)\big) = \operatorname{div}\big(D_\delta B(t)\nabla\psi(t)\big). \tag{44}$$

Subtracting (40) and (44) gives us

$$-\operatorname{div}\big(B(t)\nabla\Psi(t) - B(t+\delta)\nabla D_\delta\psi(t)\big) = \operatorname{div}\big(\big(B_t(t) - D_\delta B(t)\big)\nabla\psi(t)\big).$$

Finally to both sides we add the term $\operatorname{div}\big(B(t)\nabla D_\delta\psi(t)\big)$, which results in

$$-\operatorname{div}\big(B(t)\nabla\Phi_\delta(t)\big) = g_\delta(t), \quad \Phi_\delta(t)|_{\partial\Omega} = 0, \tag{45}$$

where

$$\begin{aligned}\Phi_\delta(t) &:= \Psi(t) - D_\delta\psi(t),\\ g_\delta(t) &:= \operatorname{div}\big(\big(B(t) - B(t+\delta)\big)\nabla D_\delta\psi(t) + \big(B_t(t) - D_\delta B(t)\big)\nabla\psi(t)\big).\end{aligned}$$

Again, problem (45) has the same form as the dual problem (32) with a special right-hand side $g_\delta(t)$, which, as we shall show, lies in $L^2(\Omega)$ and $\|g_\delta(t)\|_{L^2(\Omega)} \to 0$, as $\delta \to 0$. We can therefore apply Lemma 9, which states that $\|\Phi_\delta(t)\|_{H^2(\Omega)}$ can be estimated by $\|g_\delta(t)\|_{L^2(\Omega)}$. The continuous embedding $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$ gives us

$$\|\Phi_\delta(t)\|_{C(\overline{\Omega})} \le C\|\Phi_\delta(t)\|_{H^2(\Omega)} \le C\|g_\delta(t)\|_{L^2(\Omega)} \to 0, \text{ as } \delta \to 0, \quad t \in (0,T).$$

Hence, by the definition of $\Phi_\delta(t)$, it follows that $\psi_t(t)$ exists, $\Psi(t) = \psi_t(t)$ and (43) gives estimate (38).

It remains to estimate $\|g_\delta(t)\|_{L^2(\Omega)}$. We have:

$$\begin{aligned}&\|g_\delta(t)\|_{L^2(\Omega)}\\ &\quad\le \|B(t) - B(t+\delta)\|_{W^{1,\infty}(\Omega)}\|D_\delta\psi(t)\|_{H^2(\Omega)} + \|B_t(t) - D_\delta B(t)\|_{W^{1,\infty}(\Omega)}\|\psi(t)\|_{H^2(\Omega)}.\end{aligned} \tag{46}$$

Since $B_t \in L^\infty(0,T;W^{1,\infty}(\Omega))$, we have $B \in C(0,T;W^{1,\infty}(\Omega))$ and thus

$$\|B(t) - B(t+\delta)\|_{W^{1,\infty}(\Omega)} \to 0, \text{ as } \delta \to 0, \text{ for all } t \in (0,T). \tag{47}$$

Furthermore, for a.a. $t \in (0,T)$, $B_t(t)$ exists and lies in $W^{1,\infty}(\Omega)$. Therefore

$$\|B_t(t) - D_\delta B(t)\|_{W^{1,\infty}(\Omega)} \to 0 \text{ as } \delta \to 0, \text{ for a.a. } t \in (0,T). \tag{48}$$

Finally, we need to estimate $\|D_\delta\psi(t)\|_{H^2(\Omega)}$. Problem (44) for the unknown $D_\delta\psi(t)$ has the same form as the dual problem (32) taken at time $t+\delta$ with a special right-hand side $\operatorname{div}\big(D_\delta B(t)\nabla\psi(t)\big)$. We can therefore apply Lemma 9, which states that $\|D_\delta\psi(t)\|_{H^2(\Omega)}$ can be estimated by the term $\|\operatorname{div}\big(D_\delta B(t)\nabla\psi(t)\big)\|_{L^2(\Omega)}$:

$$\begin{aligned}\|D_\delta\psi(t)\|_{H^2(\Omega)} &\le C\|\operatorname{div}\big(D_\delta B(t)\nabla\psi(t)\big)\|_{L^2(\Omega)} \le C\|D_\delta B(t)\|_{W^{1,\infty}(\Omega)}\|\psi(t)\|_{H^2(\Omega)}\\ &\le C\big(\|B_t(t)\|_{W^{1,\infty}(\Omega)} + \|D_\delta B(t) - B_t(t)\|_{W^{1,\infty}(\Omega)}\big)\|z\|_{L^2(\Omega)} \le C\|z\|_{L^2(\Omega)},\end{aligned} \tag{49}$$

for a.a. $t \in (0,T)$ and all $\delta > 0$ sufficiently small. Estimates (46)-(49) imply $\|g_\delta(t)\|_{L^2(\Omega)} \to 0$, as $\delta \to 0$, for a.a. $t \in (0,T)$. □

Lemma 13. *There exists a constant $C > 0$ independent of h, such that for a.a. $t \in (0,T)$ and all $h \in (0,h_0)$*

$$\|\chi_t(t)\|_{L^2(\Omega)} \leq Ch^{p+1}|u_t(t)|_{H^{p+1}(\Omega)}.$$

Proof. As in the proof of Lemma 11 we can write

$$\|\chi_t(t)\|_{L^2(\Omega)} = \sup_{z\in L^2(\Omega)} \frac{(\chi_t(t),z)}{\|z\|_{L^2(\Omega)}}.$$

Let $z \in L^2(\Omega)$ be arbitrary but fixed (we note that z is independent of time). Due to (36), we have

$$\big(\chi_t(t),z\big) = \Big(\frac{\partial\chi(t)}{\partial t},z\Big) = \frac{\mathrm{d}}{\mathrm{d}t}\big(\chi(t),z\big) = \frac{\mathrm{d}}{\mathrm{d}t}A_h\big(u(t),\psi(t),\chi(t)\big). \tag{50}$$

Differentiating identity (21) with respect to time, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}A_h\big(u(t),\psi_h(t),\chi(t)\big) = 0.$$

This together with (50) gives

$$\begin{aligned}\big(\chi_t(t),z\big) &= \frac{\mathrm{d}}{\mathrm{d}t}A_h\big(u(t),\psi(t)-\psi_h(t),\chi(t)\big) = \widetilde{a}_h\big(\psi(t)-\psi_h(t),\chi(t)\big)\\ &+ A_h\Big(u(t),\frac{\partial}{\partial t}\big(\psi(t)-\psi_h(t)\big),\chi(t)\Big) + A_h\big(u(t),\psi(t)-\psi_h(t),\chi_t(t)\big),\end{aligned} \tag{51}$$

where $\widetilde{a}_h(\cdot,\cdot)$ is defined by (27). We shall now estimate individual terms in (51). Since

$$\Big|\frac{\partial\beta(u)}{\partial t}\Big| = \Big|\beta'(u)\frac{\partial u}{\partial t}\Big| \leq LC_R, \text{ in } Q_T.$$

we have $\frac{\partial\beta(u)}{\partial t} \in L^\infty(\Omega)$ for a.a. $t \in (0,T)$ and we can therefore estimate $\widetilde{a}_h\big(v,\varphi\big)$ similarly as in the proof of Lemma 7 to obtain

$$\widetilde{a}_h\big(v,\varphi\big) \leq C\big(\|v\|_{DG} + h|v|_{H^2(\Omega,\mathcal{T}_h)}\big)\|\varphi\|_{DG},$$

which yields

$$\begin{aligned}\widetilde{a}_h\big(\psi(t)-\psi_h(t),\chi(t)\big) &\leq C\big(\|\psi(t)-\psi_h(t)\|_{DG} + h|\psi(t)-\psi_h(t)|_{H^2(\Omega,\mathcal{T}_h)}\big)\|\chi(t)\|_{DG}\\ &\leq Ch^{p+1}\|z\|_{L^2(\Omega)}|u|_{H^{p+1}(\Omega)}.\end{aligned} \tag{52}$$

From Lemma 7 we immediately see that

$$A_h\Big(u(t),\tfrac{\partial}{\partial t}\big(\psi(t)-\psi_h(t)\big),\chi(t)\Big) \leq Ch|\psi_t(t)|_{H^2(\Omega,\mathcal{T}_h)}\|\chi(t)\|_{DG} \leq Ch^{p+1}\|z\|_{L^2(\Omega)}|u|_{H^{p+1}(\Omega)}. \tag{53}$$

Similarly, we obtain

$$A_h\big(u(t),\psi(t)-\psi_h(t),\chi_t(t)\big) \leq Ch|\psi(t)|_{H^2(\Omega,\mathcal{T}_h)}\|\chi_t(t)\|_{DG} \leq Ch^{p+1}\|z\|_{L^2(\Omega)}|u_t|_{H^{p+1}(\Omega)}. \tag{54}$$

Finally, we combine (51) and estimates (52)-(54), which completes the proof of Lemma 13. □

Lemma 14. *There exists a constant $C > 0$ independent of h, such that for all $h \in (0, h_0)$ and a.a. $t \in (0, T)$*

$$\|\chi(t)\|_{L^\infty(\Omega)} \quad \leq \quad Ch^p |u(t)|_{H^{p+1}(\Omega)}.$$

Proof. For a given h and t the function $\chi(t)$ is a piecewise continuous function on a given finite triangulation $\mathcal{T}_h$, thus there exists an element $K \in \mathcal{T}_h$ such that $\|\chi(t)\|_{L^\infty(\Omega)} = \|\chi(t)\|_{L^\infty(K)}$. By Lemmas 3, 4 e) and 11 we have

$$\begin{aligned}
\|\chi\|_{L^\infty(K)} &\leq \|u - \Pi_h u\|_{L^\infty(K)} + \|\Pi_h u - u^*\|_{L^\infty(K)} \\
&\leq Ch^p |u|_{H^{p+1}(\Omega)} + C_I h_K^{-1} \|\Pi_h u - u^*\|_{L^2(K)} \\
&\leq Ch^p |u|_{H^{p+1}(\Omega)} + C_I h_K^{-1} \big(\|\Pi_h u - u\|_{L^2(K)} + \|u - u^*\|_{L^2(K)} \big) \\
&\leq Ch^p |u|_{H^{p+1}(\Omega)}.
\end{aligned}$$

□

Now we shall establish an important property of the A_h-projection u^* needed in the following analysis.

Lemma 15. *There exists a constant $C_R^* > 0$ independent of h, such that for all $h \in (0, h_0)$ and a.a. $t \in (0, T)$*

$$\|\nabla u^*(t)\|_{L^\infty(\Omega)} \leq C_R^*.$$

Proof. For a given h and t the function $u_h^*(t)$ is a piecewise continuous function on a given finite triangulation $\mathcal{T}_h$, thus there exists an element $K \in \mathcal{T}_h$ such that $\|\nabla u_h^*(t)\|_{L^\infty(\Omega)} = \|\nabla u_h^*(t)\|_{L^\infty(K)}$. Due to the second inverse inequality in Lemma 3, we have

$$\begin{aligned}
\|\nabla u_h^*(t)\|_{L^\infty(K)} &\leq C_I h_K^{-1} \|\nabla u_h^*(t)\|_{L^2(K)} = C_I h_K^{-1} |u_h^*(t)|_{H^1(K)} \\
&\leq C_I h_K^{-1} |u_h^*(t) - u(t)|_{H^1(K)} + C_I h_K^{-1} |u(t)|_{H^1(K)} \\
&\leq \sqrt{2} C_I h_K^{-1} \|\chi(t)\|_{DG} + C_I h_K^{-1} |K|^{1/2} |\nabla u(t)|_{L^\infty(K)}.
\end{aligned}$$

Now we can use estimate (22) and assumption (A3) on the mesh to obtain

$$\begin{aligned}
\|\nabla u_h^*(t)\|_{L^\infty(K)} &\leq Ch_K^{-1} h^p |u(t)|_{H^{p+1}(\Omega)} + Ch_K^{-1} h_K |\nabla u(t)|_{L^\infty(K)} \\
&\leq CC_3 |u|_{H^{p+1}(\Omega)} + CC_R \leq C_R^*,
\end{aligned}$$

where $C_R^* := CC_3 \|u\|_{L^\infty(0,T;H^{p+1}(\Omega))} + CC_R < \infty$ is a constant independent of h and t. □

Lemma 16. *Let $\zeta := u^* - u_h \in S_h$. There exists a constant $C > 0$ such that for all $h \in (0, h_0)$ and a.a. $t \in (0, T)$*

$$\begin{aligned}
&A_h\big(u, u^*, \zeta\big) - A_h\big(u_h, u^*, \zeta\big) - l_h(u, \zeta) + l_h(u_h, \zeta) \\
&\qquad \leq Ch^{2(p+1)} |u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{4} \|\zeta\|^2_{DG}.
\end{aligned} \tag{55}$$

Proof. We break down (55) into individual terms A_i, defined in the sequel, and treat them separately:

$$A_h\big(u, u^*, \zeta\big) - A_h\big(u_h, u^*, \zeta\big) - l_h(u, \zeta) + l_h(u_h, \zeta) =: \sum_{i=1}^{5} A_i.$$

1) First term: Due to the Lipschitz continuity of β and Lemma 15

$$\begin{aligned}
A_1 &:= \sum_{K\in\mathcal{T}_h}\int_K \big(\beta(u)-\beta(u_h)\big)\nabla u^*\cdot\nabla\zeta \,\mathrm{d}x \\
&\le \|\nabla u^*\|_{L^\infty(\Omega)}\sum_{K\in\mathcal{T}_h}\int_K L|u-u_h||\nabla\zeta|\,\mathrm{d}x \le LC_R^*\|u-u_h\|_{L^2(\Omega)}|\zeta|_{H^1(\Omega,\mathcal{T}_h)} \\
&\le C\big(\|\chi\|_{L^2(\Omega)}+\|\zeta\|_{L^2(\Omega)}\big)|\zeta|_{H^1(\Omega,\mathcal{T}_h)}.
\end{aligned}$$

Finally, Young's inequality and Lemma 11 give

$$A_1 \le Ch^{2(p+1)}|u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{16}\|\zeta\|^2_{DG}.$$

2) Second term: Due to the Lipschitz continuity of β, Lemma 15 and Young's inequality,

$$\begin{aligned}
A_2 &:= \int_{\mathcal{F}_h^I}\big\langle\big(\beta(u)-\beta(u_h)\big)\nabla u^*\big\rangle\cdot\mathbf{n}[\zeta]\,\mathrm{d}S \le \|\nabla u^*\|_{L^\infty(\Omega)}\int_{\mathcal{F}_h^I}\big\langle L|u-u_h|\big\rangle|[\zeta]|\,\mathrm{d}S \\
&\le C\Big(\int_{\mathcal{F}_h^I}\frac{d(\Gamma)}{C_W}\big\langle|u-u_h|\big\rangle^2\,dS\Big)^{1/2}\Big(\int_{\mathcal{F}_h^I}\frac{C_W}{d(\Gamma)}[\zeta]^2\,dS\Big)^{1/2} \\
&\le C\Big(\sum_{K\in\mathcal{T}_h}\int_{\partial K} h_K|u-u_h|^2\,\mathrm{d}S\Big)^{1/2}\cdot J_h(\zeta,\zeta)^{1/2} \\
&\le C\sum_{K\in\mathcal{T}_h}\int_{\partial K} h_K|u-u_h|^2\,\mathrm{d}S + \frac{\beta_0}{64}J_h(\zeta,\zeta).
\end{aligned} \tag{56}$$

Now we estimate by Lemma 2 and Young's inequality

$$\begin{aligned}
&C\sum_{K\in\mathcal{T}_h}\int_{\partial K} h_K|u-u_h|^2\,\mathrm{d}S \\
&\qquad\le C\sum_{K\in\mathcal{T}_h} h_K\big(\|u-u_h\|_{L^2(K)}|u-u_h|_{H^1(K)} + h_K^{-1}\|u-u_h\|^2_{L^2(\Omega)}\big) \\
&\qquad\le C\big(h\|u-u_h\|_{L^2(\Omega)}|u-u_h|_{H^1(\Omega,\mathcal{T}_h)} + \|u-u_h\|^2_{L^2(\Omega)}\big) \\
&\qquad= C\big(h\|\chi+\zeta\|_{L^2(\Omega)}|\chi+\zeta|_{H^1(\Omega,\mathcal{T}_h)} + \|\chi+\zeta\|^2_{L^2(\Omega)}\big) \\
&\qquad\le Ch^{2(p+1)}|u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{64}|\zeta|^2_{H^1(\Omega,\mathcal{T}_h)}.
\end{aligned} \tag{57}$$

Finally, combining (56) and (57) results in

$$A_2 \le Ch^{2(p+1)}|u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{32}\|\zeta\|^2_{DG}.$$

3) Third term: Since u is a continuous solution, we have on each interior edge $[u^*] =$

$[u^* - u] = -[\chi]$. Using this and Lemma 14, we get

$$\begin{aligned} A_3 &:= \int_{\mathcal{F}_h^I} \big\langle \big(\beta(u) - \beta(u_h)\big)\nabla\zeta \big\rangle \cdot \mathbf{n}[u^*] \,\mathrm{d}S \le L \int_{\mathcal{F}_h^I} \big\langle |u - u_h||\nabla\zeta| \big\rangle |[\chi]| \,\mathrm{d}S \\ &\le 2L\|\chi\|_{L^\infty(\Omega)} \sum_{K\in\mathcal{T}_h} \int_{\partial K} |u - u_h||\nabla\zeta| \,\mathrm{d}S \\ &\le Ch^p |u|_{H^{p+1}(\Omega)} \Big(\sum_{K\in\mathcal{T}_h} \int_{\partial K} |u - u_h|^2 \,dS \Big)^{1/2} \Big(\sum_{K\in\mathcal{T}_h} \int_{\partial K} |\nabla\zeta|^2 \,dS \Big)^{1/2} \\ &\le C|u|_{H^{p+1}(\Omega)} \Big(\sum_{K\in\mathcal{T}_h} \int_{\partial K} h^p |u - u_h|^2 \,dS \Big)^{1/2} \Big(\sum_{K\in\mathcal{T}_h} \int_{\partial K} h^p |\nabla\zeta|^2 \,dS \Big)^{1/2}. \end{aligned} \tag{58}$$

By assumption (A3), the multiplicative trace and inverse inequalities, we have

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} h^p |\nabla\zeta|^2 \,dS \le C_M \sum_{K\in\mathcal{T}_h} h_K \big(\|\nabla\zeta\|_{L^2(K)} |\nabla\zeta|_{H^1(K)} + h_K^{-1} \|\nabla\zeta\|^2_{L^2(K)} \big) \le C|\zeta|^2_{H^1(\Omega,\mathcal{T}_h)}. \tag{59}$$

Finally, due to assumption (A3) and estimate (57), by Young's inequality we obtain from (58) and (59)

$$\begin{aligned} A_3 &\le C \sum_{K\in\mathcal{T}_h} \int_{\partial K} h_K |u - u_h|^2 \,dS + \frac{\beta_0}{64} |\zeta|^2_{H^1(\Omega,\mathcal{T}_h)} \\ &\le Ch^{2(p+1)} |u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{64} |\zeta|^2_{H^1(\Omega,\mathcal{T}_h)} + \frac{\beta_0}{64} |\zeta|^2_{H^1(\Omega,\mathcal{T}_h)} \\ &\le Ch^{2(p+1)} |u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{16} \|\zeta\|^2_{DG}. \end{aligned}$$

4) Fourth term: We can proceed similarly as in the estimation of A_2 to obtain

$$A_4 := \int_{\mathcal{F}_h^B} \big(\beta(u) - \beta(u_h)\big)\nabla u^* \cdot \mathbf{n}\zeta \,\mathrm{d}S \le Ch^{2(p+1)} |u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{32} \|\zeta\|^2_{DG}.$$

5) Fifth term: We use the fact that $u = u_D$ on $\partial\Omega$:

$$\begin{aligned} A_5 &:= \int_{\mathcal{F}_h^B} \big(\beta(u) - \beta(u_h)\big)\nabla\zeta \cdot \mathbf{n} u^* \,\mathrm{d}S - l_h(u,\zeta) + l_h(u_h,\zeta) \\ &= \int_{\mathcal{F}_h^B} \big(\beta(u) - \beta(u_h)\big)\nabla\zeta \cdot \mathbf{n}(u^* - u_D) \,\mathrm{d}S = \int_{\mathcal{F}_h^B} \big(\beta(u) - \beta(u_h)\big)\nabla\zeta \cdot \mathbf{n}(u^* - u) \,\mathrm{d}S. \end{aligned}$$

Now we can proceed similarly as in the estimate of A_3:

$$A_5 \le Ch^{2(p+1)} |u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{16} \|\zeta\|^2_{DG}.$$

From the derived estimates of $A_1 \dots A_5$ we get the desired estimate (55). □

Lemma 17. *Let u be the solution of the continuous problem (1), u_h the solution of the discrete problem (9), u^* be defined by (21), 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*

$$|b_h(u,\zeta) - b_h(u_h,\zeta)| \le C\|\zeta\|_{DG} \left(h^{p+1} |u|_{H^{p+1}(\Omega)} + \|\zeta\|_{L^2(\Omega)} \right). \tag{60}$$

Proof. The proof can be carried out similarly as in Lemma 4.3 from [15]. □

Theorem 18 (Main theorem)**.** *Let assumptions* (H) *and* (A) *be satisfied and let the constant C_W be chosen in such a way that (13) holds. Let u be the exact solution of problem (1) satisfying the regularity condition (11) and let u_h be the approximate solution defined by (9). Then the error $e_h = u - u_h$ satisfies the estimate*

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

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

Proof. Let u^* be the A_h projection defined by (21) and let χ and ζ be as in Lemmas 8 – 17, i. e. $\chi = u - u^*$, $\zeta = u^* - u_h$. Then $e_h = u - u_h = \chi + \zeta$. Let us subtract $(9, \mathrm{b})$ from (10), substitute $\zeta \in S_h$ for φ_h and use the relation

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

Then we get

$$\begin{aligned}\frac{1}{2}\,\frac{\mathrm{d}}{\mathrm{d}t}\,\|\zeta(t)\|^2_{L^2(\Omega)} &+ A_h\big(u(t), u(t), \zeta(t)\big) - A_h\big(u_h(t), u_h(t), \zeta(t)\big)\\ &= \big[b_h\big(u_h(t),\,\zeta(t)\big) - b_h\big(u(t),\,\zeta(t)\big)\big] - \big(\chi_t(t),\,\zeta(t)\big) + l_h\big(u(t), \zeta(t)\big) - l_h\big(u_h(t), \zeta(t)\big).\end{aligned} \tag{61}$$

The convective terms on the right-hand side term can be estimated by Lemma 17 and Young's inequality as follows (we omit the argument t)

$$\begin{aligned}b_h(u_h, \zeta) - b_h(u, \zeta) &\le C\|\zeta\|_{DG}\left(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\zeta\|_{L^2(\Omega)}\right)\\ &\le \frac{\beta_0}{4}\|\zeta\|^2_{DG} + \frac{C}{\beta_0}\left(h^{2(p+1)}|u|^2_{H^{p+1}(\Omega)} + \|\zeta\|^2_{L^2(\Omega)}\right).\end{aligned}$$

For the second right-hand side term in (61), by the Cauchy and Young's inequalities and Lemma 11, we have

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

Further, we treat the diffusion terms in (61):

$$\begin{aligned}A_h\big(u, u, \zeta\big) &- A_h\big(u_h, u_h, \zeta\big)\\ &= A_h\big(u, \chi, \zeta\big) + A_h\big(u, u^*, \zeta\big) - A_h\big(u_h, u^*, \zeta\big) + A_h\big(u_h, \zeta, \zeta\big)\\ &\ge A_h\big(u, u^*, \zeta\big) - A_h\big(u_h, u^*, \zeta\big) + \beta_0\|\zeta\|^2_{DG},\end{aligned} \tag{62}$$

due to the coercivity of A_h – Lemma 6 – and the definition of u^*, cf. (21). Hence, combining (61) – (62), we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\|\zeta\|^2_{L^2(\Omega)} + \frac{3\beta_0}{4}\|\zeta\|^2_{DG} \;\le\; & C\,h^{2(p+1)}\Big(\frac{1}{\beta_0}\,|u|^2_{H^{p+1}(\Omega)} + |u_t|^2_{H^{p+1}(\Omega)}\Big)\\ &+ C\Big(1 + \frac{1}{\beta_0}\Big)\|\zeta\|^2_{L^2(\Omega)} - \big[A_h\big(u, u^*, \zeta\big) - A_h\big(u_h, u^*, \zeta\big) - l_h(u, \zeta) + l_h(u_h, \zeta)\big].\end{aligned}$$

Now we apply Lemma 16 to obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\zeta\|^2_{L^2(\Omega)} + \frac{3\beta_0}{4}\|\zeta\|^2_{DG} \;\le\; C\,h^{2(p+1)}\Big(\frac{1}{\beta_0}\,|u|^2_{H^{p+1}(\Omega)} + |u_t|^2_{H^{p+1}(\Omega)}\Big) + C\Big(1+\frac{1}{\beta_0}\Big)\|\zeta\|^2_{L^2(\Omega)} + Ch^{2(p+1)}|u|^2_{H^{p+1}(\Omega)} + C\|\zeta\|^2_{L^2(\Omega)} + \frac{\beta_0}{4}\|\zeta\|^2_{DG}.$$

Finally, by rearranging we get a.e. in $(0,T)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\zeta\|^2_{L^2(\Omega)} + \beta_0\|\zeta\|^2_{DG} \;\le\; C\,h^{2(p+1)}\left(|u|^2_{H^{p+1}(\Omega)} + |u_t|^2_{H^{p+1}(\Omega)}\right) + C\Big(1+\frac{1}{\beta_0}\Big)\|\zeta\|^2_{L^2(\Omega)}. \tag{63}$$

The integration of (63) from 0 to $t \in [0,T]$ yields

$$\begin{aligned}\|\zeta(t)\|^2_{L^2(\Omega)} + \beta_0\int_0^t \|\zeta(\vartheta)\|^2_{DG}\,\mathrm{d}\vartheta \\ \le C\,h^{2(p+1)}\Big(\int_0^t |u(\vartheta)|^2_{H^{p+1}(\Omega)}\,\mathrm{d}\vartheta + \int_0^t |u_t(\vartheta)|^2_{H^{p+1}(\Omega)}\,\mathrm{d}\vartheta\Big) \\ + C\Big(1+\frac{1}{\beta_0}\Big)\int_0^t \|\zeta(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta + C\,h^{2(p+1)}|u^0|^2_{H^{p+1}(\Omega)},\end{aligned} \tag{64}$$

since

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

Now we apply Gronwall's Lemma (cf. [13]) to (64), which yields

$$\|\zeta(t)\|^2_{L^2(\Omega)} + \beta_0\int_0^t \|\zeta(\vartheta)\|^2_{DG}\,\mathrm{d}\vartheta \le C\,h^{2(p+1)}\exp\Big(\tilde{C}\Big(1+\frac{1}{\beta_0}\Big)t\Big), \tag{65}$$

where C and $\tilde{C}$ are constants independent of t and h. Since $e_h = \chi + \zeta$, to complete the proof, it is sufficient now to combine (65) with the estimate of $\|\chi(t)\|_{L^2(\Omega)}$ from Lemma 11. □

6 Conclusion

This paper is concerned with the analysis of the discontinuous Galerkin space semidiscretization of a nonstationary convection-diffusion problem with nonlinear diffusion and nonlinear convection, equipped with Dirichlet boundary conditions and an initial condition. We have proven optimal error estimates of order $O(h^{p+1})$ in the $L^\infty(0,T;L^2(\Omega))$-norm for the SIPG method under the assumptions that the piecewise polynomial approximation of degree p is used, the time derivative of the exact solution is sufficiently regular and the solution of the linearized dual problem of the form

$$-\mathrm{div}\big(\beta(u(t))\nabla\psi(t)\big) = z \quad \text{in } \Omega, \quad \psi|_{\partial\Omega} = 0.$$

possesses a solution $\psi(t) \in H^2(\Omega)$ with a time derivative $\psi_t(t) \in H^2(\Omega)$ for any $z \in L^2(\Omega)$. This is true under additional conditions on the nonlinearity $\beta(\cdot)$ and the exact solution u, provided the polygonal domain Ω is convex.

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 and of the dual problem (the case of a polygonal nonconvex domain Ω and/or mixed Dirichlet-Neumann boundary conditions).
- The extension of the derived estimates to three spatial dimensions.
- The investigation of optimal error estimates for other variants of the DGFEM for the diffusion terms, such as the nonsymmetric and incomplete interior penalty Galerkin methods (NIPG and IIPG), where the presented technique cannot be applied.

Acknowledgements This work is a part of the research projects MSM 0021620839. The research was also supported by the project LC06052 of the Ministry of Education of the Czech Republic (Jindřich Nečas Center for Mathematical Modelling).

References

[1] D. N. Arnold: An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, **19**, 742–760 (1982).

[2] D. N. Arnold, F. Brezzi, B. Cockburn, D. Marini: Discontinuos Galerkin methods for elliptic problems. *Discontinuous Galerkin methods. Theory, Computation and Applications. Lecture Notes in Computational Science and Engineering 11*, Springer, Berlin, 89–101 (2000).

[3] D. N. Arnold, F. Brezzi, B. Cockburn, D. Marini: Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, **39**, 1749–1779 (2001).

[4] I. Babuška, C. E. Baumann, J. T. Oden: A discontinuous *hp* finite element method for diffusion problems, 1-D analysis. *Comput. Math. Appl.*, **37**, 103–122 (1999).

[5] F. Bassi, S. Rebay: A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations. *J. Comput. Phys.*, **131**, 267–279 (1997).

[6] F. Bassi, S. Rebay: High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, **138**, 251–285 (1997).

[7] C. E. Baumann, J. T. Oden: A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations. *Int. J. Numer. Methods Fluids*, **31**, 79–95 (1999).

[8] P.G. Ciarlet: *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam (1979).

[9] B. Cockburn, C. W. Shu: TVB Runge–Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II. General framework, *Math. Comp.*. **52**, 411–435 (1989).

[10] C. Dawson, V. Aizinger: A discontinuous Galerkin method for three-dimensional shallow water equations.*J. Sci. Comput.* **22-23**, 245–267 (2005).

[11] V. Dolejší: On the Discontinuous Galerkin Method for the Numerical Solution of the Navier-Stokes Equations. *Int. J. Numer. Methods Fluids,* **45**, 1083–1106 (2004).

[12] V. Dolejší, M. Feistauer: A semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow. *J. Comput. Phys.*, **198**, 727–746 (2004).

[13] V. Dolejší, M. Feistauer: Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems. *Numer. Funct. Anal. Optimiz.,* **26**, 349–383 (2005).

[14] V. Dolejší, M. Feistauer, V. Kučera: On the Discontinuous Galerkin method for the Simulation of Compressible Flow with wide range of Mach numbers . *Comput. Visual. Sci.*, **10**, 17–27 (2007).

[15] V. Dolejší, M. Feistauer, V. Kučera, V. Sobotíková: An optimal $L^\infty(L^2)$-error estimate for the discontinuous Galerkin approximation of a nonlinear non-stationary convectiondiffusion problem. *IMA J. Numer. Anal.,* **28**, 496–521 (2008).

[16] V. Dolejší, M. Feistauer, V. Sobotíková: Analysis of the discontinuous Galerkin method for nonlinear convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.,* **194**, 2709-2733 (2005).

[17] M. Feistauer, V. Kučera: Analysis of the DGFEM for nonlinear convection-diffusion problems. *Electr. Trans. Num. Anal*, **32**, 33-48 (2008).

[18] M. Feistauer, V. Kučera: On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.,* **224**, 208-221 (2007).

[19] P. Grisvard: *Singularities in Boundary Value Problems*. Springer, Berlin (1992).

[20] R. Hartmann, P. Houston: Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations. *Technical Report 2001-42 (SFB 359)*, IWR Heidelberg (2001).

[21] P. Houston, I. Perugia, D. Schötzau: Mixed discontinuous Galerkin approximation of the Maxwell operator, Technical Report 2002/45, University of Leicester, Department of Mathematics. *SIAM J. Numer. Anal.*, **42**, 434–459 (2002).

[22] C. Hu, C. W. Shu: A discontinuous Galerkin finite element method for Hamilton-Jacobi equations. *SIAM J. Sci. Comput.*, **21**, 666–690 (1999).

[23] J. Jaffre, C. Johnson, A. Szepessy: Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws, *Math. Models Methods Appl. Sci.*. **5**, 367–386 (1995).

[24] C. Johnson, J. Pitkäranta: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, **46**, 1–26 (1986).

[25] O. Karakashian, C. Makridakis: A space-time finite element method for the nonlinear Schrödinger equation: the discontinuous Galerkin method. *Math. Comput.*, **67**, 479–499 (1998).

[26] A. Kufner, O. John, S. Fučík,: *Function Spaces*. Academia, Prague (1977).

[27] P. Le Saint, P.–A. Raviart: On a finite element method for solving the neutron transport equation. *Mathematical Aspects of Finite Elements in Partial Differential Equations*, Academic Press, 89–145 (1974).

[28] W. H. Reed, T. R. Hill: Triangular mesh methods for the neutron transport equation. *Technical Report LA–UR–73–479*, Los Alamos Scientific Laboratory (1973).

[29] D. Schötzau, C. Schwab, A. Toselli: Mixed hp-DGFEM for incompressible flows. *SIAM J. Numer. Anal.*, **40**, 2171–2194 (2003).

[30] S. Sun, M.F. Wheeler: $L^2(H^1)$-norm a posteriori error estimation for discontinuous Galerkin approximations of reactive transport problems. *J. Sci. Comput.*, **22-23**, 501–530 (2005).

[31] A. Toselli: HP discontinuous Galerkin approximations for the Stokes problem. *Math. Models Methods Appl. Sci.*, **12**, 1565–1597 (2002).

[32] J.J.W van der Vegt, H. Van der Ven: Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flow, part I. General formulation. *J. Comput. Phys.*, **182**, 546–585 (2002).

[33] M. F. Wheeler:, An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, **15**, 152–161 (1978).

Analysis of space-time discontinuous Galerkin method for nonlinear convection-diffusion problems

Miloslav Feistauer, Václav Kučera, Karel Najzar, and Jaroslava Prokopová

Charles University Prague, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Praha 8, Czech Republic

Published in February 2011,
Numerische Mathematik

Abstract

The paper presents the theory of the discontinuous Galerkin finite element method for the space-time discretization of a nonlinear nonstationary convection-diffusion initial-boundary value problem. The discontinuous Galerkin method is applied separately in space and time using, in general, different space grids on different time levels and different polynomial degrees p and q in space and time dicretization. In the space discretization the nonsymmetric, symmetric and incomplete interior and boundary penalty (NIPG, SIPG, IIPG) approximation of diffusion terms is used. The paper is concerned with the proof of error estimates in "$L^2(L^2)$"- and "DG"-norm formed by the "$L^2(H^1)$"-seminorm and penalty terms. Special space-time interpolation and a special technique has been applied for obtaining optimal error estimates with respect to the time step.

Keywords: nonstationary nonlinear convection-diffusion equation; space-time discontinuous Galerkin finite element discretization; NIPG, SIPG and IIPG treatment of diffusion terms; error estimates

1 Introduction

In a number of complex problems from science and technology (aerospace engineering, turbomachinery, oil recovery, meteorology, environmental protection etc.) we meet the requirement to apply new efficient, robust, reliable and highly accurate numerical methods. It is necessary to develop techniques that allow to realize numerical approximations of strongly nonlinear singularly perturbed systems in domains with a complex geometry, whose solutions contain internal or boundary layers.

An excellent candidate to overcome the mentioned difficulties is the discontinuous Galerkin finite element (DGFE) method, which has become rather popular for the solution of a number of problems.

The DGFE method uses piecewise polynomial approximations of the sought solution on a finite element mesh without any requirement on the continuity between neighbouring elements and can be considered as a generalization of the finite volume and finite element methods. It allows to construct higher order schemes in a natural way and is

suitable for the approximation of discontinuous solutions of conservation laws or solutions of singularly perturbed convection-diffusion problems having steep gradients. This method uses advantages of the finite element method and finite volume schemes with an approximate Riemann solver and can be applied on unstructured grids, which are generated for most complex geometries.

The original DGFE method was first used in [47] for the solution of a neutron transport linear equation and analyzed theoretically in [44] and later in [41]. Nearly simultaneously the DGFE techniques were developed for the numerical solution of elliptic problems or parabolic problems ([58], [2]). Further, the DGFE method was applied to *transport-reaction problems ([12])*, nonlinear conservation laws ([16], [40]), convection-diffusion linear or nonlinear problems ([18], [10], [18], [17], [34], [32]), compressible flow ([7], [8], [9], [20], [22], [36], [57]), simulation of compressible low Mach number flows at incompressible limit ([24], [33]), solution of incompressible viscous flow ([53], [56]), porous media flow ([54]), shallow water flow ([19]), the Hamilton-Jacobi equations ([38]), the Schrödinger equation ([42]) and the Maxwell equations ([37]). Theoretical analysis of various types of the DGFE method applied to elliptic problems can be found, e.g. in [5], [3] and [4]. In [48], DGFE analysis is performed in the case of a parabolic problem with a nonlinear diffusion. In [39], analysis of *hp*-version of the DGFE method applied to stationary advection-diffusion-reaction equations is analyzed.

In the discretization of nonstationary problems, one often uses the *space semidiscretization,* also called the *method of lines.* In this approach, the DGFE discretization with respect to space variables is applied only, whereas time remains continuous. This leads to a large system of ordinary differential equations, which can be solved numerically by a suitable ODE solver. (See, e.g., [48], [10], [16], [25], [26], [23].) In CFD and conservation laws, explicit schemes are often used, which are however conditionally stable. Therefore, it is suitable to apply implicit or semi-implicit methods. In [48] implicit θ-schemes are analyzed, [23] is concerned with the analysis of a semi-implicit linearized scheme for a nonlinear convection-diffusion problem and in [22] and [33] an efficient semi-implicit method for the solution of the compressible Euler equations was developed. However, these methods have low order of accuracy in time. As for higher-order time discretization methods, we can mention the well-known Crank-Nicolson scheme, which is second-order in time. In computational fluid dynamics, Runge-Kutta methods are very popular. However, they are conditionally stable and in connection with the DGFEM the time step is strongly limited by the CFL stability condition. An example of unconditionally stable method is the technique using the backward difference formula (BDF). It was used for the solution of compressible flow, e.g. in [22] and analyzed theoretically in the case of a scalar nonlinear convection-diffusion equation in [27].

The numerical simulation of strongly nonstationary transient problems requires the application of numerical schemes of high order of accuracy in space as well as in time. In the paper [6], a time discretization of arbitrary order was proposed and analyzed. Unfortunately, it is applicable to linear parabolic problems only. In the framework of the space DG semidiscretization, the well-known Runge-Kutta discontinuous Galerkin methods were developed, see e.g. [18]. They are applicable to the numerical solution of a wide class of problems, including nonlinear conservation laws and nonlinear convection-diffusion problems, but they are conditionally stable.

One possibility, how to construct an unconditionally stable numerical schemes of high-order of accuracy is to use the discontinuous Galerkin discretization with respect to both space and time. The discontinuous Galerkin time discretization was introduced and analyzed, e.g. in [28] for the solution of ordinary differential equations. In [30],

[29], [1], [51] and [52] the solution of parabolic problems is carried out with the aid of conforming finite elements in space combined with the DG time discretization. See also the monograph [55]. The works [40], and [57] apply on the other hand to the full DG discretization in the space-time domain. This requires to construct the mesh in the space-time cylinder, which may be quite complicated task for 3D problems.

In this paper we are concerned with the space-time discontinuous Galerkin discretization applied separately in space and in time for the numerical solution of a nonstationary nonlinear convection-diffusion equation. The time interval is split into subintervals and on each time level a different space mesh may be used in general. This approach is suitable particularly in the case when the space mesh adaptivity is performed in the course of increasing time. Moreover, the triangulations used for the space discretization may be nonconforming with hanging nodes. In the discontinuous Galerkin formulation we use the nonsymmetric, symmetric or incomplete version of the discretization of the diffusion terms and interior and boundary penalty (i.e., NIPG, SIPG or IIPG versions). For the space and time discretization, piecewise polynomial approximations of different degrees p and q, respectively, are used. The main subject of the paper is the derivation of error estimates of the space-time DGFE method for the nonstationary initial-boundary value problem with nonlinear convection and linear diffusion. We do not consider a singularly perturbed case with dominating convection, but assume that the diffusion coefficient is a fixed positive constant of order $O(1)$. Under the assumption that the triangulations on all time levels are uniformly shape regular, and the exact solution has some regularity properties, error estimates are derived for the space-time DGFE method.

The structure of the paper is as follows: First, the continuous problem is formulated and the main assumptions are introduced. Further, the discontinuous Galerkin discretization in space and time is described. In the next section, some auxiliary results concerning properties of forms appearing in the definition of the approximate solution are obtained and the abstract error estimate is derived. Then the error estimates of the DG space-time discretization are proven. Finally an outlook of the future work is given.

2 Continuous problem

Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) be a bounded polyhedral domain and $T > 0$. We consider the following initial-boundary value problem: Find $u : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

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

$$u\big|_{\partial\Omega\times(0,T)} = u_D, \tag{2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{3}$$

We assume that $\varepsilon > 0$ and $f_s \in C^1(\mathbb{R}), \quad |f_s'| \le C, \ s = 1, \dots, d$. This means that the fluxes f_s are Lipschitz-continuous in $\mathbb{R}$.

Using techniques from [50], it is possible to prove the existence and uniqueness of a weak solution to problem (1) – (3).

We use the standard notation of function spaces (see, e.g. [43]). If ω is a bounded domain, then we define the Lebesgue spaces

$$\begin{aligned} L^\infty(\omega) &= \{\text{measurable functions } \varphi;\ \|\varphi\|_{L^\infty(\omega)} = \operatorname{ess\,sup}_{x\in\omega} |\varphi(x)| < \infty\}, \\ L^2(\omega) &= \{\text{measurable functions } \varphi;\ \|\varphi\|_{L^2(\omega)} = \Big(\int_\omega |\varphi|^2 \, dx \Big)^{1/2} < \infty\} \end{aligned}$$

and the Sobolev space

$$H^k(\omega) = \{\varphi \in L^2(\omega);\ \|\varphi\|_{H^k(\omega)} = \big(\sum_{|\alpha|\le k} \|D^\alpha \varphi\|^2_{L^2(\omega)}\big)^{1/2} < \infty\},$$

with the seminorm

$$|\varphi|_{H^k(\omega)} = \Big(\sum_{|\alpha|=k} \|D^\alpha \varphi\|^2_{L^2(\omega)}\Big)^{1/2}.$$

We also use the Bochner spaces. Let X be a Banach space with a norm $\|\cdot\|_X$ and a seminorm $|\cdot|_X$ and let s be an integer. Then we define:

$$C([0,T];X) = \{\varphi : [0,T] \to X,\ \text{continuous}, \|\varphi\|_{C([0,T];X)} = \sup_{t\in[0,T]} \|\varphi\|_X < \infty\},$$

$$L^2(0,T;X) = \Big\{\varphi : (0,T) \to X,\ \text{strongly measurable}, \|\varphi\|^2_{L^2(0,T;X)} = \int_0^T \|\varphi\|^2_X\, dt < \infty\Big\},$$

$$H^s(0,T;X) = \Big\{\varphi \in L^2(0,T;X); \|\varphi\|^2_{H^s(0,T;X)} = \int_0^T \sum_{\alpha=0}^{s} \Big\|\frac{\partial^\alpha \varphi}{\partial t^\alpha}\Big\|^2_X < \infty\Big\}.$$

Moreover, we set

$$|\varphi|_{C([0,T];X)} = \sup_{t\in[0,T]} |\varphi|_X,$$

$$|\varphi|_{L^2(0,T;X)} = \Big(\int_0^T |\varphi|^2_X\, dt\Big)^{1/2},$$

$$|\varphi|_{H^s(0,T;X)} = \Big(\int_0^T \Big|\frac{\partial^s \varphi}{\partial t^s}\Big|^2_X\Big)^{1/2}.$$

3 Discretization

3.1 Construction of a mesh in Q_T

In the time interval $[0,T]$ we shall construct a partition formed by time instants $0 = t_0 < \cdots < t_M = T$ and denote $I_m = (t_{m-1}, t_m),\ \tau_m = t_m - t_{m-1}$. We have $[0,T] = \bigcup_{i=1}^M \bar{I}_m,\ I_m \cap I_n = \emptyset$ for $m \neq n$.

For each I_m we consider a partition $\mathcal{T}_{h,m}$ of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed $d-$dimensional simplices (triangles for $d = 2$ and tetrahedra for $d = 3$) with mutually disjoint interiors. We shall call $\mathcal{T}_{h,m}$ a triangulation of Ω. We do not require the standard properties of $\mathcal{T}_{h,m}$ used in the finite element method. This means that we admit the so-called hanging nodes (and in 3D also hanging edges). The partitions $\mathcal{T}_{h,m}$ are in general different for different m.

Let $K, K' \in \mathcal{T}_{h,m}$. We say that K and K' are *neighbouring elements*, if the set $\partial K \cap \partial K'$ has positive $(d-1)$-dimensional measure. We say that $\Gamma \subset K$ is a *face* of K, if it is a maximal connected open subset either of $\partial K \cap \partial K'$, where K' is a neighbouring element to K, or of $\partial K \cap \partial\Omega$. By $\mathcal{F}_{h,m}$ we denote the system of all faces of all elements $K \in \mathcal{T}_{h,m}$. Further, we define the set of all inner faces by $\mathcal{F}^I_{h,m} = \{\Gamma \in \mathcal{F}_{h,m};\ \Gamma \subset \Omega\}$ and by $\mathcal{F}^B_{h,m} = \{\Gamma \in \mathcal{F}_{h,m};\ \Gamma \subset \partial\Omega\}$the set of all boundary faces. Obviously, $\mathcal{F}_{h,m} = \mathcal{F}^I_{h,m} \cup \mathcal{F}^B_{h,m}$.

For each $\Gamma \in \mathcal{F}_{h,m}$ we define a unit normal vector $\boldsymbol{n}_\Gamma$. We assume that for $\Gamma \in \mathcal{F}^B_{h,m}$ the normal $\boldsymbol{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega$. For each face $\Gamma \in \mathcal{F}^I_{h,m}$ the orientation of $\boldsymbol{n}_\Gamma$ is arbitrary but fixed. See Figure 1.

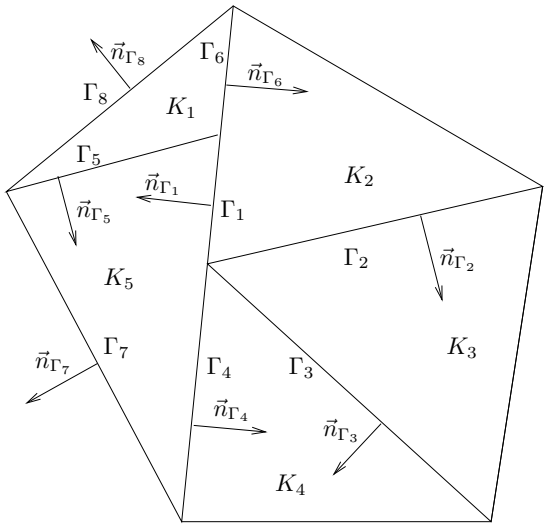


Figure 1: Example of elements $K_l,\ l = 1, \dots, 5$, and faces $\Gamma_l,\ l = 1, \dots, 8$, with the corresponding normals $\boldsymbol{n}_{\Gamma_l}$.

In our further considerations we shall use the following notation. For an element $K \in \mathcal{T}_{h,m}$ we set $h_K = \mathrm{diam}(K),\ h_m = \max_{K \in \mathcal{T}_{h,m}} h_K,\ h = \max_{m=1,\dots,M} h_m$. By ρ_K we denote the radius of the largest d-dimensional ball inscribed into K and by $|K|$ we denote the d-dimensional Lebesgue measure of K. Further, by $d(\Gamma)$ we denote the diameter of $\Gamma \in \mathcal{F}_{h,m}$. Finally, we set $\tau = \max_{m=1,\dots,M} \tau_m$.

3.2 Forms defined on spaces of discontinuous functions

For a function φ defined in $\bigcup_{m=1}^{M} I_m$ we denote

$$\varphi_m^{\pm} = \varphi\,(t_m \pm) = \lim_{t \to t_m \pm} \varphi(t), \quad \{\varphi\}_m = \varphi\,(t_m +) - \varphi\,(t_m -)\,. \tag{4}$$

Over a triangulation $\mathcal{T}_{h,m}$ we define the *broken Sobolev spaces*

$$H^k(\Omega, \mathcal{T}_{h,m}) = \{v; v|_K \in H^k(K)\ \forall\, K \in \mathcal{T}_{h,m}\} \tag{5}$$

equipped with the seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_{h,m})} = \Big(\sum_{K \in \mathcal{T}_{h,m}} |v|^2_{H^k(K)} \Big)^{1/2}. \tag{6}$$

For each face $\Gamma \in \mathcal{F}_{h,m}^I$ there exist two neighbours $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_{h,m}$ such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial K_\Gamma^{(R)}$. We use convention that $\boldsymbol{n}_\Gamma$ is the outer normal to the element $K_\Gamma^{(L)}$ and the inner normal to the element $K_\Gamma^{(R)}$. For $v \in H^1(\Omega, \mathcal{T}_{h,m})$ and $\Gamma \in \mathcal{F}_{h,m}^I$ we

introduce the following notation:

$$\begin{aligned} v|_\Gamma^{(L)} &= \text{the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \\ v|_\Gamma^{(R)} &= \text{the trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma, \\ \langle v \rangle_\Gamma &= \frac{1}{2}\left(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}\right), \\ [v]_\Gamma &= v|_\Gamma^{(L)} - v|_\Gamma^{(R)}. \end{aligned} \tag{7}$$

Now, let $\Gamma \in \mathcal{F}_{h,m}^B$ and $K_\Gamma^{(L)} \in \mathcal{T}_{h,m}$ be such an element that $\Gamma \subset K_\Gamma^{(L)} \cap \partial\Omega$. For $v \in H^1(\Omega, \mathcal{T}_{h,m})$ we define $v|_\Gamma^{(R)}$ by extrapolation, i.e.

$$v|_\Gamma^{(R)} := v|_\Gamma^{(L)} = \text{ the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma. \tag{8}$$

If $[\cdot]_\Gamma$ and $\langle\cdot\rangle_\Gamma$ appear in an integral $\int_\Gamma \ldots \mathrm{d}S$, where $\Gamma \in \mathcal{F}_{h,m}^I$, we usually omit the subscript Γ and write simply $[\cdot]$ and $\langle\cdot\rangle$. Moreover, if $\Gamma \in \mathcal{F}_{h,m}^B$ and $v \in H^1(\Omega, \mathcal{T}_{h,m})$, then $\int_\Gamma v\,\mathrm{d}S$ means $\int_\Gamma v|_\Gamma^{(L)}\,\mathrm{d}S$.

Let $C_W > 0$ be a fixed constant. We introduce the notation

$$\begin{aligned} h(\Gamma) &= \frac{h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}}}{2C_W} \quad \text{for } \Gamma \in \mathcal{F}_{h,m}^I, \\ h(\Gamma) &= \frac{h_{K_\Gamma^{(L)}}}{C_W} \qquad\qquad \text{for } \Gamma \in \mathcal{F}_{h,m}^B. \end{aligned} \tag{9}$$

By $(\cdot,\cdot)$ we denote the scalar product in $L^2(\Omega)$ and by $\|\cdot\|$ we denote the norm in $L^2(\Omega)$. If $\overline{u}$, $\varphi \in H^2(\Omega, \mathcal{T}_{h,m})$, we define the forms

$$\begin{aligned} a_{h,m}(\overline{u},\varphi) = \varepsilon &\sum_{K\in\mathcal{T}_{h,m}} \int_K \nabla\,\overline{u}\cdot\nabla\,\varphi\,\mathrm{d}x \\ &- \varepsilon \sum_{\Gamma\in\mathcal{F}_{h,m}^I} \int_\Gamma \big(\langle\nabla\overline{u}\rangle\cdot\boldsymbol{n}_\Gamma[\varphi] + \theta\langle\nabla\varphi\rangle\cdot\boldsymbol{n}_\Gamma\,[\overline{u}]\big)\mathrm{d}S \\ &- \varepsilon \sum_{\Gamma\in\mathcal{F}_{h,m}^B} \int_\Gamma \big(\nabla\overline{u}\cdot\boldsymbol{n}_\Gamma\,\varphi + \theta\,\nabla\varphi\cdot\boldsymbol{n}_\Gamma\,\overline{u}\big)\mathrm{d}S, \end{aligned} \tag{10}$$

$$J_{h,m}(\overline{u},\varphi) = \sum_{\Gamma\in\mathcal{F}_{h,m}^I} h(\Gamma)^{-1}\int_\Gamma [\overline{u}]\,[\varphi]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_{h,m}^B} h(\Gamma)^{-1}\int_\Gamma \overline{u}\,\varphi\,\mathrm{d}S, \tag{11}$$

$$A_{h,m} = a_{h,m} + \varepsilon J_{h,m}, \tag{12}$$

$$\begin{aligned} b_{h,m}(\overline{u},\varphi) = &- \sum_{K\in\mathcal{T}_{h,m}} \int_K \sum_{s=1}^d f_s(\overline{u})\,\frac{\partial\varphi}{\partial x_s}\,\mathrm{d}x \\ &+ \sum_{\Gamma\in\mathcal{F}_{h,m}^I} \int_\Gamma H\left(\overline{u}|_\Gamma^{(L)}, \overline{u}|_\Gamma^{(R)}, \boldsymbol{n}_\Gamma\right)[\varphi]\big|_\Gamma\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_{h,m}^B} \int_\Gamma H\left(\overline{u}|_\Gamma^{(L)}, \overline{u}|_\Gamma^{(L)}, \boldsymbol{n}_\Gamma\right)\varphi|_\Gamma^{(L)}\,\mathrm{d}S. \end{aligned} \tag{13}$$

Here H is a numerical flux. We assume that it has the following properties.

(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:

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

(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*:

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

Finally, the right-hand side form is defined on the basis of data:

$$\ell_{h,m}(\varphi) = (g,\varphi) + \varepsilon \sum_{\Gamma \in \mathcal{F}_{h,m}^B} \Big(\frac{C_W}{h_{K_\Gamma^{(L)}}} \int_\Gamma u_D\, \varphi \,\mathrm{d}S - \theta \int_\Gamma \nabla\varphi \cdot \boldsymbol{n}_\Gamma u_D \,\mathrm{d}S \Big). \tag{14}$$

In the above forms we take $\theta = -1,\ \theta = 0,\ \theta = 1$ and obtain the nonsymmetric (NIPG), incomplete (IIPG) and symmetric (SIPG) variants of the approximation of the diffusion terms, respectively.

In the space $H^1(\Omega, \mathcal{T}_{h,m})$, the following norm will be used:

$$\|\varphi\|_{\mathrm{DG},m} \;=\; \Big(\sum_{K \in \mathcal{T}_{h,m}} |\varphi|^2_{H^1(K)} + J_{h,m}(\varphi,\varphi) \Big)^{1/2}. \tag{15}$$

3.3 Discrete problem

Let $p,\ q \ge 1$ be integers. For each $m = 1,\dots,M$ we define the finite-dimensional space

$$S^p_{h,m} = \left\{ \varphi \in L^2(\Omega); \varphi|_K \in P^p(K)\ \forall\, K \in \mathcal{T}_{h,m} \right\}. \tag{16}$$

By Π_m we denote the $L^2(\Omega)$-projection on $S^p_{h,m}$, i.e., if $\varphi \in L^2(\Omega)$, then $\Pi_m\varphi \in S^p_{h,m}$ and

$$(\Pi_m\varphi - \varphi, \psi) = 0, \quad \forall\, \psi \in S^p_{h,m}. \tag{17}$$

The approximate solution will be sought in the space

$$S^{p,q}_{h,\tau} = \Big\{ \varphi \in L^2(Q_T); \varphi\big|_{I_m} = \sum_{i=0}^{q} t^i\, \varphi_i \quad \text{with } \varphi_i \in S^p_{h,m},\ m = 1,\dots,M \Big\}. \tag{18}$$

In what follows we shall use the notation $U' = \partial U/\partial t, u' = \partial u/\partial t, D^{q+1} = \partial^{q+1}/\partial t^{q+1}$.

Definition 1. *We say that the function U is an approximate solution of problem (1) – (3), if $U \in S^{p,q}_{h,m}$ and*

$$\int_{I_m} \big((U',\varphi) + A_{h,m}(U,\varphi) + b_{h,m}(U,\varphi) \big)\,\mathrm{d}t + \big(\{U\}_{m-1}, \varphi^+_{m-1} \big) = \int_{I_m} \ell_{h,m}(\varphi)\,\mathrm{d}t \tag{19}$$
$$\forall\, \varphi \in S^{p,q}_{h,\tau}, \quad \forall\, m = 1,\dots,M, \quad U_0^- = \Pi_1 u^0.$$

It is possible to show that the exact sufficiently regular solution u satisfies the identity

$$\int_{I_m} \big((u',\varphi) + A_{h,m}(u,\varphi) + b_{h,m}(u,\varphi)\big)\,\mathrm{d}t + \big(\{u_{m-1}\},\varphi_{m-1}^+\big) = \int_{I_m} \ell_{h,m}(\varphi)\,\mathrm{d}t, \quad \forall\,\varphi \in S_{h,\tau}^{p,q}, \quad \forall\, m = 1,\dots,M, \tag{20}$$

if we set $u(0-) = u(0)$.

Remark 1. *It is also possible to consider $q = 0$. In this case, scheme (19) represents a version of the backward Euler method. Since it can be analyzed in a similar way as, for example, in [23], we shall be concerned only with $q \geq 1$.*

In the error analysis we shall use the $S_{h,\tau}^{p,q}$-interpolation π of functions $v \in L^2(Q_T)$ defined by

$$\begin{aligned} &\text{a)} \quad \pi\, v \in S_{h,\tau}^{p,q}, \\ &\text{b)} \quad (\pi\, v)\,(t_m-) = \Pi_m\, v(t_m-), \\ &\text{c)} \quad \int_{I_m} (\pi v - v,\, \varphi^*)\,\mathrm{d}t = 0, \quad \forall\,\varphi^* \in S_{h,\tau}^{p,q-1}, \quad \forall\, m = 1,\dots,M. \end{aligned} \tag{21}$$

In [32], Lemma 4, it was proven that πu is uniquely determined. Moreover, by [32], Lemma 9,

$$\pi u|_{I_m} = \pi(\Pi_m u)|_{I_m}. \tag{22}$$

Our main goal will be the derivation of the estimation of the error $e = U - u$, which can be expressed in the form

$$e = \xi + \eta, \tag{23}$$

where

$$\xi = U - \pi u \in S_{h,\tau}^{p,q}, \quad \eta \;=\; \pi u - u. \tag{24}$$

Then, in virtue of (19) and (20),

$$\begin{aligned} \int_{I_m} \big((\xi',\varphi) + A_{h,m}(\xi,\varphi)\big)\mathrm{d}t + \big(\{\xi_{m-1}\},\varphi_{m-1}^+\big) &= \int_{I_m} \big(b_{h,m}(u,\varphi) - b_{h,m}(U,\varphi)\big)\mathrm{d}t \\ - \int_{I_m} \big((\eta',\varphi) + A_{h,m}(\eta,\varphi)\big)\mathrm{d}t - \big(\{\eta\}_{m-1},\varphi_{m-1}^+\big), &\quad \forall\varphi \in S_{h,\tau}^{p,q}. \end{aligned} \tag{25}$$

4 Abstract error estimate

In this section we shall be concerned with the derivation of error estimates in terms of interpolation error.

4.1 Assumptions on the triangulation

In our further considerations, by C and c we shall denote positive generic constants, independent of $h, \tau, K, \varepsilon, u, U$, which can attain different values in different places. In

the sequel, we shall consider a system of triangulations $\mathcal{T}_{h,m}$, $m = 1, \ldots, M$, $h \in (0, h_0)$, which is *shape regular* and *locally quasiuniform*:

$$\frac{h_K}{\rho_K} \leq C_R, \quad K \in \mathcal{T}_{h,m}, \quad m = 1, \ldots, M, \quad h \in (0, h_0), \tag{26}$$

$$h_K \leq C_Q h_{K'}, \quad \text{for neighbouring elements } K, K' \in \mathcal{T}_{h,m}. \tag{27}$$

Then there exist positive constants C_-, C_+ such that

$$C_- h_K \leq h(\Gamma) \leq C_+ h_K, \quad \Gamma \in \mathcal{F}_{h,m}, \Gamma \subset K \in \mathcal{T}_{h,m},\ h \in \mathcal{T}_{h,m},\ m = 1, \ldots, M. \tag{28}$$

4.2 Auxiliary results

In the analysis of the DGFEM we use the following important tools.

Multiplicative trace inequality: There exists a constant $C_M > 0$ independent of v, h, K and M such that

$$\begin{aligned} \|v\|^2_{L^2(\partial K)} &\leq C_M \left(\|v\|_{L^2(K)} \, |v|_{H^1(K)} + h_K^{-1} \|v\|^2_{L^2(K)} \right), \\ v &\in H^1(K),\ K \in \mathcal{T}_{h,m},\ h \in (0, h_0),\ m = 1, \ldots, M. \end{aligned} \tag{29}$$

Inverse inequality: There exists a constant $C_I > 0$ independent of v, h, K and M such that

$$|v|_{H^1(K)} \leq C_I h_K^{-1} \|v\|_{L^2(K)}, \qquad v \in P^p(K),\ K \in \mathcal{T}_{h,m},\ h \in (0, h_0),\ m = 1, \ldots, M. \tag{30}$$

(For proofs, see, e.g. [25] and [11].)

Coercivity of the form $A_{h,m}$: It holds

$$A_{h,m}(\xi, \xi) \geq \frac{\varepsilon}{2} \|\xi\|^2_{\mathrm{DG},m} \tag{31}$$

provided

$$\begin{aligned} &C_W > 0 \text{ for NIPG}, \\ &C_W \geq C_M (1 + C_I)\,(1 + C_Q) \text{ for IIPG}, \\ &C_W \geq 2 C_M (1 + C_I)\,(1 + C_Q) \text{ for SIPG}. \end{aligned} \tag{32}$$

(See, [31].)

Consistency of $b_{h,m}$: For any $\varphi \in S^{p,q}_{h,\tau}$ and $k > 0$,

$$\begin{aligned} |b_{h,m}(u, \varphi) - b_{h,m}(U, \varphi)| &\leq C \|\varphi\|_{\mathrm{DG},m} \left(\|\xi\|^2 + \tilde{\sigma}^2_m(\eta) \right)^{1/2} \\ &\leq \frac{\varepsilon}{k} \|\varphi\|^2_{\mathrm{DG},m} + \frac{C}{\varepsilon} \left(\|\xi\|^2 + \tilde{\sigma}^2_m(\eta) \right), \end{aligned} \tag{33}$$

where

$$\tilde{\sigma}^2_m(\eta) = \sum_{K \in \mathcal{T}_{h,m}} \left(\|\eta\|^2_{L^2(K)} + h_K^2 |\eta|^2_{H^1(K)} \right). \tag{34}$$

(The constant C in the last expression depends, of course, on k.) The proof can be carried out in a similar way as in [21] or [26].

4.3 Derivation of estimates for ξ

Let us substitute $\varphi := \xi$ in (25) and analyze individual terms. A simple calculation yields

$$2\int_{I_m}(\xi',\xi)\,\mathrm{d}t + 2\left(\{\xi\}_{m-1},\xi_{m-1}^+\right) = \int_{I_m}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2\mathrm{d}t + 2\left(\{\xi\}_{m-1},\xi_{m-1}^+\right) \tag{35}$$
$$= \|\xi_m^-\|^2 - \|\xi_{m-1}^+\|^2 + 2\left(\xi_{m-1}^+ - \xi_{m-1}^-,\xi_{m-1}^+\right)$$

and

$$2\left(\xi_{m-1}^+ - \xi_{m-1}^-,\xi_{m-1}^+\right) = \left\|\xi_{m-1}^+\right\|^2 + \left\|\{\xi\}_{m-1}\right\|^2 - \left\|\xi_{m-1}^-\right\|^2. \tag{36}$$

Hence,

$$2\int_{I_m}\left(\xi',\xi\right)\mathrm{d}t + 2\left(\{\xi\}_{m-1},\xi_{m-1}^+\right) = \left\|\xi_m^-\right\|^2 - \left\|\xi_{m-1}^-\right\|^2 + \left\|\{\xi\}_{m-1}\right\|^2. \tag{37}$$

Further, we shall be concerned with estimates of the right-hand side of (25). In the same way as in [26], Lemma 9, using Cauchy inequality, multiplicative trace inequality and inverse inequality, we can show that for $\varphi \in S_{h,\tau}^{p,q}$ we have

$$\left|A_{h,m}(\eta,\varphi)\right| \le C_A\varepsilon\|\varphi\|_{\mathrm{DG},m}\sigma_m(\eta), \tag{38}$$

where

$$\sigma_m^2(\eta) = \|\eta\|_{\mathrm{DG},m}^2 + \sum_{K\in\mathcal{T}_{h,m}} h_K^2|\eta|_{H^2(K)}^2. \tag{39}$$

By Young's inequality, for $k > 0$,

$$|A_{h,m}(\eta,\varphi)| \le \frac{\varepsilon}{k}\,\|\varphi\|_{\mathrm{DG},m}^2 + C\varepsilon\sigma_m^2(\eta). \tag{40}$$

Now (25), where we set $\varphi := \xi$, relation (37) and estimates (31), (33), (40) imply that

$$\left\|\xi_m^-\right\|^2 - \left\|\xi_{m-1}\right\|^2 + \left\|\{\xi\}_{m-1}^-\right\|^2 + \varepsilon\int_{I_m}\|\xi\|_{\mathrm{DG},m}^2\,\mathrm{d}t \tag{41}$$
$$\le -2\int_{I_m}(\eta',\xi)\,\mathrm{d}t - 2\left(\{\eta\}_{m-1},\xi_{m-1}^+\right) + \frac{2\varepsilon}{k}\int_{I_m}\|\xi\|_{\mathrm{DG},m}^2\,\mathrm{d}t$$
$$+\ \frac{C}{\varepsilon}\int_{I_m}\|\xi\|^2\,\mathrm{d}t + C\int_{I_m}\left(\varepsilon\sigma_m^2(\eta) + \frac{1}{\varepsilon}\tilde{\sigma}_m^2(\eta)\right)\mathrm{d}t.$$

Further, we shall be concerned with the expression

$$\int_{I_m}(\eta',\xi)\,\mathrm{d}t + \left(\{\eta\}_{m-1},\varphi_{m-1}^+\right).$$

Integration by parts yields

$$\int_{I_m}(\eta',\xi)\,\mathrm{d}t = \left(\eta_m^-,\xi_m^-\right) - \left(\eta_{m-1}^+,\xi_{m-1}^+\right) - \int_{I_m}(\eta,\xi')\,\mathrm{d}t. \tag{42}$$

Since $\xi' \in S_{h,\tau}^{p,q-1}$ and $\eta = \pi u - u$, it follows from the definition of π that

$$\int_{I_m}(\eta,\xi')\,\mathrm{d}t = 0.$$

Thus,

$$\begin{aligned}\int_{I_m} (\eta', \xi)\,\mathrm{d}t + \big(\{\eta\}_{m-1}, \xi_{m-1}^+\big) \\ = \big(\eta_m^-, \xi_m^-\big) - \big(\eta_{m-1}^+, \xi_{m-1}^+\big) + \big(\eta_{m-1}^+, \xi_{m-1}^+\big) - \big(\eta_{m-1}^-, \xi_{m-1}^+\big)\,.\end{aligned} \tag{43}$$

Further, since ξ_m^-, $\xi_{m-1}^- \in S_{h,m}^p$, (21), b) and the definition of Π_m imply that

$$\big(\eta_m^-, \xi_m^-\big) = 0 \text{ and } (\eta_{m-1}^-, \xi_{m-1}^-) = 0. \tag{44}$$

Moreover,

$$\begin{aligned}\big|\big(\eta_{m-1}^-, \xi_{m-1}^+\big)\big| &= \big|\big(\eta_{m-1}^-, \xi_{m-1}^+ - \xi_{m-1}^-\big)\big| = \big|\big(\eta_{m-1}^-, \{\xi\}_{m-1}\big)\big| \\ &\le \frac{1}{2}\big(\big\|\{\xi\}_{m-1}\big\|^2 + \big\|\eta_{m-1}^-\big\|^2\big).\end{aligned} \tag{45}$$

From (43) – (45) we find that

$$\Big|\int_{I_m} (\eta', \xi)\,\mathrm{d}t + \big(\{\eta\}_{m-1}, \xi_{m-1}^+\big)\Big| \le \frac{1}{2}\big\|\{\xi\}_{m-1}\big\|^2 + \frac{1}{2}\big\|\eta_{m-1}^-\big\|^2. \tag{46}$$

This and (41) imply that

$$\begin{aligned}\big\|\xi_m^-\big\|^2 - \big\|\xi_{m-1}^-\big\|^2 + \varepsilon\Big(1 - \frac{2}{k}\Big)\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2\,\mathrm{d}t \\ \le \frac{C}{\varepsilon}\int_{I_m} \|\xi\|^2\,\mathrm{d}t + 2\big\|\eta_{m-1}^-\big\|^2 + C\int_{I_m} R_m(\eta)\,\mathrm{d}t,\end{aligned} \tag{47}$$

where

$$R_m(\eta) = \varepsilon\sigma_m^2(\eta) + \frac{1}{\varepsilon}\tilde{\sigma}_m^2(\eta). \tag{48}$$

In what follows, it will be necessary to estimate the terms with η and $\int_{I_m} \|\xi\|^2\,\mathrm{d}t$.

4.4 Estimation of $\displaystyle\int_{I_m} \|\xi\|^2\,\mathrm{d}t$

By $\mathcal{P}^q$ we shall denote the set of polynomials in $t \in \mathbb{R}$ of degree $\le q$. In the interval $(0, 1]$ we shall consider the Gauss–Radau quadrature formula

$$\int_0^1 \varphi(t)\,\mathrm{d}t \approx \sum_{i=1}^{q+1} w_i\,\varphi(\vartheta_i), \tag{49}$$

where $0 < \vartheta_1 < \cdots < \vartheta_{q+1} = 1$ are the Radau integration points and $w_i > 0$ are the Radau weights. (We can refer, for example, to formulas from [46], transformed from the interval $[-1, 1)$ to $(0, 1]$.) Formula (49) is transformed to the interval $(t_{m-1}, t_m]$, which yields

$$\int_{I_m} \varphi(t)\,\mathrm{d}t \approx \tau_m \sum_{i=1}^{q+1} w_i \varphi\big(t^{m,i}\big), \tag{50}$$

where $t^{m,i} = t_{m-1} + \tau_m\,\vartheta_i$. Formulas (49), (50) are exact for polynomials of degree $\le 2q$. In [1], the following result was proven:

Lemma 1. *Let $p \in \mathcal{P}^q$ and let $\tilde{p} \in \mathcal{P}^q$ be the Lagrange interpolation of the function $\tau_m\, p(t)/(t - t_{m-1})$ at the points $t^{m,i}$, $i = 1, \dots, q+1$:*

$$\tilde{p}(t^{m,i}) = \tau_m\, p\big(t^{m,i}\big)\big/\big(t^{m,i} - t_{m-1}\big) = p\big(t^{m,i}\big)\,\vartheta_i^{-1},\ \ i = 1, \dots, q+1.$$

Then

$$\int_{I_m} p'\,\tilde{p}\,\mathrm{d}t + p(t_{m-1})\,\tilde{p}(t_{m-1}) = \frac{1}{2}\Big(p^2(t_m) + \sum_{i=1}^{q+1} w_i\,\vartheta_i^{-2}\,p^2(t^{m,i})\Big). \tag{51}$$

Now, by $\tilde{\xi}$ we shall denote the Lagrange interpolation of $\tau_m\xi(t)/(t - t_{m-1})$ at the points $t^{m,i}$, $i = 1, \dots, q+1$. This means that for each $x \in \Omega$, the function $\tilde{\xi}(\cdot, x)$ is a polynomial (in t) of degree $\le q$. In what follows we shall denote

$$\|\xi\|_m^2 = \tau_m \sum_{i=1}^{q+1} w_i\,\vartheta_i^{-1}\big\|\xi(t^{m,i})\big\|^2. \tag{52}$$

Let us set $\varphi := \tilde{\xi}$ in (25). Then we get

$$\begin{aligned}
&\underbrace{\int_{I_m} (\xi', \tilde{\xi})\,\mathrm{d}t + \big(\xi_{m-1}^+, \tilde{\xi}_{m-1}^+\big)}_{\text{(a)}} + \underbrace{\int_{I_m} A_{h,m}(\xi, \tilde{\xi})\,\mathrm{d}t}_{\text{(b)}} \\
&\quad = \underbrace{\big(\xi_{m-1}^-, \tilde{\xi}_{m-1}^+\big)}_{\text{(c)}} - \underbrace{\int_{I_m} (\eta', \tilde{\xi})\,\mathrm{d}t - \big(\{\eta\}_{m-1}, \tilde{\xi}_{m-1}^+\big)}_{\text{(d)}} - \underbrace{\int_{I_m} A_{h,m}(\eta, \tilde{\xi})\,\mathrm{d}t}_{\text{(e)}} \\
&\qquad + \underbrace{\int_{I_m} \Big(b_{h,m}(u, \tilde{\xi}) - b_{h,m}(U, \tilde{\xi})\,\mathrm{d}t\Big)}_{\text{(f)}}.
\end{aligned} \tag{53}$$

In what follows, we shall analyze individual terms (a) – (f).

(a) By Fubini's theorem and (51),

$$\begin{aligned}
&\int_{I_m} \big(\xi', \tilde{\xi}\big)\,\mathrm{d}t + \big(\xi_{m-1}^+, \tilde{\xi}_{m-1}^+\big) = \int_\Omega \Big(\int_{t_{m-1}}^{t_m} \xi'\,\tilde{\xi}\,\mathrm{d}t + \xi_{m-1}^+\,\tilde{\xi}_{m-1}^+\Big)\mathrm{d}x \\
&\quad = \int_\Omega \frac{1}{2}\Big((\xi_m^-)^2 + \sum_{i=1}^{q+1} w_i\,\vartheta_i^{-2}\big(\xi(t^{m,1})\big)^2\Big)\mathrm{d}x = \frac{1}{2}\Big(\big\|\xi_m^-\big\|^2 + \sum_{i=1}^{q+1} w_i\,\vartheta_i^{-2}\big\|\xi(t^{m,i})\big\|^2\Big).
\end{aligned} \tag{54}$$

Hence, since $\vartheta_i^{-1} \ge 1$, in view of the notation (52), we get the inequality

$$\int_{I_m} \big(\xi', \tilde{\xi}\big)\,\mathrm{d}t + \big(\xi_{m-1}^+, \tilde{\xi}_{m-1}^+\big) \ge \frac{1}{2}\Big(\|\xi_m^-\|^2 + \frac{1}{\tau_m}\|\xi\|_m^2\Big). \tag{55}$$

(b) We use the following lemma:

Lemma 2. *Under assumptions (32) we have*

$$\int_{I_m} A_{h,m}(\xi, \tilde{\xi})\,\mathrm{d}t \ge \frac{\varepsilon}{2}\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2\mathrm{d}t. \tag{56}$$

Proof. In view of (10) and (12),

$$\begin{aligned}\int_{I_m} A_{h,m}(\xi,\tilde{\xi})\,\mathrm{d}t &= \varepsilon\int_{I_m}\sum_{K\in\mathcal{T}_{h,m}}\int_K \nabla\xi\cdot\nabla\tilde{\xi}\,\mathrm{d}x\mathrm{d}t\\ &-\varepsilon\int_{I_m}\sum_{\Gamma\in\mathcal{F}^I_{h,m}}\int_\Gamma\Big(\langle\nabla\xi\rangle\cdot\boldsymbol{n}_\Gamma[\tilde{\xi}]+\theta\langle\nabla\tilde{\xi}\rangle\cdot\boldsymbol{n}_\Gamma[\xi]\Big)\,\mathrm{d}S\mathrm{d}t\\ &-\varepsilon\int_{I_m}\sum_{\Gamma\in\mathcal{F}^B_{h,m}}\int_\Gamma\Big(\nabla\xi\cdot\boldsymbol{n}_\Gamma\tilde{\xi}+\theta\nabla\tilde{\xi}\cdot\boldsymbol{n}_\Gamma\xi\Big)\,\mathrm{d}S\mathrm{d}t+\ \varepsilon\int_{I_m}J_{h,m}(\xi,\tilde{\xi})\,\mathrm{d}t.\end{aligned}$$

The expressions $\xi\big|_\Gamma$, $[\xi]_\Gamma$, $\tilde{\xi}\big|_\Gamma$, $[\tilde{\xi}]_\Gamma$, $\nabla\xi$ and $\nabla\tilde{\xi}$ are polynomials in t of degree $\le q$. Hence, $\int_K\nabla\xi\cdot\nabla\tilde{\xi}\,\mathrm{d}x$, $\int_\Gamma[\xi]_\Gamma[\tilde{\xi}]_\Gamma\,\mathrm{d}S$, $\int_\Gamma\langle\nabla\xi\rangle\cdot\boldsymbol{n}[\tilde{\xi}]\,\mathrm{d}S$, $J_{h,m}(\xi,\tilde{\xi})$, etc. are polynomials in t of degree $\le 2q$. Therefore, we can express the integrals $\int_{I_m}\cdots\mathrm{d}t$ with the aid of the integration formula (50). We also use the relations $\tilde{\xi}(t^{m,i})=\xi(t^{m,i})\vartheta_i^{-1}$, $\nabla\tilde{\xi}(t^{m,i})=\nabla\xi(t^{m,i})\vartheta_i^{-1}$, $[\tilde{\xi}(t^{m,i})]=[\xi(t^{m,i})\vartheta_i^{-1}]$. Then , by (10) we get

$$\begin{aligned}&\int_{I_m}A_{h,m}(\xi,\tilde{\xi})\,\mathrm{d}t \qquad (57)\\ &=\varepsilon\,\tau_m\sum_{i=1}^{q+1}w_i\Big(\sum_{K\in\mathcal{T}_{h,m}}\int_K\nabla\xi(t^{m,i})\cdot\nabla\tilde{\xi}(t^{m,i})\mathrm{d}x+J_{h,m}(\xi(t^{m,i}),\tilde{\xi}(t^{m,i}))\,\mathrm{d}x\\ &\quad-\ \varepsilon\,\tau_m\sum_{\Gamma\in\mathcal{F}^I_{h,m}}\int_\Gamma\Big(\langle\nabla\xi(t^{m,i})\rangle\cdot\boldsymbol{n}_\Gamma[\tilde{\xi}(t^{m,i})]+\theta\langle\nabla\tilde{\xi}(t^{m,i})\rangle\cdot\boldsymbol{n}_\Gamma[\xi(t^{m,i})]\Big)\mathrm{d}S\\ &\quad-\ \varepsilon\,\tau_m\sum_{\Gamma\in\mathcal{F}^B_{h,m}}\int_\Gamma\Big(\nabla\xi(t^{m,i})\cdot\boldsymbol{n}_\Gamma\tilde{\xi}(t^{m,i})\ +\theta\nabla\tilde{\xi}(t^{m,i})\cdot\boldsymbol{n}_\Gamma\xi(t^{m,i})\Big)\mathrm{d}S\Big)\\ &=\tau_m\sum_{i=1}^{q+1}\vartheta_i^{-1}w_i\left(a_{h,m}(\xi(t^{m,i}),\xi(t^{m,i}))+\varepsilon J_{h,m}(\xi(t^{m,i}),\xi(t^{m,i}))\right).\end{aligned}$$

In virtue of the results from [31], under assumptions (32), we have

$$a_{h,m}(\xi(t^{m,i}),\xi(t^{m,i}))+\varepsilon J_{h,m}(\xi(t^{m,i}),\xi(t^{m,i}))\ge\frac{\varepsilon}{2}\|\xi(t^{m,i})\|^2_{DG,m},\quad i=1,\dots,q+1. \qquad (58)$$

If we use (57), (58), inequality $\vartheta_i^{-1}\ge 1$ and take into account that $\|\xi\|^2_{DG,m}$ is a polynomial in t of degree $\le 2q$, we find that

$$\begin{aligned}\int_{I_m}A_{h,m}(\xi,\tilde{\xi})\,\mathrm{d}t\ &\ge\ \frac{\varepsilon}{2}\tau_m\sum_{i=1}^{q+1}\vartheta_i^{-1}w_i\,\|\xi(t^{m,i})\|^2_{DG,m}\\ &\ge\ \frac{\varepsilon}{2}\tau_m\sum_{i=1}^{q+1}w_i\,\|\xi(t^{m,i})\|^2_{DG,m}=\frac{\varepsilon}{2}\int_{I_m}\|\xi\|^2_{DG,m}\,\mathrm{d}t,\end{aligned}$$

what we wanted to prove. □ □

(c) By the Cauchy inequality,

$$\left|(\xi^-_{m-1},\tilde{\xi}^+_{m-1})\right|\le\left\|\xi^-_{m-1}\right\|\left\|\tilde{\xi}^+_{m-1}\right\|. \qquad (59)$$

Lemma 3. *There exists a constant c_1 independent of h_K, τ_m, ξ such that*

$$\left\|\tilde{\xi}^+_{m-1}\right\|^2 \le \frac{c_1}{\tau_m}\,\|\xi\|^2_m \tag{60}$$

Proof. The function $\tilde{\xi}$ is the Lagrange interpolant to $\tau_m\,\xi(t)/(t-t_{m-1})$ at points $t^{m,i} = t_{m-1} + \tau_m\,\vartheta_i$, $i = 1,\dots,q+1$. This means that

$$\tilde{\xi}(t) \;=\; \tau_m \sum_{i=1}^{q+1} \frac{\xi(t^{m,i})}{t^{m,i}-t_{m-1}} \prod_{\substack{j=1\\ j\neq i}}^{q+1} \frac{t-t^{m,j}}{t^{m,i}-t^{m,j}} = \tau_m \sum_{i=1}^{q+1} \frac{\xi(t^{m,i})}{\tau_m\,\vartheta_i} \prod_{\substack{j=1\\ j\neq i}}^{q+1} \frac{t-t_{m-1}-\tau_m\,\vartheta_i}{\tau_m(\vartheta_i-\vartheta_j)}.$$

Setting $t = t_{m-1}$, we get

$$\tilde{\xi}^+_{m-1} = \sum_{i=1}^{q+1} \xi(t^{m,i})\,\vartheta_i^{-1} \prod_{\substack{j=1\\ j\neq i}}^{q+1} \frac{-\vartheta_j}{\vartheta_i-\vartheta_j}$$

and, thus, since $\vartheta_i^{-1} \le \vartheta_1^{-1}$,

$$\begin{aligned}\left\|\tilde{\xi}^+_{m-1}\right\|^2 &\le C(q) \sum_{i=1}^{q+1} \vartheta_i^{-1}\,\vartheta_1^{-1} \left\|\xi(t^{m,i})\right\|^2 \left(\prod_{\substack{j=1\\ j\neq i}}^{q+1} \frac{\vartheta_j}{\vartheta_i-\vartheta_j}\right)^2 \\ &\le \tilde{C}(q) \sum_{i=1}^{q+1} \vartheta_i^{-1} \left\|\xi(t^{m,i})\right\|^2 \left(\prod_{\substack{j=1\\ j\neq i}}^{q+1} \frac{\vartheta_j}{\vartheta_i-\vartheta_j}\right)^2.\end{aligned} \tag{61}$$

The Radau weights are defined as

$$w_i = \int_0^1 \prod_{\substack{j=1\\ j\neq i}}^{q+1} \frac{z-\vartheta_j}{\vartheta_i-\vartheta_j}\,\mathrm{d}z.$$

By [46], $w^* := \min_{i=1,\dots,q+1} w_i > 0$. Moreover, let us set

$$w^{**} := \max_{i=1,\dots,q+1} \left(\prod_{\substack{j=1\\ j\neq i}}^{q+1} \frac{\vartheta_j}{\vartheta_i-\vartheta_j}\right)^2.$$

Hence, since $w_i \ge w^*$, using (54), we get

$$\left\|\tilde{\xi}^+_{m-1}\right\|^2 \le \tilde{C}(q) \sum_{i=1}^{q+1} \vartheta_i^{-1} \|\xi(t^{m,i})\|^2 \frac{w^{**}\,w^*}{w^*} \le c_1 \sum_{i=1}^{q+1} \vartheta_i^{-1} \left\|\xi(t^{m,i})\right\|^2 w_i = \frac{c_1}{\tau_m}\,\|\xi\|^2_m,$$

with $c_1 = \tilde{C}(q) w^{**}/w^*$. □ □

(d) Integration by parts implies that

$$\begin{aligned}&\int_{I_m} \left(\eta', \tilde{\xi}\right) \mathrm{d}t + \left(\{\eta\}_{m-1}, \tilde{\xi}^+_{m-1}\right) \\ &\quad = -\int_{I_m} \left(\eta, \tilde{\xi}'\right) \mathrm{d}t + \left(\eta^-_m, \tilde{\xi}^-_m\right) - \left(\eta^+_{m-1}, \tilde{\xi}^+_{m-1}\right) + \left(\eta^+_{m-1}, \tilde{\xi}^+_{m-1}\right) - \left(\eta^-_{m-1}, \tilde{\xi}^+_{m-1}\right).\end{aligned} \tag{62}$$

Since $\tilde{\xi}' \in S_{h,\tau}^{p,q-1}$, in virtue of (21), c),

$$\int_{I_m} \big(\eta, \tilde{\xi}'\big)\,\mathrm{d}t = 0. \tag{63}$$

Further, $\tilde{\xi}_m^- \in S_{h,m}^p$ and thus,

$$\big(\eta_m^-, \tilde{\xi}_m^-\big) = 0. \tag{64}$$

It follows from (62) – (64) that

$$\int_{I_m} \big(\eta', \tilde{\xi}\big)\,\mathrm{d}t + \big(\{\eta\}_{m-1}, \xi_{m-1}^+\big) = -\big(\eta_{m-1}^-, \tilde{\xi}_{m-1}^+\big) \;\le\; \big\|\eta_{m-1}^-\big\|\,\big\|\tilde{\xi}_{m-1}^+\big\|. \tag{65}$$

(e) We use the following lemma:

Lemma 4. *If $k > 0$, then there exists a constant $C > 0$ such that*

$$\Big|\int_{I_m} A_{h,m}(\eta, \tilde{\xi})\,\mathrm{d}t\Big| \le \frac{\varepsilon}{k}\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2\,\mathrm{d}t + C\,\varepsilon\int_{I_m} \sigma_m^2(\eta)\,\mathrm{d}t. \tag{66}$$

Proof. Using (40) with $\varphi := \tilde{\xi}$, we get

$$\Big|\int_{I_m} A_{h,m}(\eta, \tilde{\xi})\,\mathrm{d}t\Big| \le \frac{\varepsilon}{k}\int_{I_m} \big\|\tilde{\xi}\big\|_{\mathrm{DG},m}^2\,\mathrm{d}t + C\varepsilon\int_{I_m} \sigma_m^2(\eta)\,\mathrm{d}t. \tag{67}$$

Now we shall estimate $\int_{I_m} \|\tilde{\xi}\|_{\mathrm{DG},m}^2\,\mathrm{d}t$. The function $\tilde{\xi}(t) = \sum_{j=0}^q \alpha_j\,t^j$, where $\alpha_j \in S_{h,m}^p$ is the Radau interpolation of the function $\tau_m\,\xi(t)/(t - t_{m-1})$. Hence,

$$\|\tilde{\xi}(t^{m,i})\|_{\mathrm{DG},m}^2 = \|\xi(t^{m,i})\|_{\mathrm{DG},m}^2\,\vartheta_i^{-2}, \quad i = 1, \dots, q+1,$$

and $\|\tilde{\xi}(t)\|_{\mathrm{DG},m}^2$ is a polynomial in t of degree $\le 2q$. Thus, we get

$$\begin{aligned}\int_{I_m} \big\|\tilde{\xi}(t)\big\|_{\mathrm{DG},m}^2\,\mathrm{d}t &= \tau_m\sum_{i=1}^{q+1} w_i\big\|\tilde{\xi}(t^{m,i})\big\|_{\mathrm{DG},m}^2 = \tau_m\sum_{i=1}^{q+1} w_i\,\vartheta_i^{-2}\big\|\xi(t^{m,i})\big\|_{\mathrm{DG},m}^2 \\ &\le \vartheta_1^{-2}\tau_m\sum_{i=1}^{q+1} w_i\big\|\xi(t^{m,i})\big\|_{\mathrm{DG},m}^2 = \vartheta_1^{-2}\int_{I_m}\|\xi\|_{\mathrm{DG},m}^2\,\mathrm{d}t.\end{aligned}$$

Hence,

$$\int_{I_m} \big\|\tilde{\xi}\big\|_{\mathrm{DG},m}^2\,\mathrm{d}t \le C\int_{I_m} \big\|\xi\big\|_{\mathrm{DG},m}^2\,\mathrm{d}t. \tag{68}$$

From (67) and (68) we get estimate (66), which we wanted to prove. □ □

(f) By (33) and (68),

$$\Big|\int_{I_m} b_{h,m}(u, \tilde{\xi}) - b_{h,m}(U, \tilde{\xi})\,\mathrm{d}t\Big| \le \frac{\varepsilon}{k}\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2\mathrm{d}t + \frac{C}{\varepsilon}\Big(\int_{I_m}\|\xi\|^2\mathrm{d}t + \int_{I_m}\tilde{\sigma}_m^2(\eta)\mathrm{d}t\Big). \tag{69}$$

Now we prove the desired estimate.

Lemma 5. *There exist constants $C, C^* > 0$ such that*

$$\int_{I_m} \|\xi\|^2 \,\mathrm{d}t \le C\,\tau_m\Big(\big\|\xi_{m-1}^-\big\|^2 + \big\|\eta_{m-1}^-\big\|^2 + \int_{I_m} R_m(\eta)\,\mathrm{d}t\Big), \tag{70}$$

provided

$$0 < \tau_m \le C^*\varepsilon. \tag{71}$$

Proof. If we proceed similarly as in the proof of (68), using (52) and the inequalities $1 \le \vartheta_i^{-1} \le \vartheta_1^{-1}$, we get

$$\begin{aligned}\int_{I_m} \|\xi\|^2 \,\mathrm{d}t &= \tau_m \sum_{i=1}^{q+1} w_i \big\|\xi(t^{m,i})\big\|^2 \le \|\xi\|_m^2, \\ \|\xi\|_m^2 &\le \vartheta_1^{-1}\,\tau_m \sum_{i=1}^{q+1} w_i \big\|\xi(t^{m,i})\big\|^2 = \vartheta_1^{-1}\int_{I_m} \|\xi\|^2 \,\mathrm{d}t.\end{aligned} \tag{72}$$

Now, estimates (53), (55), (56), (59), (60), (65), (66) and (69) yield

$$\begin{aligned}&\frac{1}{2}\|\xi_m^-\|^2 + \frac{1}{2}\,\frac{1}{\tau_m}\|\xi\|_m^2 + \frac{\varepsilon}{2}\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2 \,\mathrm{d}t \\ &\quad\le \big\|\xi_{m-1}^-\big\|\,\|\xi\|_m \sqrt{\frac{c_1}{\tau_m}} + \big\|\eta_{m-1}^-\big\|\,\|\xi\|_m \sqrt{\frac{c_1}{\tau_m}} + \frac{2\varepsilon}{k}\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2 \,\mathrm{d}t \\ &\qquad + \frac{C}{\varepsilon}\int_{I_m} \|\xi\|^2 \,\mathrm{d}t + C\,\varepsilon \int_{I_m} \sigma_m^2(\eta)\,\mathrm{d}t + \frac{C}{\varepsilon}\int_{I_m} \tilde\sigma_m^2(\eta)\,\mathrm{d}t.\end{aligned}$$

This, (48), (72), Young's inequality and the choice $k := 8$ imply that

$$\begin{aligned}&\big\|\xi_m^-\big\|^2 + \frac{\varepsilon}{2}\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2 \,\mathrm{d}t + \Big(\frac{1}{2\tau_m} - \frac{\tilde C}{\varepsilon}\Big)\int_{I_m} \|\xi\|^2 \,\mathrm{d}t \\ &\quad\le C\Big(\big\|\xi_{m-1}^-\big\|^2 + \big\|\eta_{m-1}^-\big\|^2 + \int_{I_m} R_m(\eta)\,\mathrm{d}t\Big).\end{aligned} \tag{73}$$

Let us put $C^* = 1/(4\tilde C)$, where $\tilde C$ is the constant from (73), and assume that (71) holds. Then $\frac{1}{2\tau_m} - \frac{\tilde C}{\varepsilon} \ge \frac{1}{4\tau_m}$ and (73) implies (70). □ □

Summarizing estimates (47) with $k := 8$ and (70), we find that for $m = 1, \dots, M$,

$$\big\|\xi_m^-\big\|^2 + \frac{\varepsilon}{2}\int_{I_m} \|\xi\|_{\mathrm{DG},m}^2 \,\mathrm{d}t \le \Big(1 + \frac{c}{\varepsilon}\,\tau_m\Big)\,\big\|\xi_{m-1}^-\big\|^2 + C\big\|\eta_{m-1}^-\big\|^2 + C\int_{I_m} R_m(\eta)\,\mathrm{d}t, \tag{74}$$

with constants c, $C > 0$.

Finally, we come to the *abstract error estimate*.

Theorem 6. *Let (71) hold. Then there exist constants $C, c > 0$ such that the error $e = U - u$ satisfies the estimate for all $m = 1, \dots, M$:*

$$\begin{aligned}&\|e_m^-\|^2 + \frac{\varepsilon}{2}\sum_{j=1}^{m}\int_{I_j} \|e\|_{DG,j}^2 \,\mathrm{d}t \\ &\quad\le C\exp(ct_m/\varepsilon)\Big(\sum_{j=1}^{m}\|\eta_j^-\|^2 + \sum_{j=1}^{m}\int_{I_j} R_j(\eta)\,\mathrm{d}t\Big) + 2\Big(\|\eta_m^-\|^2 + \varepsilon\sum_{j=1}^{m}\int_{I_j} \|\eta\|_{DG,j}^2 \,\mathrm{d}t\Big).\end{aligned} \tag{75}$$

Proof. The application of the discrete Gronwall's lemma to (74) gives the estimate

$$\begin{aligned} &\left\|\xi_m^-\right\|^2 + \frac{\varepsilon}{2}\sum_{j=1}^{m}\int_{I_j}\|\xi\|_{\mathrm{DG},j}^2\,\mathrm{d}t \\ &\quad\le C\exp(ct_m/\varepsilon)\Big(\left\|\xi_0^-\right\|^2 + \sum_{j=1}^{m}\left\|\eta_j^-\right\|^2 + \sum_{j=1}^{m}\int_{I_j}R_j(\eta)\,\mathrm{d}t\Big), \end{aligned} \tag{76}$$

for $m = 1, \dots, M$. In view of the definition of U_0^-, we have $\xi_0^- = 0$. Now, if we use the relation $e = \xi + \eta$ and the inequalities

$$\begin{aligned} \|e\|^2 &\le 2(\|\xi\|^2 + \|\eta\|^2), \\ \|e\|_{DG,j}^2 &\le 2(\|\xi\|_{DG,j}^2 + \|\eta\|_{DG,j}^2), \end{aligned} \tag{77}$$

from (76) we immediately get (75). □ □

5 Interpolation error bounds and error estimation in terms of h and τ

This section will be devoted to obtaining error estimates in dependence on the mesh sizes τ_m and h_m. They will be obtained on the basis of estimate (76), the relations

$$\begin{aligned} &e = U - u = \xi + \eta, \qquad \pi\,u\big|_{I_m} = \pi\,(\Pi_m u)\,\big|_{I_m}, \\ &\eta\big|_{I_m} = (\pi u - u)\big|_{I_m} = \eta^{(1)} + \eta^{(2)}, \text{ with } \eta^{(1)} = (\Pi_m u - u)\big|_{I_m},\ \eta^{(2)} = (\pi(\Pi_m u) - \Pi_m u)\,\big|_{I_m} \end{aligned} \tag{78}$$

and estimates of individual terms on the right-hand side of (76) containing η, which will be proven in the sequel. To this end, we assume that the exact solution satisfies the regularity condition

$$u \in H^{q+1}\big(0,T;H^1(\Omega)\big) \cap C([0,T];H^{p+1}(\Omega)) \tag{79}$$

and that the meshes satisfy conditions (26), (27), (31) and (71).

Obviously, $C([0,T];H^{p+1}(\Omega)) \subset L^2(0,T;H^{p+1}(\Omega))$. Moreover, let

$$\tau_m \ge Ch_m^2, \quad m = 1, \dots, M. \tag{80}$$

Let us note that this assumption is not necessary, if the meshes are not time-dependent, i.e. all meshes $\mathcal{T}_{h,m},\ m = 1, \dots, M$, are identical.

If $r \ge 1$ is integer and $\mu = \min(r,p)$, then for $m = 1, \dots, M$ and any $v \in H^{r+1}(\Omega)$ we have the standard estimates

$$\begin{aligned} \|\Pi_m v - v\|_{L^2(K)} &\le C\,h_K^{\mu+1}|v|_{H^{r+1}(K)}, \\ |\Pi_m v - v|_{H^1(K)} &\le C\,h_K^{\mu}|v|_{H^{r+1}(K)}, \\ |\Pi_m v - v|_{H^2(K)} &\le C\,h_K^{\mu-1}|v|_{H^{r+1}(K)}, \end{aligned} \tag{81}$$

for $K \in \mathcal{T}_{h,m},\ h \in (0,h_0)$ and

$$\begin{aligned} &\text{a)} \quad \|\Pi_m v\|_{L^2(K)} \le \|v\|_{L^2(K)} \text{ for } v \in L^2(K),\ K \in \mathcal{T}_{h,m},\ h \in (0,h_0), \\ &\text{b)} \quad |\Pi_m v|_{H^1(K)} \le C|v|_{H^1(K)} \text{ for } v \in H^1(K),\ K \in \mathcal{T}_{h,m},\ h \in (0,h_0). \end{aligned} \tag{82}$$

It is possible to find that

$$D^{q+1}(\Pi_m u) = \Pi_m(D^{q+1}u). \tag{83}$$

Actually, by (17), $\Pi_m u(\cdot,t) \in S_{h,m}^p$ and for all $t \in I_m$,

$$\int_\Omega \big(\Pi_m u(x,t) - u(x,t)\big)\,\varphi(x)\,\mathrm{d}x = 0, \quad \forall\,\varphi \in S_{h,m}^p. \tag{84}$$

The differentiation with respect to t yields

$$\int_\Omega \big(D^{q+1}(\Pi_m u(x,t)) - D^{q+1}u(x,t)\big)\,\varphi(x)\,\mathrm{d}x = 0, \quad \forall\,\varphi \in S_{h,m}^p. \tag{85}$$

Moreover, obviously $D^{q+1}(\Pi_m u(t)) \in S_{h,m}^p$ and thus, (83) holds.

Similarly we can prove that

$$D^{q+1}(\nabla\Pi_m u) = \nabla\Pi_m(D^{q+1}u). \tag{86}$$

5.1 Time interpolation

Lemma 7. *Let $\varphi \in C((t_{m-1},t_m], S_{h,m}^p), m = 1\ldots,M$. Then for each $x \in K$, $K \in \mathcal{T}_{h,m}$, $t \in I_m$, $m = 1,\ldots,M$ we have*

$$\pi\,\varphi(x,t) = \tilde{P}_m\varphi(x,t), \tag{87}$$

where $\tilde{P}_m$ is defined in the following way: For $\omega \in C\big((t_{m-1},t_m]\big)$,

$$\begin{aligned} &a) \quad \tilde{P}_m\omega \in \mathcal{P}^q(I_m), \\ &b) \quad \int_{I_m} \big(\tilde{P}_m\omega(t) - \omega(t)\big)t^j\,\mathrm{d}t = 0, \quad \forall\, j = 0,\ldots,q-1, \\ &c) \quad \tilde{P}_m\omega(t_m-) = \omega(t_m-). \end{aligned} \tag{88}$$

Proof. Let $m \in \{1,\ldots,M\}$. From the definition of the operators π and $\tilde{P}_m$ it follows that for each $K \in \mathcal{T}_{h,m}$ the functions $\pi\varphi$ and $\tilde{P}_m\varphi$ are on $K \times I_m$ polynomials of degree $\le q$ in $t \in I_m$ and of degree $\le p$ in $x \in K$. Moreover, $\pi\varphi(x,t_m-) = \varphi(x,t_m-) = \tilde{P}_m\varphi(x,t_m-)$ for all $x \in K$. Obviously, condition (21), c) is equivalent to

$$\begin{aligned} &\int_{I_m}\left(\int_K (\pi\varphi(x,t) - \varphi(x,t))\,\sigma(x)\,\mathrm{d}x\right) t^j\,\mathrm{d}t = 0, \\ &\qquad \forall\, j = 0,\ldots,q-1, \quad \forall\,\sigma \in P^p(K), \quad \forall\, K \in \mathcal{T}_{h,m}. \end{aligned} \tag{89}$$

Further, by (88), for any $K \in \mathcal{T}_{h,m}$,

$$\int_{I_m}\Big(\tilde{P}_m\varphi(x,t) - \varphi(x,t)\Big)\,t^j\,\mathrm{d}t = 0, \quad \forall\, j = 0,\ldots,q-1, \quad \forall\, x \in K. \tag{90}$$

Let $\sigma \in P^p(K)$. Then (90) and Fubini's theorem imply that

$$\begin{aligned} 0 &= \int_K\left(\int_{I_m} (\tilde{P}_m\varphi(x,t) - \varphi(x,t))\,t^j\,\mathrm{d}t\right)\sigma(x)\,\mathrm{d}x \\ &= \int_{I_m}\left(\int_K (\tilde{P}_m\varphi(x,t) - \varphi(x,t))\sigma(x)\,\mathrm{d}x\right)t^j\,\mathrm{d}t, \\ &\qquad \forall\, j = 0,\ldots,q-1, \quad \forall\,\sigma \in P^p(K), \quad \forall\, K \in \mathcal{T}_{h,m}. \end{aligned} \tag{91}$$

Comparing (91) with (89) and taking into account the fact that the operator π is uniquely determined by conditions (21), we immediately get (87).

□

Lemma 8. *If $\omega \in H^{q+1}(I_m)$, then*

$$\left\|\tilde{P}_m\omega - \omega\right\|^2_{L^2(I_m)} \le C\,\tau_m^{2q+2}\left\|D^{q+1}\omega\right\|^2_{L^2(I_m)}, \tag{92}$$

where $C > 0$ is a constant independent of ω, m and τ_m.

Proof. We proceed in several steps.

1) We transform the reference interval $[0,1]$ onto the interval $[t_{m-1}, t_m]$ by the mapping

$$t = t_m - \vartheta\tau_m, \quad \vartheta \in [0,1]. \tag{93}$$

If $\omega \in H^{q+1}(I_m)$ and $s(\vartheta) = \omega(t_m - \vartheta\tau_m)$, then $s \in H^{q+1}(0,1)$ and

$$(\tilde{P}_m\omega)(t_m - \vartheta\tau_m) = (Ps)(\vartheta), \tag{94}$$

where the operator P is defined by

$$\begin{aligned} &\text{a)} \quad Ps \in \mathcal{P}^q(0,1), \\ &\text{b)} \quad \int_0^1 (Ps(\vartheta) - s(\vartheta))\,\vartheta^j\,\mathrm{d}\vartheta = 0 \quad \forall\, j = 0,\dots,q-1, \\ &\text{c)} \quad Ps(0+) = s(0+). \end{aligned} \tag{95}$$

Moreover, if we set

$$Z_m(t) = \tilde{P}_m\omega(t) - \omega(t),\ t \in (t_{m-1}, t_m), \quad z(\vartheta) = Ps(\vartheta) - s(\vartheta),\ \vartheta \in (0,1), \tag{96}$$

we have

$$z(\vartheta) = Z_m(t_m - \vartheta\tau_m), \quad D^{q+1}z(\vartheta) = (-1)^{q+1}\,\tau_m^{q+1}\,D^{q+1}Z_m(t_m - \vartheta\tau_m), \quad \vartheta \in (0,1). \tag{97}$$

By the substitution theorem,

$$\begin{aligned} \|z\|^2_{L^2(0,1)} &= \frac{1}{\tau_m}\|Z_m\|^2_{L^2(I_m)}, \\ \|D^{q+1}z\|^2_{L^2(0,1)} &= \tau_m^{2q+1}\|D^{q+1}Z_m\|^2_{L^2(I_m)}. \end{aligned} \tag{98}$$

2) Since conditions (95), a)–c) determine the values of the operator P uniquely, it is clear that

$$Pr = r \quad \text{for} \quad r \in \mathcal{P}^q(0,1). \tag{99}$$

Now we prove that the operator P is a continuous mapping of the space $H^{q+1}(0,1)$ into $L^2(0,1)$. Let $u_n \in H^{q+1}(0,1)$, $n = 1,2,\dots$ and $u_n \to 0$ in $H^{q+1}(0,1)$ as $n \to \infty$. The continuous imbedding $H^{q+1}(0,1) \hookrightarrow C([0,1])$ implies that

$$u_n \rightrightarrows 0 \quad \text{in } [0,1] \tag{100}$$

and hence, by (95), c),

$$Pu_n(0) \to 0. \tag{101}$$

For $j = 0,\dots,q-1$ we have

$$\int_0^1 (Pu_n - u_n)\,(\vartheta)\,\vartheta^j\,\mathrm{d}\vartheta = 0.$$

This and (100) imply that

$$\int_0^1 Pu_n(\vartheta)\,\vartheta^j\,\mathrm{d}\vartheta = \int_0^1 u_n(\vartheta)\,\vartheta^j\,\mathrm{d}\vartheta \to 0, \quad j=0,\dots,q-1. \tag{102}$$

Since $Pu_n \in \mathcal{P}^q(0,1)$, we can write

$$Pu_n(\vartheta) = \sum_{i=1}^{q} c_i^{(n)}\,\vartheta^i + (Pu_n)(0), \quad \vartheta\in[0,1]. \tag{103}$$

Integration yields

$$\int_0^1 Pu_n(\vartheta)\,\vartheta^j\,\mathrm{d}\vartheta = \int_0^1 \sum_{i=1}^{q} c_i^{(n)}\,\vartheta^{i+j}\,\mathrm{d}\vartheta + (Pu_n)(0)\int_0^1 \vartheta^j\,\mathrm{d}\vartheta \tag{104}$$

$$= \sum_{i=1}^{q} c_i^{(n)}\,\frac{1}{i+j+1} + (Pu_n)(0)\,\frac{1}{j+1}, \quad j=0,\dots,q-1.$$

Using (101), (102), (104) and the fact that the matrix $\left(\frac{1}{i+j}\right)_{i,j=1}^{q}$ is nonsingular (cf. [35]), we find that

$$c_i^{(n)} \to 0 \quad \text{for} \quad i=1,\dots,q \quad \text{as} \quad n\to\infty.$$

Thus, $Pu_n \rightrightarrows 0$ in $[0,1]$ and $Pu_n \to 0$ in $L^2(0,1)$.

3) The above results allow us to apply Theorem 3.1.4 from [11] and get the estimate

$$\|z\|_{L^2(0,1)} \le C\|D^{q+1}z\|_{L^2(0,1)} \tag{105}$$

with a constant $C>0$ independent of $z\in H^{q+1}(0,1)$. This and (98) imply that

$$\|Z_m\|_{L^2(I_m)} \le C\,\tau_m^{2q+2}\|D^{q+1}Z_m\|_{L^2(I_m)}. \tag{106}$$

Taking into account that $D^{q+1}\,\tilde{P}_m\omega = 0$, we immediately get (92). □ □

In Appendix we give a direct proof of estimate (92) without the use of Theorem 3.1.4 from [11].

Lemmas 7 and 8 imply that for $\varphi \in H^{q+1}(I_m, S_{h,m}^p)$ we have

$$\|\pi\varphi(x,\cdot)-\varphi(x,\cdot)\|_{L^2(I_m)}^2 \le C\,\tau_m^{2q+2}\big\|D^{q+1}\varphi(x,\cdot)\big\|_{L^2(I_m)}^2,\ x\in K,\ K\in\mathcal{T}_{h,m},\ m=1,\dots,M. \tag{107}$$

5.2 Estimates of terms with η

Our further goal is to estimate the expressions

$$\|\eta_m^-\|^2,\ \int_{I_m}\|\eta\|_{L^2(K)}^2\,\mathrm{d}t,\ \int_{I_m}|\eta|_{H^1(K)}^2\,\mathrm{d}t,\ \int_{I_m}|\eta|_{H^2(K)}^2\,\mathrm{d}t,\ J_{h,m}(\eta,\eta).$$

By (78),

$$\|\eta\|_{L^2(K)}^2 \le 2\|\eta^{(1)}\|_{L^2(K)}^2 + 2\|\eta^{(2)}\|_{L^2(K)}^2, \tag{108}$$
$$|\eta|_{H^s(K)}^2 \le 2|\eta^{(1)}|_{H^s(K)}^2 + 2|\eta^{(2)}|_{H^s(K)}^2,\ s=1,2.$$

Lemma 9. *The following estimates hold for $K \in \mathcal{T}_{h,m},\ m = 1, \dots, M$:*

$$\begin{aligned}
\|\eta_m^-\|^2 &\le Ch^{p+1}|u(t_j)|_{H^{p+1}(\Omega)}, && (109)\\
\int_{I_m} \|\eta^{(1)}\|^2_{L^2(K)}\mathrm{d}t &\le C\, h_K^{2(p+1)}|u|^2_{L^2(I_m,H^{p+1}(K))}, && (110)\\
\int_{I_m} |\eta^{(1)}|^2_{H^1(K)}\mathrm{d}t &\le C\, h_K^{2p}|u|^2_{L^2(I_m,H^{p+1}(K))}, && (111)\\
h_K^2\int_{I_m} |\eta^{(1)}|^2_{H^2(K)}\mathrm{d}t &\le C\, h_K^{2p}|u|^2_{L^2(I_m,H^{p+1}(K))}. && (112)
\end{aligned}$$

Proof. It is enough to use (81). □ □

The derivation of estimates of terms with $\eta^{(2)}$ is more complicated.

Lemma 10. *For $K \in \mathcal{T}_{h,m},\ m = 1, \dots, M$, we have*

$$\begin{aligned}
\int_{I_m} \|\eta^{(2)}\|^2_{L^2(K)}\mathrm{d}t &\le C\, \tau_m^{2(q+1)}|u|^2_{H^{q+1}(I_m,L^2(K))}, && (113)\\
\int_{I_m} |\eta^{(2)}|^2_{H^1(K)}\mathrm{d}t &\le C\, \tau_m^{2(q+1)}|u|^2_{H^{q+1}(I_m,H^1(K))}, && (114)\\
h_K^2\int_{I_m} |\eta^{(2)}|^2_{H^2(K)}\mathrm{d}t &\le C\, \tau_m^{2(q+1)}|u|^2_{H^{q+1}(I_m,H^1(K))}.
\end{aligned}$$

Proof. a) The use of Fubini's theorem and relations (87), (83), (107), (82), a) and (92) yield the relations

$$\begin{aligned}
\int_{I_m} \|\eta^{(2)}\|^2_{L^2(K)} &= \int_{I_m}\Big(\int_K |\eta^{(2)}|^2\mathrm{d}x\Big)\mathrm{d}t\\
&= \int_K\Big(\int_{I_m} |\eta^{(2)}|^2\mathrm{d}t\Big)\mathrm{d}x = \int_K \|\tilde{P}_m(\Pi_m u) - \Pi_m u\|^2_{L^2(I_m)}\mathrm{d}x\\
&\le C\,\tau_m^{2q+2}\int_K \|D^{q+1}(\Pi_m u)\|^2_{L^2(I_m)}\mathrm{d}x\\
&= C\,\tau_m^{2q+2}\int_{I_m}\Big(\int_K |\,D^{q+1}(\Pi_m u)|^2\mathrm{d}x\Big)\mathrm{d}t\\
&= C\,\tau_m^{2q+2}\int_{I_m}\Big(\int_K |\Pi_m(D^{q+1}u)\|^2\mathrm{d}x\Big)\mathrm{d}t\\
&\le C\,\tau_m^{2q+2}\int_{I_m}\Big(\int_K |D^{q+1}u|^2\mathrm{d}x\Big)\mathrm{d}t\\
&= C\,\tau_m^{2q+2}|u|^2_{H^{q+1}(I_m,L^2(K))}.
\end{aligned}$$

b) Further, due to Fubini's theorem, (87), (107), (86) and (82), b), we find that

$$\begin{aligned}
\int_{I_m} |\eta^{(2)}|_{H^1(K)}\mathrm{d}t &= \int_{I_m} \Big(\int_K \Big| \nabla \Big(\Pi_m u - \tilde{P}_m(\Pi_m u) \Big) \Big|^2 \,\mathrm{d}x \Big) \mathrm{d}t \\
&= \int_K \Bigg(\int_{I_m} \sum_{j=1}^{d} \Bigg(\frac{\partial}{\partial x_j}(\Pi_m u) - \tilde{P}_m \left(\frac{\partial}{\partial x_j}(\Pi_m u) \right) \Bigg)^2 \mathrm{d}t \Bigg) \mathrm{d}x \\
&\le C\, \tau_m^{2q+2} \int_K |\nabla(\Pi_m u)|^2_{H^{q+1}(I_m)} \mathrm{d}x \\
&= C\, \tau_m^{2q+2} \int_K \Big(\int_{I_m} \left| D^{q+1} \nabla(\Pi_m u) \right|^2 \mathrm{d}t \Big) \mathrm{d}x \\
&= C\, \tau_m^{2q+2} \int_{I_m} \Big(\int_K \left| \nabla(\Pi_m D^{q+1} u) \right|^2 \mathrm{d}x \Big) \mathrm{d}t \\
&= C\, \tau_m^{2q+2} \int_{I_m} \left| \Pi_m \left(D^{q+1} u \right) \right|^2_{H^1(K)} \mathrm{d}t \\
&\le C\, \tau_m^{2q+2} \int_{I_m} \left| D^{q+1} u \right|^2_{H^1(K)} \mathrm{d}t = C\, \tau_m^{2q+2} |u|^2_{H^{q+1}(I_m, H^1(K))}.
\end{aligned}$$

c) Using a similar process as in b) and (30), we find that

$$\begin{aligned}
\int_{I_m} |\eta^{(2)}|_{H^2(K)}\mathrm{d}t &\le C\, \tau_m^{2q+2} \int_{I_m} \left| \Pi_m \left(D^{q+1} u \right) \right|^2_{H^2(K)} \mathrm{d}t \\
&\le C\, \tau_m^{2q+2} h_K^{-2} \int_{I_m} \left| D^{q+1} u \right|^2_{H^1(K)} \mathrm{d}t \\
&= C\, \tau_m^{2q+2} h_K^{-2} |u|_{H^{q+1}(I_m, H^1(K))}.
\end{aligned}$$

This yields (115). □
□

Finally, we shall be concerned with the estimation of $\int_{I_m} J_{h,m}(\eta, \eta)\,\mathrm{d}t$. It holds

$$J_{h,m}(\eta,\eta) \le C\big(J_{h,m}(\Pi_m u - u, \Pi_m u - u) + J_{h,m}(\pi(\Pi_m u) - \Pi_m u, \pi(\Pi_m u) - \Pi_m u) \big). \quad (115)$$

Using the multiplicative trace inequality (29) and (81), in the same way as in [21] we get

$$\int_{I_m} J_{h,m}\big(\Pi_m u - u, \Pi_m u - u\big)\,\mathrm{d}t \le C\, h^{2p} \big| u \big|^2_{L^2(I_m, H^{p+1}(\Omega))}. \quad (116)$$

Further, we shall estimate the expression

$$\int_{I_m} J_{h,m}\big(\pi(\Pi_m u) - \Pi_m u,\, \pi(\Pi_m u) - \Pi_m u\big)\,\mathrm{d}t.$$

We proceed in two steps.

(I) Let $\Gamma \in \mathcal{F}_{h,m}^I$, i. e. $\Gamma \subset \Omega$. If we set $\varphi := \Pi_m u$ and use the relation $\big[\tilde{P}_m \varphi - \varphi\big] = \tilde{P}_m[\varphi] - [\varphi]$ and estimate (107), we find that

$$\begin{aligned}
\int_{I_m} \Big(\int_\Gamma [\pi(\Pi_m u) - \Pi_m u]^2 \,\mathrm{d}S \Big)\,\mathrm{d}t &= \int_\Gamma \left\| \tilde{P}_m[\varphi(x,\cdot)] - [\varphi(x,\cdot)] \right\|^2_{L^2(I_m)} \mathrm{d}S \quad (117) \\
&\le C\, \tau_m^{2q+2} \int_\Gamma \left\| D^{q+1}[\varphi(x,\cdot)] \right\|^2_{L^2(I_m)} \mathrm{d}S.
\end{aligned}$$

If we take into account that

$$D^{q+1}[\varphi(x,\cdot)] = [D^{q+1}\varphi(x,\cdot)],\ [D^{q+1}u] = 0, \tag{118}$$

and use Fubini's theorem, we obtain

$$\int_{I_m}\Big(\int_\Gamma [\pi(\Pi_m u) - \Pi_m u]^2\,\mathrm{d}S\Big)\,\mathrm{d}t = \int_\Gamma\Big(\int_{I_m}[\pi(\Pi_m u) - \Pi_m u]^2\,\mathrm{d}t\Big)\,\mathrm{d}S \tag{119}$$
$$\le C\,\tau_m^{2q+2}\int_\Gamma\Big(\int_{I_m}\big|D^{q+1}[\varphi(x,t)]\big|^2\,\mathrm{d}t\Big)\,\mathrm{d}S = C\,\tau_m^{2q+2}\int_{I_m}\Big(\int_\Gamma [D^{q+1}(\Pi_m u - u)]^2\,\mathrm{d}S\Big)\,\mathrm{d}t.$$

The application of the multiplicative trace inequality implies that

$$\sum_{\Gamma\in\mathcal{F}_{h,m}^I}\int_\Gamma \big[D^{q+1}(\Pi_m u - u)\big]^2\,\mathrm{d}S \tag{120}$$
$$\le\ C\sum_{K\in\mathcal{T}_{h,m}}\int_{\partial K}\big[D^{q+1}(\Pi_m u - u)\big]^2\,\mathrm{d}S = C\sum_{K\in\mathcal{T}_{h,m}}\big\|D^{q+1}(\Pi_m u - u)\big\|^2_{L^2(\partial K)}$$
$$\le\ C\sum_{K\in\mathcal{T}_{h,m}}\Big(\big\|D^{q+1}(\Pi_m u - u)\big\|_{L^2(K)}\big|D^{q+1}(\Pi_m u - u)\big|_{H^1(K)} + h_K^{-1}\big\|D^{q+1}(\Pi_m u - u)\big\|^2_{L^2(K)}\Big).$$

By (83),

$$D^{q+1}(\Pi_m u - u) = \Pi_m(D^{q+1}u) - D^{q+1}u. \tag{121}$$

In virtue of (79), $D^{q+1}u \in L^2(I_m, H^1(\Omega))$. This and the approximation properties (81) of Π_m imply that

$$\begin{aligned}\|\Pi_m(D^{q+1}u) - D^{q+1}u\|_{L^2(K)} &\le Ch_K|D^{q+1}u|_{H^1(K)},\\ |\Pi_m(D^{q+1}u) - D^{q+1}u|_{H^1(K)} &\le C|D^{q+1}u|_{H^1(K)}.\end{aligned} \tag{122}$$

Summarizing (28), (119), (120), (121) and (122), we get

$$\int_{I_m}\Big(\sum_{\Gamma\in\mathcal{F}_{h,m}^I} h(\Gamma)^{-1}\int_\Gamma\big[\pi(\Pi_m u) - \Pi_m u\big]^2\,\mathrm{d}S\Big)\,\mathrm{d}t \tag{123}$$
$$\le\ C\,\tau_m^{2q+2}\int_{I_m}\sum_{K\in\mathcal{T}_{h,m}}\big|D^{q+1}u\big|^2_{H^1(K)}\,\mathrm{d}t$$
$$=\ C\,\tau_m^{2q+2}\sum_{K\in\mathcal{T}_{h,m}}|u|^2_{H^{q+1}(I_m,H^1(K))}.$$

(II) In what follows, we shall assume that $\Gamma\in\mathcal{F}_{h,m}^B$, i. e. $\Gamma\subset\partial\Omega\cap\partial K$ for some $K\in\mathcal{T}_{h,m}$, and estimate the expression

$$\beta\ :=\ \int_{I_m}\Big(h(\Gamma)^{-1}\int_\Gamma|\pi(\Pi_m u) - \Pi_m u|^2\,\mathrm{d}S\Big)\,\mathrm{d}t. \tag{124}$$

Proceeding in a similar way as above, we find that

$$\beta\ \le\ C\,\tau_m^{2q+2}h(\Gamma)^{-1}\int_\Gamma\big\|D^{q+1}(\Pi_m u)\big\|^2_{L^2(I_m)}\,\mathrm{d}S \tag{125}$$
$$=\ C\,\tau_m^{2q+2}h(\Gamma)^{-1}\int_{I_m}\Big(\int_\Gamma\big|D^{q+1}(\Pi_m u)\big|^2\,\mathrm{d}S\Big)\,\mathrm{d}t$$
$$=\ C\,\tau_m^{2q+2}h(\Gamma)^{-1}\int_{I_m}\Big(\int_\Gamma\big|\Pi_m(D^{q+1}u)\big|^2\,\mathrm{d}S\Big)\,\mathrm{d}t.$$

If we apply the multiplicative trace inequality, we get the estimate of the part of $J_{h,m}\big(\pi(\Pi_m u)-\Pi_m u,\, \pi(\Pi_m u)-\Pi_m u\big)$ corresponding to $\Gamma \subset \partial\Omega\cap\partial K$ of order $O\big(\tau_m^{2q+2}h_K^{-2}\big)$. If $h_K \sim \tau_m$, we loose the order of accuracy $O(\tau_m^2)$ and the resulting error estimate is of order $O(\tau_m^q)$, which would be suboptimal. This drawback will be cured in the following way.

Let the Dirichlet data $u_D = u_D(x,t)$ have behaviour in t as a polynomial of degree $\le q$. In other words, we assume that

$$u_D(x,t) = \sum_{j=0}^{q} \psi_j(x)\, t^j, \tag{126}$$

where $\psi_j \in H^{p+1/2}(\partial\Omega)$ for $j = 0,\dots,q$. Hence, $D^{q+1}u|_{\partial\Omega} = D^{q+1}u_D = 0$. This and (125) imply that

$$(+) \;\le\; C\,\tau_m^{2q+2} \int_{I_m} \Big(h(\Gamma)^{-1} \int_\Gamma \big|\Pi_m(D^{q+1}u) - D^{q+1}u\big|^2 \,\mathrm{d}S\Big)\,\mathrm{d}t. \tag{127}$$

Again we use the multiplicative trace inequality and estimates (122) and get the estimate

$$\int_{I_m} \Big(\sum_{\Gamma\in\mathcal{F}_{h,m}^B} h(\Gamma)^{-1} \int_\Gamma \big|\pi(\Pi_m u) - \Pi_m u\big|^2 \,\mathrm{d}s\Big)\,\mathrm{d}t \le C\,\tau_m^{2q+2} \sum_{K\in\mathcal{T}_{h,m}} |u|^2_{H^{q+1}(I_m,H^1(K))}. \tag{128}$$

The above results can be summarized in the following way.

Lemma 11. *If we consider scheme (19) and the Dirichlet data is defined by (126), then*

$$|J_{h,m}(\eta,\eta)| \le C \sum_{K\in\mathcal{T}_{h,m}} \Big(h_K^{2p}|u|^2_{L^2(I_m,H^{p+1}(K))} + \tau_m^{2q+2}|u|^2_{H^{q+1}(I_m,H^1(K))}\Big). \tag{129}$$

5.3 Main result

In this section we shall conclude the analysis of the error estimate.

Theorem 12. *Let u be the exact solution of problem (1) – (3) satisfying the regularity condition (79). Let U be the approximate solution to problem (1) – (3) obtained by scheme (19) in the case that the Dirichlet data u_D is defined by (126), over spatial meshes $\mathcal{T}_{h,m}$ and time partition I_m, $m = 1,\dots,M$, satisfying conditions (26), (27), (71) and (80). Then there exist constants C, $c > 0$ independent of h, τ, m, ε, u such that*

$$\begin{aligned}
&\|e_m^-\|^2 + \frac{\varepsilon}{2}\sum_{j=1}^{m}\int_{I_j} \|e\|^2_{DG,j}\,\mathrm{d}t \qquad (130)\\
\le\;& C\exp(ct_m/\varepsilon)\Big(\sum_{j=1}^{m}\big(h_j^{2p}|u|^2_{L^2(I_j;H^{p+1}(\Omega))} + \tau_j^{2q+1}|u|^2_{H^{q+1}(I_j;H^1(\Omega))}\big)\Big(\varepsilon+\frac{1}{\varepsilon}\Big)\\
&+ h^{2p}|u|^2_{C([0,T];H^{p+1}(\Omega))}\Big) + C\,h^{2p+2}|u|^2_{C([0,T];H^{p+1}(\Omega))}\\
&+ C\,\varepsilon\sum_{j=1}^{m}\Big(h_j^{2p}|u|^2_{L^2(I_j;H^{p+1}(\Omega))} + \tau_j^{2q+2}|u|_{H^{q+1}(I_j;H^1(\Omega))}\Big),\\
&\quad m = 1,\dots,M,\ h\in(0,h_0),
\end{aligned}$$

or simply,

$$\begin{aligned} &\|e_m^-\| + \frac{\varepsilon}{2}\sum_{j=1}^{m}\int_{I_j}\|e\|_{DG,j}^2\,\mathrm{d}t \qquad (131)\\ \le\ & C\exp(ct_m/\varepsilon)\Big(\big(h^{2p}|u|^2_{L^2(0,T;H^{p+1}(\Omega))}+\tau^{2q+2}|u|_{H^{q+1}(0,T;H^1(\Omega))}\big)\Big(\varepsilon+\frac{1}{\varepsilon}\Big)\\ &+h^{2p}|u|^2_{C([0,T]+H^{p+1}(\Omega))}\Big),\quad m=1,\dots,M,\ h\in(0,h_0). \end{aligned}$$

Proof. In order to prove (130), we start from (75) and estimate the terms containing η. In virtue of (48), (39), (34),

$$\begin{aligned} &R_j(\eta)=\varepsilon\,\sigma_j^2(\eta)+\frac{1}{\varepsilon}\,\tilde\sigma_j^2(\eta) \qquad (132)\\ &=\varepsilon\Big(\sum_{K\in\mathcal{T}_{h,j}}\big(|\eta|^2_{H^1(K)}+h_K^2|\eta|^2_{H^2(K)}\big)+J_{h,j}(\eta,\eta)\Big)+\frac{1}{\varepsilon}\sum_{K\in\mathcal{T}_{h,j}}\Big(\|\eta\|^2_{L^2(K)}+h_K^2|\eta|^2_{H^1(K)}\Big). \end{aligned}$$

Now, (132) together with (108) and Lemmas 9 and 10 yield the estimate

$$\int_{I_j}R_j(\eta)\,\mathrm{d}t\le C\Big(\varepsilon+\frac{1}{\varepsilon}\Big)\sum_{K\in\mathcal{T}_{h,j}}\Big(h_K^{2p}|u|^2_{L^2(I_j,H^{p+1}(K))}+\ \tau_j^{2q+2}|u|^2_{H^{q+1}(I_j,H^1(K))}\Big). \quad (133)$$

This and the inequality $h_K\le h_j$ lead to

$$\int_{I_j}R_j(\eta)\,\mathrm{d}t\ \ \le\ \ C\Big(\varepsilon+\frac{1}{\varepsilon}\Big)\Big(h_j^{2p}|u|^2_{L^2(I_j,H^{p+1}(\Omega))}+\ \tau_j^{2q+2}|u|^2_{H^{q+1}(I_j,H^1(\Omega))}\Big). \quad (134)$$

Similarly, we get

$$\int_{I_j}\|\eta\|_{DG,j}\,\mathrm{d}t\le h_j^{2p}|u|^2_{L^2(I_j,H^{p+1}(\Omega))}+\ \tau_j^{2q+2}|u|^2_{H^{q+1}(I_j,H^1(\Omega))}. \quad (135)$$

Further, by (109) and (80),

$$\sum_{j=1}^{m}\|\eta_j^-\|^2\le C\sum_{j=1}^{M}\tau_j\,h_j^{2p}|u(t_j)|^2_{H^{p+1}(\Omega)}\le C\,Th^{2p}|u|^2_{C([0,T];H^{p+1}(\Omega))}. \quad (136)$$

Finally, using (76) and (134) – (136), we arrive at estimates (130) and (131), which we wanted to prove. □ □

Remark 2. *As we see, estimate (130) is not uniform with respect to $\varepsilon\to 0$. Just on the contrary, the constant in this estimate behaves as $C\exp(cT/\varepsilon)$, which blows up to ∞ as $\varepsilon\to 0$. This is a consequence of the application of Young's inequality necessary for the treatment of nonlinear terms, and Gronwall's lemma. The question, how to avoid this bad behaviour of the error estimate, remains open.*

5.4 The case of identical meshes on all time levels

If all meshes $\mathcal{T}_{h,m},\ m=1\dots,M$, are identical, which means that $\mathcal{T}_{h,m}=\mathcal{T}_h$ for all $m=1,\dots,M$, then all spaces $S^p_{h,m}$ and forms $a_{h,m},b_{h,m},\dots$ are also identical: $S^p_{h,m}=S^p_h$,

$a_{h,m} = a_h, b_{h,m} = b_h, \dots$ for all $m = 1, \dots, M$. This implies that $\{\xi\}_{m-1} \in S_h^p$ and by (24), (21), a), and (17), we have $(\eta_{m-1}^-, \{\xi\}_{m-1}) = 0$. Hence,

$$\int_{I_m} (\eta', \xi)\,\mathrm{d}t + \big(\{\eta\}_{m-1}, \xi_{m-1}^+\big) = 0. \tag{137}$$

Moreover, similarly it is possible to show that the expression $\sum_{j=1}^m \left\|\eta_j^-\right\|^2$ does not appear in estimate (76) and instead of (75) we get the estimate

$$\begin{aligned} &\|e_m^-\|^2 + \frac{\varepsilon}{2}\sum_{j=1}^m \int_{I_j} \|e\|_{DG,j}^2\,\mathrm{d}t \\ &\le C\exp(ct_m/\varepsilon)\Big(\sum_{j=1}^m \int_{I_j} R_j(\eta)\,\mathrm{d}t\Big) + 2\|\eta_m^-\|^2 + 2\varepsilon\sum_{j=1}^m \int_{I_j} \|\eta\|_{DG,j}^2\,\mathrm{d}t, \quad m = 1, \dots, M. \end{aligned} \tag{138}$$

Due to the fact that $\sum_{j=1}^m \left\|\eta_j^-\right\|^2$ does not appear in the abstract error estimate (75), assumption (80) can be omitted in the process of the derivation of the error estimates (130) and (131). This leads us to the following result.

Theorem 13. *Let u be the exact solution of problem (1) – (3) satisfying the regularity condition (79). Let U be the approximate solution to problem (1) – (3) obtained by scheme (19) in the case that the Dirichlet data u_D is defined by (126), over spatial meshes $\mathcal{T}_{h,m} = \mathcal{T}_h$ for all $m = 1, \dots, M$, and time partition I_m, $m = 1, \dots, M$, satisfying conditions (26), (27) and (71). Then there exist constants C, $c > 0$ independent of h, τ, m, ε, u such that error estimates (130) and (131) hold.*

5.5 $L^2(Q_T)$-error estimate

Finally, we shall be concerned with the $L^2(L^2)$-error estimate, i. e., the error estimate in the norm of the space $L^2(Q_T)$.

Theorem 14. *Let u be the exact solution of problem (1) – (3) satisfying the regularity condition (79). Let U be the approximate solution to problem (1) – (3) obtained by scheme (19) in the case that the Dirichlet data u_D is defined by (126), over spatial meshes $\mathcal{T}_{h,m}$ and time partition I_m, $m = 1, \dots, M$, satisfying conditions (26), (27), (71) and (80). Then there exist constants C, $c > 0$ independent of h, τ, m, ε, u such that*

$$\begin{aligned} \|e\|_{L^2(Q_T)}^2 \le{}& C\left(h^{2p+2} + e^{cT/\varepsilon}h^{2p}\right)|u|^2_{C([0,T],H^{p+1}(\Omega))} \\ &+ C\Big(\varepsilon + \frac{1}{\varepsilon}\Big)\left(1 + e^{cT/\varepsilon}\right)\left(h^{2p}|u|^2_{L^2(0,T;H^{p+1}(\Omega))} + \tau^{2q+2}|u|^2_{H^{q+1}(0,T;H^1(\Omega))}\right) \\ &+ C\big(h^{2p+2}|u|^2_{L^2(0,T;H^{p+1}(\Omega))} + \tau^{2q+2}|u|^2_{H^{q+1}(0,T;L^2(\Omega))}\big). \end{aligned} \tag{139}$$

Proof. It follows from (70) that

$$\int_0^T \|\xi\|^2\,\mathrm{d}t \le C\sum_{m=1}^M \tau_m\Big(\|\xi_{m-1}^-\|^2 + \|\eta_{m-1}^-\|^2 + \int_{I_m} R_m(\eta)\,\mathrm{d}t\Big). \tag{140}$$

This and (77) yield

$$\int_0^T \|e\|^2\,\mathrm{d}t \le C\sum_{m=1}^M \tau_m\Big(\|\xi_{m-1}^-\|^2 + \|\eta_{m-1}^-\|^2 + \int_{I_m} R_m(\eta)\,\mathrm{d}t\Big) + 2\int_0^T \|\eta\|^2\,\mathrm{d}t. \tag{141}$$

Now we use (76) with $m := m-1 < M$, $\xi_0 = 0$, $\eta_0^- = \Pi_1 u^0 - u^0$ and get

$$\|e\|^2_{L^2(Q_T)} = \int_0^T \|e\|^2 \,\mathrm{d}t \le C \sum_{m=1}^{M} \tau_m \Big(\|\eta^-_{m-1}\|^2 + \int_{I_m} R_m(\eta)\,\mathrm{d}t \tag{142}$$
$$+ \; C\, e^{cT/\varepsilon} \Big(\sum_{j=1}^{M} \|\eta_j^-\|^2 + \sum_{j=1}^{M} \int_{I_j} R_j(\eta)\,\mathrm{d}t \Big) \Big) + 2\|\eta\|^2_{L^2(Q_T)}.$$

Further, by (133), (136) and (109),

$$\sum_{j=1}^{M} \int_{I_j} R_j(\eta)\,\mathrm{d}t \le C\Big(\varepsilon + \frac{1}{\varepsilon}\Big) \Big(h^{2p} |u|^2_{L^2(0,T;H^{p+1}(\Omega))} + \; \tau^{2q+2} |u|^2_{H^{q+1}(0,T;H^1(\Omega))} \Big), \tag{143}$$

$$\sum_{j=1}^{M} \|\eta_j^-\|^2 \le C\, h^{2p} |u|^2_{C([0,T]+H^{p+1}(\Omega))}, \tag{144}$$

$$\|\eta^-_{m-1}\|^2 \le C\, h^{2p+2} |u|^2_{C([0,T]+H^{p+1}(\Omega))}, \tag{145}$$

$$\int_{I_m} R_m(\eta)\,\mathrm{d}t \le C\Big(\varepsilon + \frac{1}{\varepsilon}\Big) \Big(h_m^{2p} |u|^2_{L^2(I_m;H^{p+1}(\Omega))} + \; \tau_m^{2q+2} |u|^2_{H^{q+1}(I_m,H^1(\Omega))} \Big). \tag{146}$$

Moreover, (108), (110) and (113) imply that

$$\|\eta\|^2_{L^2(Q_T)} = \sum_{m=1}^{M} \int_{I_m} \|\eta\|^2 \,\mathrm{d}t \tag{147}$$
$$\le C \sum_{m=1}^{M} \Big(h^{2p+2} |u|^2_{L^2(I_m,H^{p+2}(\Omega))} + \tau^{2q+2} |u|^2_{H^{q+1}(I_m;L^2(\Omega))} \Big)$$
$$= C \Big(h^{2p+2} |u|^2_{L^2(0,T;H^{p+1}(\Omega))} + \tau^{2q+2} |u|^2_{H^{q+1}(0,T;L^2(\Omega))} \Big).$$

From estimates (142) – (147) we get

$$\|e\|^2_{L^2(Q_T)} \le C \sum_{m=1}^{M} \tau_m \Big(h^{2p+2} |u|^2_{C([0,T];H^{p+1})} \tag{148}$$
$$+ \; \Big(\varepsilon + \frac{1}{\varepsilon}\Big) \Big(h_m^{2p} |u|^2_{L^2(I_m,h^{p+1}(\Omega))} + \tau_m^{2q+2} |u|^2_{H^{q+1}(I_m,H^1(\Omega))} \Big) \Big)$$
$$+ \; e^{cT/\varepsilon} \Big(h^{2p} |u|^2_{C([0,T],H^{p+1}(\Omega))}$$
$$+ \Big(\varepsilon + \frac{1}{\varepsilon}\Big) \Big(h^{2p} |u|^2_{L^2(0,T;H^{p+1})} + \tau^{2q+2} |u|_{H^{q+1}(0,T;H^{1(\Omega)})} \Big) \Big)$$
$$+ \; C \Big(h^{2p+2} |u|^2_{L^2(0,T;H^{p+1}(\Omega))} + \tau^{2q+2} |u|^2_{H^{q+1}(0,T;L^2(\Omega))} \Big).$$

This and the relation $\sum_{m=1}^{M} \tau_m = T$ yield the final estimate (139). □ □

Remark 3. *Similarly as in Section 5.4, it is possible to formulate the $L^2(L^2)$-error estimate in the case of identical space meshes on all time levels without assumption (80).*

6 Appendix: Alternative proof of Lemma 8

Here we prove Lemma 8 without the use of Theorem 3.1.4 from [11].

Lemma 15. *Let $s \in C^\infty([0,1])$, $s(0) = 0$ and*

$$\int_0^1 \vartheta^i s(x)\,\mathrm{d}\vartheta = 0, \quad i = 0, \cdots, q-1. \tag{149}$$

Then

$$\|s\|_{L^2(0,1)} \le C\,\|D^{q+1}s\|_{L^2(0,1)}, \tag{150}$$

where C is a constant independent of the function s.

Proof. Let us develop the function with the aid of the Taylor formula with integral remainder:

$$s(\vartheta) = s(0) + \cdots + \frac{s^{(q)}(0)}{q!}\,\vartheta^q + \int_0^\vartheta \frac{(\vartheta-\tau)^q}{q!} s^{(q+1)}(\tau)\,\mathrm{d}\tau, \ \ \vartheta \in [0,1]. \tag{151}$$

In the space $L^2(0,1)$ we choose an orthonormal system of polynomials φ_i, $i = 0, 1, \ldots$, such that φ_i is a polynomial of degree i and $\varphi_i(0) \neq 0$. (At the end of Appendix we show how this system can be constructed.) Obviously,

$$\int_0^1 \vartheta^i s(\vartheta)\,\mathrm{d}\vartheta = 0, \ \ i = 0, \ldots, q-1 \quad \Longleftrightarrow \quad \int_0^1 \varphi_i(\vartheta) s(\vartheta)\,\mathrm{d}\vartheta = 0, \ \ i = 0, \ldots, q-1. \tag{152}$$

In virtue of the properties of the system φ_i, $i = 0, 1, \ldots$, the expansion (151) can be written in the form

$$s(\vartheta) = \sum_{i=0}^{q} c_i \varphi_i(\vartheta) + \int_0^\vartheta \frac{(\vartheta-\tau)^q}{q!} s^{(q+1)}(\tau)\,\mathrm{d}\tau, \ \ \vartheta \in [0,1], \tag{153}$$

where c_i are constants depending on the values $s(0), s'(0), \ldots, s^{(q)}(0)$. From assumption (151) and equivalence (152) for $j = 0, \ldots, q-1$ we get

$$0 = \int_0^1 \varphi_j(\vartheta) s(\vartheta)\,\mathrm{d}\vartheta = c_j + \int_0^1 \varphi_j(\vartheta) \int_0^\vartheta \frac{(\vartheta-\tau)^q}{q!} s^{(q+1)}(\tau)\,\mathrm{d}\tau\,\mathrm{d}\vartheta.$$

The use of Fubini's theorem yields

$$c_j = -\int_0^1 \varphi_j(\vartheta) \int_0^\vartheta \frac{(\vartheta-\tau)^q}{q!} s^{(q+1)}(\tau)\,\mathrm{d}\tau\,\mathrm{d}\vartheta = -\int_0^1 \psi_j(\tau) s^{(q+1)}(\tau)\,\mathrm{d}\tau, \tag{154}$$

where

$$\psi_j(\tau) = \frac{1}{q!}\int_\tau^1 \varphi_j(\vartheta)(\vartheta-\tau)^q\,\mathrm{d}\vartheta, \quad j = 0, \ldots, q-1.$$

Since $\varphi_q(0) \neq 0$, from the assumption that $s(0) = 0$ and expansion (153) we get

$$c_q = -\frac{1}{\varphi_q(0)} \sum_{i=0}^{q-1} c_i \varphi_i(0) = \int_0^1 \psi_q(\tau) s^{(q+1)}(\tau)\,\mathrm{d}\tau$$

with

$$\psi_q(\tau) = \frac{1}{\varphi_q(0)} \sum_{j=0}^{q-1} \varphi_j(0)\psi_j(\tau).$$

Substituting in expansion (153) for $c_i,\ i = 0, \dots, q$, we find that

$$s(\vartheta) = \int_0^1 k(\vartheta, \tau) s^{(q+1)}(\tau)\, \mathrm{d}\tau, \tag{155}$$

where

$$k(\vartheta,\tau) = \begin{cases} \frac{(\vartheta-\tau)^q}{q!} + \varphi_q(\vartheta)\psi_q(\tau) - \sum_{i=0}^{q-1} \varphi_i(\vartheta)\psi_i(\tau) & \text{for} \quad 1 \geq \vartheta > \tau \geq 0, \\ \varphi_q(\vartheta)\psi_q(\tau) - \sum_{i=0}^{q-1} \varphi_i(\vartheta)\psi_i(\tau) & \text{for} \quad 1 \geq \tau > \vartheta \geq 0. \end{cases}$$

The function $k(\vartheta, t)$ is continuous on the set $[0,1] \times [0,1]$ and from (155) we get (150). □ □

Lemma 16. *Let $s \in H^{q+1}(0,1)$, $s(0) = 0$ and*

$$\int_0^1 \vartheta^i s(\vartheta)\, \mathrm{d}\vartheta = 0, \quad i = 0, \dots, q-1. \tag{156}$$

Then

$$\|s\|_{L^2(0,1)} \leq C\, \|D^{q+1} s\|_{L^2(0,1)}, \tag{157}$$

where $C > 0$ is a constant independent of the function s.

Proof. The space $C^\infty([0,1])$ is dense in $H^{q+1}(0,1)$. Therefore, there exists a sequence $s_n \in C^\infty([0,1]$ such that

$$\lim_{n\to\infty} \|s_n - \omega\|_{H^{q+1}(0,1)} = 0.$$

This and Lemma 15 imply that

$$\begin{aligned} &\|s_n\|_{L^2(0,1)} \to \|s\|_{L^2(0,1)},\ \|D^{q+1} s_n\|_{L^2(0,1)} \to \|D^{q+1} s\|_{L^2(0,1)} \quad \text{as } n \to \infty, \\ &\|s_n\|_{L^2(0,1)} \leq C \|D^{q+1} s_n\|_{L^2(0,1)}, \quad n = 1, 2, \dots. \end{aligned}$$

Hence, (157) holds. □ □

Now, we can finish the proof of Lemma 8. Using the notation in (96), we have $z \in H^{q+1}(0,1)$, $z(0) = 0$ and

$$\int_0^1 z(\vartheta)\vartheta^j\, \mathrm{d}\vartheta = 0 \quad \text{for } j = 0, \dots, q-1.$$

By Lemma 16, the function z satisfies (157) (i.e. (105)) and, in virtue of (98), estimate (106) holds, which implies estimate (92). This finishes the alternative proof of Lemma 8.

Finally, we show, how to construct in the space $L^2(0,1)$ the system of orthonormal polynomials $\varphi_i,\ i = 0, 1, \dots$, such that φ_i is a polynomial of degree i satisfying $\varphi_i(0) \neq 0$. It is possible to put

$$\varphi_i(\vartheta) = \sqrt{2}\, P_i(2\vartheta - 1), \quad i = 0, 1, \dots,$$

where P_i is the Legendre polynomial of degree i, defined as

$$P_i(\vartheta) = \sqrt{i+\frac{1}{2}}\,\frac{1}{2^i\,i!}\frac{\mathrm{d}^i}{\mathrm{d}\vartheta^i}\,(\vartheta^2-1)^i, \quad i=0,1,\dots\,.$$

The system P_i, $i=0,1\dots$, is a complete orthonormal basis in the space $L^2(-1,1)$. It is possible to verify that φ_i, $i=0,1\dots$, form a complete orthonormal basis in $L^2(0,1)$ and

$$\varphi_i(\vartheta) = \sqrt{2i+1}\,\frac{1}{i!}\frac{\mathrm{d}^i}{\mathrm{d}\vartheta^i}\,(\vartheta^2-\vartheta)^i, \quad \vartheta\in[0,1].$$

Since

$$\frac{\mathrm{d}^i}{\mathrm{d}\vartheta^i}\,(\vartheta^2-\vartheta)^i|_{\vartheta=0} = (-1)^i\,i!,$$

for all $i=0,1,\dots$ we have

$$\varphi_i(0) = (-1)^i\,\sqrt{2i+1} \neq 0.$$

Conclusion

In this paper we have presented a detailed analysis of error estimates of the space-time discontinuous Galerkin discretization of an initial-boundary value problem for a nonstationary convection-diffusion equation with nonlinear convection and Dirichlet boundary condition. In the space discretization NIPG, IIPG and SIPG versions of the diffusion terms with polynomial approximation of degree $p \geq 1$ is used on each time level. In time the discontinuous approximations of degree $q \geq 1$, $q \neq p$ in general are used. On different time levels, different space meshes may be used. The derived estimates in $L^2(H^1)$-norm are optimal in space and time. The error estimate in $L^2(L^2)$-norm is optimal in time, but suboptimal in space. The technique applied in this paper can be extended to the case the discontinuous Galerkin time semidiscretization combined with the *hp* discontinuous Galerkin space discretization. Of course, the analysis of this case would be still more technical. The error estimates have been obtained with the aid of a "parabolic machinery" for problems with "dominating diffusion". This means that the results are not applicable to conservation laws, solved by the finite difference or finite volume methods and treated e.g. in [13] – [15].

There are the following subjects for further work:

- derivation of optimal error estimates in space and time in the case of the SIPG method,
- numerical realization of the discrete problem and the demonstration of results by numerical experiments,
- analysis of the effect of numerical integration in space and time integrals,
- extension of the results to problems with nonlinear convection as well as diffusion,
- analysis of the combination of the time DGFEM with other space DG discretizations, as e.g. the LDG method (cf. [10], [17]),
- application of the space-time DGFEM to the numerical solution of some technically relevant problems, as, e. g. interaction of compressible flow with structures.

Acknowledgements This work was a part of the project No. MSM 0021620839 financed by the Ministry of Education of the Czech Republic and was partly supported under the Grant No. 201/08/0012 of the Czech Grant Agency of the Czech Republic.

References

[1] Akrivis, G., Makridakis, C.: Galerkin time-stepping methods for nonlinear parabolic equations. *ESAIM: Math. Modelling and Numer. Anal.*, **38**, 261–289 (2004).

[2] Arnold, D.N.: An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, **19**, 742–760 (1982).

[3] Arnold, D.N., Brezzi, F., Cockburn, B., Marini, D.: Discontinuos Galerkin methods for elliptic problems. In: Discontinuous Galerkin methods. *Theory, Computation and Applications. Lecture Notes in Computational Science and Engineering 11*, Springer, Berlin, 89–101 (2000).

[4] Arnold, D.N., Brezzi, F., Cockburn, B., Marini, D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, **39**, 1749–1779 (2001).

[5] Babuška, I., Baumann, C.E., Oden, J.T.: A discontinuous *hp* finite element method for diffusion problems, 1-D analysis. *Comput. Math. Appl.*, **37**, 103–122 (1999).

[6] Baker, G. A., Bramble, J. H., Thomée, V.: Single step Galerkin approximations for parabolic problems. *Math. Comp.* **31**, 818–847 (1977).

[7] Bassi, F., Rebay, S.: A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations. *J. Comput. Phys.*, **131**, 267–279 (1997).

[8] Bassi, F., Rebay, S.: High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, **138**, 251–285 (1997).

[9] Baumann, C.E., Oden, J.T.: A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations. *Int. J. Numer. Methods Fluids*, **31**, 79–95 (1999).

[10] Castillo, P., Cockburn, B., Schötzau, D., Schwab, C.: Optimal a priori estimates for the *hp*-version of the local discontinuous Galerkin method for convection-difussion problems. *Math. Comp.*, **71**, 455–478 (2001).

[11] Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam (1979).

[12] Cockburn, B., Dong, B., Guzmán, J.: Optimal convergence of the original DG method for the transport-reaction equation on special meshes. *SIAM J. Numer. Anal.* **46**, 1250-1265 (2008).

[13] Cockburn, B., Gremaud, P.-A.: A priori error estimates for numerical methods for scalar conservation laws. Part I: The general approach. *Math. Comp.* **65**, 533–573 (1996).

[14] Cockburn, B., Gremaud, P.-A.: A priori error estimates for numerical methods for scalar conservation laws. Part II: Flux-splitting monotone schemes on irregular Cartesion grids. *Math. Comp.* **66**, 547–572 (1997).

[15] Cockburn, B., Gremaud, P.-A., Xiangrong Yang, J.: A priori error estimates for numerical methods for scalar conservation laws. Part III: Multidimensional flux-splitting monotone schemes on non-Cartesion grids. *SIAM J. Numer. Anal.* **35**, 1775–1803 (1998).

[16] Cockburn, B., Shu, C.-W.: TVB Runge–Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II. General framework. *Math. Comp.*, **52**, 411–435 (1989).

[17] Cockburn, B., Shu, C.-W.: The local discontinuous Gelerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.* **35**, 2440–2463.

[18] Cockburn, B., Shu, C.-W.: Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. Review article. *J. Sci. Comput.* **16**, 173–261 (2001).

[19] Dawson, C., Aizinger, V.: A discontinuous Galerkin method for three-dimensional shallow water equations. *J. Sci. Comput.*, **22-23**, 245–267 (2005).

[20] Dolejší, V.: Semi-implicit interior penalty discontinuous Galerkin methods for viscous compressible flows. *Commun. Comput. Phys.*, **4**, 231–274 (2008).

[21] Dolejší, V., Feistauer, M.: Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems. *Numer. Func. Anal. Optimiz.*, **26**, 2709–2733 (2005).

[22] Dolejší, V., Feistauer, M.: A semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow. *J. Comput. Phys.*, **198**, 727–746 (2004).

[23] Dolejší, V., Feistauer, M., Hozman, J.: Analysis of semi-implicit DGFEM for nonlinear convection-diffusion problems on nonconforming meshes. *Comput. Methods Appl. Mech. Engrg.*, **196**, 2813-2827 (2007).

[24] Dolejší, V., Feistauer, M., Kučera, V.: On the discontinuous Galerkin method for the simulation of compressible flow with wide range of Mach numbers. *Comput. Visual. Sci.*, **10**, 17-27 (2007).

[25] Dolejší, V., Feistauer, M., Schwab, C.: A finite volume discontinuous Galerkin scheme for nonlinear convection-diffusion problems. *Calcolo*, **39**, 1–40 (2002).

[26] Dolejší, V., Feistauer, M., Sobotíková, V.: A discontinuous Galerkin method for nonlinear convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, **194**, 2709-2733 (2005).

[27] Dolejší, V., Vlasák, M.: Analysis of a BDF-DGFE scheme for nonlinear convection-diffusion problems. *Numer. Math.*, **110**, 405-447 (2008).

[28] Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: *Computational Differential Equations. Cambridge University Press*, Cambridge (1996).

[29] Eriksson, K., Johnson, C.: Adaptive finite element methods for parabolic problems I: A linear model problem. *SIAM J. Numer. Anal.* **28**, 43–77 (1991).

[30] Estep, D., Larsson, S.: The discontinuous Galerkin method for semilinear parabolic problems, *Math. Modelling and Numer. Anal.*, **27**, 35–54 (1993).

[31] Feistauer, M.: Optimal error estimates in the DGFEM for nonlinear convection-diffusion problems. *Numerical Mathematics and Advanced Applications, ENUMATH 2007*, Springer, Heidelberg, 323-330 (2008).

[32] Feistauer, M., Hájek, J., Švadlenka, K.: Space-time discontinuous Galerkin method for solving nonstationary linear convection-diffusion-reaction problems. *Appl. Math.* **52**, 197–234 (2007).

[33] Feistauer, M., Kučera, V.: On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.* **224**, 208-221 (2007).

[34] Feistauer, M., Švadlenka, K.: Discontinuous Galerkin method of lines for solving nonstationary singularly perturbed linear problems. *J. Numer. Math.*, **2**, 97–117 (2004).

[35] Gregory, R.T., Karney, D.L.: *A Collection of Matrices for Testing Computational Algorithms*. Wiley (1969).

[36] Hartmann, R., Houston, P.: Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations. *Technical Report 2001-42 (SFB 359)*, IWR Heidelberg.

[37] Houston, P., Perugia, I., Schötzau, D.: Mixed discontinuous Galerkin approximation of the Maxwell operator. *Technical Report 2002/45, University of Leicester, Department of Mathematics*, 2002 (SIAM J. Numer. Anal., to appear).

[38] Hu, C., Shu, C.-W.: A discontinuos Galerkin finite element method for Hamilton-Jacobi equations. *SIAM J. Sci. Comput.*, **21**, 666–690 (1999).

[39] Houston, P., Schwab, C., Süli, E.: Discontinuous *hp*-finite element methods for advection-diffusion problems. *SIAM J. Numer. Anal.*, **39**, 2133–2163 (2002).

[40] Jaffre, J., Johnson, C., Szepessy, A.: Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *Math. Models Methods Appl. Sci.*, **5**, 367–386 (1995).

[41] Johnson, C., Pitkäranta, J.: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, **46**, 1–26 (1986).

[42] Karakashian, O., Makridakis, C.: A space-time finite element method for the nonlinear Schrödinger equation: the discontinuous Galerkin method. *Math. Comput.*, **67**, 479–499 (1998).

[43] Kufner, A., John, O., Fučík, S.: *Function Spaces*. Academia, Praha (1977).

[44] Le Saint, P., Raviart, P.-A.: On a finite element method for solving the neutron transport equation. *Mathematical Aspects of Finite Elements in Partial Differential Equations*, Academic Press, 89–145 (1974).

[45] Oden, J.T., Babuška, I., Baumann, C.E.: A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys.*, **146**, 491–519 (1998).

[46] Ralston, A.: *A First Course in Numerical Analysis*. McGraw-Hill, Inc., New York (1965).

[47] Reed, W.H., Hill, T.R.: Triangular mesh methods for the neutron transport equation. *Technical Report LA–UR–73–479*, Los Alamos Scientific Laboratory, 1973.

[48] Rivière, B., Wheeler, M.F.: A discontinuous Galerkin method applied to nonlinear parabolic equations. In: Discontinuous Galerkin Methods. Theory, Computation and Applications. *Lecture Notes in Computational Science and Engineering 11*, Springer, Berlin, 231–244 (2000).

[49] Roos, H.-G., Stynes, M., Tobiska, L.: *Robust Numerical Methods for Singularly Perturbed Differential Equations*. Springer, Berlin (2008).

[50] Roubíček, T.: *Nonlinear Partial Differential Equations with Applications*. Birkhäuser, Basel (2005).

[51] Schötzau, D.: hp-DGFEM for Parabolic Evolution Problems. Applications to Diffusion and Viscous Incompressible Fluid Flow. *PhD Dissertation ETH No. 13041*, Zürich (1999).

[52] Schötzau, D., Schwab, C.: An *hp* a priori error analysis of the Discontinuous Galerkin time-stepping method for initial value problems. *Calcolo*, **37**, 207–232 (2000).

[53] Schötzau, D., Schwab, C., Toselli A.: Mixed hp-DGFEM for incompressible flows. *SIAM J. Numer. Anal.*, **40**, 2171–2194 (2003).

[54] Sun, Shuyn, Wheeler, M.F.: $L^2(H^1)$-norm a posteriori error estimation for discontinuous Galerkin approximations of reactive transport problems. *J. Sci. Comput.*, **22-23**, 501–530 (2005).

[55] Thomée, V.: *Galerkin Finite Element Methods for Parabolic Problems*. Springer, Berlin (2006).

[56] Toselli, A.: HP discontinuous Galerkin approximations for the Stokes problem. *Math. Models Methods Appl. Sci*, **12**, 1565–1597 (2002).

[57] Van der Vegt, J.J.W., Van der Ven, H.: Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flow, part I. General formulation. *J. Comput. Phys.*, **182**, 546–585 (2002).

[58] Wheeler, M.F.: An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, **15**, 152–161 (1978).

On diffusion-uniform error estimates for the DG method applied to singularly perturbed problems

Václav Kučera

Charles University Prague, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Praha 8, Czech Republic

Published in April 2014,
IMA Journal of Numerical Analysis

Abstract

This paper is concerned with the analysis of the discontinuous Galerkin (DG) finite element method applied to a nonstationary nonlinear convection-diffusion problem on quasi-uniform triangulations. Using the technique of Zhang and Shu (2004), we prove apriori error estimates which are uniform with respect to the diffusion coefficient $\varepsilon \to 0$ and valid even in the purely convective case. Zhang and Shu perform their analysis for various explicit schemes using an argument which relies heavily on mathematical induction. We extend the analysis to the method of lines using continuous mathematical induction and a nonlinear Gronwall-type lemma. For an implicit scheme, we prove that standard arguments cannot prove the desired estimates without additional assumptions. For this purpose, we use a suitable continuation of the discrete implicit solution and again use continuous mathematical induction to prove error estimates under a CFL-like condition. Finally, we extend the analysis from globally Lipschitz continuous convective nonlinearities to the locally Lipschitz continuous case.

Keywords: nonlinear convection-diffusion equation; discontinuous Galerkin finite element method; apriori error estimates; continuous mathematical induction; continuation.

1 Introduction

The numerical solution of nonstationary convection-diffusion problems plays an important role in many areas of applied mathematics ranging from fluid dynamics and heat transfer on one side to image processing on the other side. In the numerical treatment of such problems many difficulties arise due to the occurrence of internal and boundary layers, where steep gradients or discontinuities appear. Many numerical methods have been devised to overcome such difficulties. The finite volume (FV) method, which is often used, is based on piecewise constant approximations. It has good stability properties in the vicinity of discontinuities, however it has a low order of accuracy and its generalization to higher order methods is rather sophisticated. On the other hand, the finite element (FE) method with a high order of accuracy is suitable mainly for elliptic problems and various stabilization techniques (e.g. streamline diffusion or Galerkin

least squares methods) must be employed to avoid spurious oscillations in the solution of convection-diffusion problems with dominating convection.

A natural generalization of the FV and FE methods is the discontinuous Galerkin finite element method (DGFEM). This method uses advantages of FV as well as FE methods: it is based on piecewise polynomial discontinuous approximations, where boundary fluxes are evaluated with the aid of a numerical flux. The use of discontinuous functions allows to capture discontinuities and steep gradients, while the use of higher degree polynomials ensures a higher order of approximation in regions, where the solution is smooth.

Originally, the DGFEM was proposed for the solution of a neutron transport linear equation in [35] and analyzed theoretically in [33] and [29]. As for the numerical solution of elliptic and parabolic problems, discontinuous Galerkin methods are proposed and analyzed in the works [42] and [1] with further theoretical analysis in [5], [2] and [3]. The DGFEM was applied to nonlinear conservation laws ([13], [28]) and the numerical solution of compressible flow ([6], [7], [8], [15], [24], [41], [19], [23]) as well as incompressible viscous flow ([36], [40]), porous media flow ([37]), shallow water flow ([14]), the Hamilton-Jacobi equations ([27]), the Schrödinger equation ([30]) and the Maxwell equations ([25]).

This work is concerned with the analysis of the discontinuous Galerkin (DG) finite element method applied to the nonstationary singularly perturbed convection-diffusion problem defined in $\Omega \subset \mathbb{R}^d$ with mixed boundary conditions on quasi-uniform triangulations. Our aim is to derive apriori error estimates in the $L^\infty(L^2)$-norm which are uniform with respect to the diffusion coefficient $\varepsilon \to 0$ and are valid even for the limiting case $\varepsilon = 0$. In the case of linear advection-diffusion this has been done e.g. in [10], [26], [39]. In the nonlinear purely convective case, for various explicit time discretizations, such an error analysis was presented in a series of papers of Zhang and Shu starting with [44]. The typical result for a k-th order explicit scheme for a convective problem is of the form:

Lemma 1. *Let $\mathbf{f} \in [C^2(\mathbb{R})]^d$ and the polynomial order used is $p > (1+d)/2$. Then for sufficiently small h and τ satisfying some CFL-like condition, the error e_h^n of the DG scheme at time level t_n satisfies*

$$\|e_h^n\|_{L^2(\Omega)} \le C(h^{p+1/2} + \tau^k), \quad n = 0, \cdots, N, \tag{1}$$

where $C > 0$ is independent of h, τ, n.

The proof relies on an elegant estimate of the convective terms derived in [44] for the 1D case for periodic or compactly supported solutions under the assumption that the numerical flux is an *E-flux*. Using this estimate, if we know *apriori* that $\|e_h^n\| = O(h^{1+d/2}), n = 0, \cdots, N$, then we may prove the improved estimate (1). A bootstrapping argument using mathematical induction is then applied in order to eliminate the *apriori* assumption.

Since the proof relies heavily on mathematical induction, the technique cannot be directly applied to estimates for the method of lines (no discrete structure with respect to time) and implicit discretizations (not enough *apriori* information about the solution on the next time level). This is a paradoxical situation, since the estimates *per se* are simpler than for the explicit scheme (there are fewer terms to estimate), but their rigorous application is not straightforward.

The structure of the paper is as follows. In Section 2, we introduce the continuous problem, which we discretize in Section 3. In Section 4, we review the necessary results

for our analysis, including properties of the numerical flux, which we assume to be an *E-flux*.

Originally, in [44] the fundamental estimates of convective terms are derived for periodic boundary conditions or compactly supported solutions. In Section 5, we generalize these estimates to the case of mixed Dirichlet-Neumann boundary conditions for $\mathbf{f} \in (C^2(\mathbb{R}) \cap W^{2,\infty}(\mathbb{R}))^d$. Improved estimates are obtained for $\mathbf{f} \in (C^3(\mathbb{R}) \cap W^{3,\infty}(\mathbb{R}))^d$ and Dirichlet conditions on the whole boundary $\partial\Omega$.

In Section 7, we use the estimates of the convective terms to prove similar estimates as Lemma 1 for the method of lines, i.e. space semidiscretization. We apply two different techniques. First, we use the so-called *continuous mathematical induction*, [11], instead of standard mathematical induction in the bootstrapping argument. This is a technique that we shall also use in the implicit case. Alternatively, we prove the same result using a nonlinear variant of Gronwall's inequality. The resulting error estimate is valid only for higher order degrees in space, i.e. $p > 1 + d/2$ or $p > (1 + d)/2$ if $\mathbf{f}$ has higher regularity.

In Section 8, we first prove that for the implicit scheme that an analogy of Lemma 1 cannot be obtained from the error equation and the considered estimates of its individual terms without additional assumptions. Hence, we need to supply more information about the properties of the problem and its (discrete) solution in order to derive the desired error estimates. To this end, we introduce a *continuation* $\widetilde{e}_h : [0, T] \to L^2(\Omega)$ of the error $e_h^n, n = 0, \cdots, N$ constructed by means of a suitable modification of the discrete problem. By definition, estimates for this continuated solution directly imply estimates for the original implicit solution. The fact that $\tilde{e}_h$ is continuous with respect to time and that it relates to the structure of the problem allows us to prove estimates for $\tilde{e}_h$ via continuous mathematical induction. These estimates directly give us the desired estimates for e_h^n. A principal artefact of using the estimates from [44] is that even in the case of an implicit scheme we obtain a rather restrictive CFL-like condition $\tau = O(h^{1+d/2})$ and $\tau = O(h^{(1+d)/2})$ if $\mathbf{f}$ has higher regularity. Furthermore, the result is not valid for the lowest order approximation degrees (we need $p > 1+d/2$ or $p > (1+d)/2$ if $\mathbf{f}$ has higher regularity). Such restrictive assumptions are purely an artefact of the proof due to the nonlinearity of the problem. For linear problems, we may expect the standard $\tau = O(h)$ condition in various time-discretization schemes, [10], [38].

The results of Sections 7 and 8 were derived under the assumption that the convective nonlinearities $\mathbf{f} \in (C^2(\mathbb{R}) \cap W^{2,\infty}(\mathbb{R}))^d$, i.e. under *global* Lipschitz continuity and boundedness. In [44], the *global* Lipschitz case is treated by modifying $\mathbf{f}$ sufficiently far away from the compact set of values attained by u in such a way that these modified nonlinearities are globally Lipschitz continuous and bounded. While this procedure does not change the original problem, one obtains a completely new discrete problem and unless the original discrete problem has a solution in $L^\infty(Q_T)$, one cannot guarantee that the modified and original discrete problems have the same solution. In Section 9, we show how to prove the error estimates from Sections 7 and 8 directly for a *locally* Lipschitz $\mathbf{f} \in (C^2(\mathbb{R}))^d$ without modifying the original equation.

2 Continuous problem

Let $\Omega \subset \mathbb{R}^d, d \in \mathbb{N}$, be a bounded open polygonal (polyhedral) domain with Lipschitz-continuous boundary $\partial\Omega$. For $0 < T < +\infty$, we set $Q_T := \Omega \times (0, T)$. We treat the

following nonlinear convection-diffusion problem. Find $u : Q_T \to \mathbb{R}$ such that

$$\text{a)} \quad \frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) = \varepsilon \Delta u + g \quad \text{in } Q_T, \tag{2}$$

$$\text{b)} \quad u\big|_{\Gamma_D \times (0,T)} = u_D, \tag{3}$$

$$\text{c)} \quad \varepsilon \frac{\partial u}{\partial n}\Big|_{\Gamma_N \times (0,T)} = g_N, \tag{4}$$

$$\text{d)} \quad u(x,0) = u^0(x), \quad x \in \Omega. \tag{5}$$

The diffusion coefficient $\varepsilon \geq 0$ 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 $\partial\Omega = \Gamma_N \cup \Gamma_D$.

We assume that the convective fluxes $\mathbf{f} = (f_1, \cdots, f_d) \in (C_b^2(\mathbb{R}))^d = (C^2(\mathbb{R}) \cap W^{2,\infty}(\mathbb{R}))^d$, hence $\mathbf{f}$ and $\mathbf{f}' = (f_1', \cdots, f_d')$ are *globally Lipschitz continuous*. For improved estimates via Lemma 8, we shall assume the continuity and boundedness of $\mathbf{f}'''$ along with global Lipschitz continuity of $\mathbf{f}'' = (f_1'', \cdots, f_d'')$, i.e. $\mathbf{f} \in (C_b^3(\mathbb{R}))^d = (C^3(\mathbb{R}) \cap W^{3,\infty}(\mathbb{R}))^d$. In Section 9, we shall extend the error analysis, assuming only local Lipschitz continuity and boundedness, i.e. $\mathbf{f} \in (C^2(\mathbb{R}))^d$ and $\mathbf{f} \in (C^3(\mathbb{R}))^d$. As already mentioned, by $\mathbf{f}', \mathbf{f}''$ and $\mathbf{f}'''$ we will denote the vector of component-wise derivatives of $\mathbf{f}$.

Remark 1. *In [44],* local *Lipschitz continuity is treated by modifying* $\mathbf{f}$ *sufficiently far away from the compact set of values attained by* u *in such a way that these modified nonlinearities are globally Lipschitz continuous along with their first derivatives. It is argued that this procedure does not change the solution of (2). However one obtains a completely new discrete problem. Unless one knows* apriori *that the original discrete problem has a solution in* $L^\infty(Q_T)$*, one cannot guarantee that the modified and original discrete problems have the same solution. In Section 9, we show how to prove error estimates directly for a* locally *Lipschitz* $\mathbf{f}$ *without modifying the original equation.*

We use standard notation of function spaces. Let $G \subset \mathbb{R}^d$ be a bounded domain with a Lipschitz-continuous boundary ∂G. By $\overline{G}$ we denote the closure of G. Let $k \in \{0, 1, 2, \dots\}$ and $p \in [1, \infty]$. We use the standard Lebesgue and Sobolev spaces $L^p(G)$, $L^p(\partial G)$, $W^{k,p}(G)$, $H^k(G) = W^{k,2}(G)$. Further, we use the 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 $[0,T]$ with values in X (see e.g. [31]). The symbols $\|\cdot\|_X$ and $|\cdot|_X$ denote a norm and a seminorm in a space X, respectively. By $(\cdot\,,\cdot)$ we denote the standard $L^2(\Omega)-$scalar product and by $\|\cdot\|$ the $L^2(\Omega)$-norm. By $\|\cdot\|_\infty$, we denote the $L^\infty(\Omega)$-norm. For simplicity of notation, we shall drop the argument Ω in Sobolev norms, e.g. $\|\cdot\|_{H^{p+1}}$ denotes the $H^{p+1}(\Omega)-$norm. We will also denote the Bochner norms over the whole interval $[0,T]$ in concise form, e.g. $\|u\|_{L^\infty(H^{p+1})}$ denotes the $L^\infty(0,T;H^{p+1}(\Omega))$-norm.

3 Discretization

3.1 Finite element mesh

Let $\mathcal{T}_h$ be triangulation of Ω, i.e. a partition of $\overline{\Omega}$ into a finite number of closed simpices with mutually disjoint interiors. We do not require the standard conforming properties of $\mathcal{T}_h$ used in the finite element method, i.e. we admit so-called hanging nodes. For $K \in \mathcal{T}_h$ we set $h_K = \operatorname{diam}(K)$, $h = \max_{K \in \mathcal{T}_h} h_K$. By ρ_K we denote the radius of the

largest ball inscribed into K and by $|K|$ we denote the $d-$dimensional Lebesgue measure of K.

Let $K, K' \in \mathcal{T}_h$. We say that K and K' are *neighbors*, if the set $\partial K \cap \partial K'$ has nonzero $(d-1)$-dimensional measure. We say that $\Gamma \subset K$ is a *face* (or *edge* in $\mathbb{R}^2$ and *node* in $\mathbb{R}^1$) of K, if it is a maximal connected open subset either of $\partial K \cap \partial K'$, where K' is a neighbour of K, or of $\partial K \cap \partial\Omega$. By $\mathcal{F}_h$ we denote the system of all faces of all elements $K \in \mathcal{T}_h$.

Further, we define the set of all inner faces, Dirichlet boundary faces, Neumann boundary faces and all boundary faces, respectively, as

$$\mathcal{F}_h^I = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \Omega\}, \qquad \mathcal{F}_h^D = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \Gamma_D\},$$
$$\mathcal{F}_h^N = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \Gamma_N\}, \quad \mathcal{F}_h^B = \mathcal{F}_h^D \cup \mathcal{F}_h^N.$$

Obviously, $\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^B$.

For each $\Gamma \in \mathcal{F}_h$ we define a fixed unit normal vector $\mathbf{n}_\Gamma$. We assume that for $\Gamma \in \mathcal{F}_h^B$ the normal $\mathbf{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega$. Finally, by $|\Gamma|$ we denote the $(d-1)$-dimensional measure of Γ.

3.2 Spaces of discontinuous functions

Over a triangulation $\mathcal{T}_h$ we define the broken Sobolev spaces

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

equipped with the seminorm

$$|v|_{H^k(\mathcal{T}_h)} = \Big(\sum_{K\in\mathcal{T}_h} |v|^2_{H^k(K)}\Big)^{1/2}.$$

For each face $\Gamma \in \mathcal{F}_h^I$ there exist two neighbours $K_\Gamma^{(L)}$, $K_\Gamma^{(R)} \in \mathcal{T}_h$ such that $\Gamma \subset K_\Gamma^{(L)} \cap K_\Gamma^{(R)}$. We use the convention that $\mathbf{n}_\Gamma$ is the outer normal to the element $K_\Gamma^{(L)}$. For $v \in H^1(\mathcal{T}_h)$ and $\Gamma \in \mathcal{F}_h^I$ we introduce:

$$v|_\Gamma^{(L)} = \text{ the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \qquad v|_\Gamma^{(R)} = \text{ the trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma,$$
$$\langle v\rangle_\Gamma = \tfrac{1}{2}\big(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}\big), \qquad [v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}.$$

The value $[v]_\Gamma$ depends on the orientation of $\mathbf{n}_\Gamma$, but the values $\langle v\rangle_\Gamma$ and $[v]_\Gamma\mathbf{n}_\Gamma$ are independent of this orientation. Now, let $\Gamma \in \mathcal{F}_h^B$ and $K_\Gamma^{(L)} \in \mathcal{T}_h$ be such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial\Omega$. For $v \in H^1(\mathcal{T}_h)$ we set

$$v_\Gamma = v|_\Gamma^{(L)} = \text{ the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma,$$

the value $v|_\Gamma^{(R)}$, will be defined depending on the context by boundary conditions, cf. Section 3.4.

If $[\cdot]_\Gamma$ and $\langle\cdot\rangle_\Gamma$ appear in an integral of the form $\int_\Gamma \dots \,\mathrm{d}S$, we omit the subscript Γ and write simply $[\cdot]$ and $\langle\cdot\rangle$. For simplicity we shall use the following notation:

$$\int_{\mathcal{F}_h^I} \dots \,\mathrm{d}S = \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \dots \,\mathrm{d}S$$

and similarly for $\mathcal{F}_h, \mathcal{F}_h^D, \mathcal{F}_h^N$ and $\mathcal{F}_h^B$.

Let $p \geq 1$ be an integer. The approximate solution will be sought in the space of discontinuous piecewise polynomial functions

$$S_h = \{v;\, v|_K \in P^p(K), \forall K \in \mathcal{T}_h\},$$

where $P^p(K)$ denotes the space of all polynomials on K of degree $\leq p$.

3.3 Discontinuous Galerkin space semidiscretization

We introduce the following forms defined for $v, \varphi \in H^2(\mathcal{T}_h)$.
Diffusion form:

$$\begin{aligned} a_h(v,\varphi) = \sum_{K\in\mathcal{T}_h} \int_K \nabla v\cdot\varphi \,\mathrm{d}x - \int_{\mathcal{F}_h^I} \langle\nabla v\rangle\cdot\mathbf{n}[\varphi]\,\mathrm{d}S - \Theta\int_{\mathcal{F}_h^I} \langle\nabla\varphi\rangle\cdot\mathbf{n}[v]\,\mathrm{d}S \\ - \int_{\mathcal{F}_h^D} \nabla v\cdot\mathbf{n}\varphi\,\mathrm{d}S - \Theta\int_{\mathcal{F}_h^D} \nabla\varphi\cdot\mathbf{n}v\,\mathrm{d}S. \end{aligned} \tag{6}$$

Further we define the *interior and boundary penalty jump terms*:

$$J_h(v,\varphi) = \int_{\mathcal{F}_h^I} \sigma[v][\varphi]\,\mathrm{d}S + \int_{\mathcal{F}_h^D} \sigma v\varphi\,\mathrm{d}S \tag{7}$$

and the *right-hand side form*:

$$l_h(\varphi)(t) = \int_\Omega g(t)\varphi\,\mathrm{d}x + \int_{\mathcal{F}_h^N} g_N(t)\varphi\,\mathrm{d}S - \varepsilon\Theta\int_{\mathcal{F}_h^D} \nabla\varphi\cdot\mathbf{n}u_D(t)\,\mathrm{d}S + \varepsilon\int_{\mathcal{F}_h^D} \sigma u_D(t)\varphi\,\mathrm{d}S. \tag{8}$$

The parameter σ in (7) and (8) is constant on every edge and defined by

$$\sigma|_\Gamma = \frac{C_W}{|\Gamma|}, \quad \forall\ \Gamma \in \mathcal{F}_h, \tag{9}$$

where $C_W > 0$ is a constant, which must be chosen large enough to ensure coercivity of the diffusion form – cf. Lemma 9. Depending on the value of Θ in definitions (6) and (8), we obtain the *symmetric, incomplete* and *nonsymmetric interior penalty* variants of the diffusion a right-hand side forms, by taking $\Theta = 1, 0, -1$ respectively.

Finally we define the *convective form*

$$b_h(v,\varphi) = -\sum_{K\in\mathcal{T}_h} \int_K \mathbf{f}(v)\cdot\nabla\varphi\,\mathrm{d}x + \int_{\mathcal{F}_h^I} H(v^{(L)}, v^{(R)}, \mathbf{n})[\varphi]\,\mathrm{d}S + \int_{\mathcal{F}_h^B} H(v^{(L)}, v^{(R)}, \mathbf{n})\varphi^{(L)}\,\mathrm{d}S. \tag{10}$$

For $\Gamma \in \mathcal{F}_h^B$, the value $v^{(R)}$ is defined by boundary conditions, cf. Section 3.4. The form b_h approximates convective terms with the aid of a numerical flux $H(v, w, \mathbf{n})$. We assume that H is defined in $\mathbb{R}^2 \times B_1$, where $B_1 = \{\mathbf{n} \in \mathbb{R}^2; |\mathbf{n}| = 1\}$ and has the following properties:

(H1) $H(v, w, \mathbf{n})$ is *Lipschitz-continuous* with respect to v, w:

$$|H(v,w,\mathbf{n}) - H(v^*,w^*,\mathbf{n})| \leq C_L(|v-v^*| + |w-w^*|), \quad \forall v,\ w,\ v^*,\ w^* \in \mathbb{R},\ \mathbf{n} \in B_1.$$

(H2) $H(v, w, \mathbf{n})$ is *consistent*:

$$H(v,v,\mathbf{n}) = \mathbf{f}(v)\cdot\mathbf{n}, \quad \forall v \in \mathbb{R},\ \mathbf{n} \in B_1.$$

(H3) $H(v, w, \mathbf{n})$ is *conservative*:

$$H(v, w, \mathbf{n}) = -H(w, v, -\mathbf{n}), \quad \forall v,\ w \in \mathbb{R},\ \mathbf{n} \in B_1.$$

(H4) $H(v, w, \mathbf{n})$ is an *E-flux*:

$$\big(H(v, w, \mathbf{n}) - \mathbf{f}(q)\cdot \mathbf{n}\big)(v - w) \geq 0, \quad \forall v,\ w \in \mathbb{R},\ \mathbf{n} \in B_1 \text{ and all } q \text{ between } v, w.$$

Condition $(H4)$ was introduced by Osher in [34], cf. also [4]. This is a generalization of the concept of monotone numerical fluxes.

Lemma 2. *Let $H(v, w, \mathbf{n})$ be a nondecreasing function of its first argument, and a nonincreasing function of its second argument (i.e. H is a monotone numerical flux). Furthermore, let H be consistent. Then H is an E-flux.*

Proof. Without loss of generality assume that $v \leq q \leq w$. Then we have, due to the consistency of H,

$$\begin{aligned}&\big(H(v, w, \mathbf{n}) - \mathbf{f}(q)\cdot \mathbf{n}\big)(v - w)\\ &\quad = \big(H(v, w, \mathbf{n}) - H(q, w, \mathbf{n}) + H(q, w, \mathbf{n}) - H(q, q, \mathbf{n})\cdot \mathbf{n}\big)(v - w) \geq 0,\end{aligned}$$

since H is nondecreasing in its first and nonincreasing in its second arguments. □

Remark 2. *Many commonly used numerical fluxes are monotone, e.g. Lax-Friedrichs, Godunov, Engquist-Osher and the Roe flux with entropy fix, cf. [4].*

3.4 Boundary conditions

In the work [44], boundary conditions are avoided by assuming either periodic boundary conditions, or compactly supported solutions (i.e. $u = 0$ on a neighborhood of $\partial\Omega$). In our work, we shall treat general Dirichlet and Neumann boundary conditions, however there are subtleties involved.

In the definition of form b_h, it is necessary to specify the state $v|_\Gamma^{(R)}$ on edges $\Gamma \in \mathcal{F}_h^B$. This is a delicate task due to the convective nature of the problem for $\varepsilon = 0$ or $\varepsilon \ll 1$. It is common practice to prescribe a Dirichlet boundary condition on 'inflow' edges, where the characteristics enter the domain and a Neumann boundary condition on 'outflow' edges, where characteristics leave the domain. The question arises whether to define Γ_D and Γ_N by the exact solution u or by u_h. Our analysis allows for both possibilities, our only requirement is that either $\mathbf{f}'(u(x,t)).\mathbf{n} \geq 0$ or $\mathbf{f}'(u_h(x,t)).\mathbf{n} \geq 0$ on Γ_N. In other words, Γ_N is (a subset of) the outflow boundary for either the exact, or numerical solution.

3.4.1 Boundary conditions depending on u

We assume that the sets Γ_D and Γ_N, which appear in the definition of b_h, are defined by the exact solution u. Specifically, we assume that

$$\begin{aligned}&\Gamma_N^{(t)} \subseteq \{x \in \partial\Omega; \mathbf{f}'(u(x,t)).\mathbf{n} \geq 0\},\\ &\Gamma_D^{(t)} := \partial\Omega \setminus \Gamma_N.\end{aligned} \tag{11}$$

The first condition in (11) is necessary in our analysis. We note that $\Gamma_N^{(t)}$ and $\Gamma_D^{(t)}$ may depend on t, however for convenience we shall drop the superscript (t) in the notation. In (10), we define

$$v|_\Gamma^{(R)} = \begin{cases} u_D, & \text{for } \Gamma \subset \Gamma_D, \\ v|_\Gamma^{(L)}, & \text{for } \Gamma \subset \Gamma_N. \end{cases}$$

3.4.2 Boundary conditions depending on u_h

We assume that the sets Γ_D and Γ_N, which appear in the definition of b_h, are defined by the approximate solution u_h. Specifically, we assume that

$$\begin{aligned} &\Gamma_{N,h}^{(t)} \subseteq \{x \in \partial\Omega; \mathbf{f}'(u_h(x,t)).\mathbf{n} \geq 0\}, \\ &\Gamma_{D,h}^{(t)} := \partial\Omega \setminus \Gamma_N. \end{aligned} \tag{12}$$

Again, the first condition in (12) is necessary in our analysis. As in the previous case, $\Gamma_{N,h}^{(t)}$ and $\Gamma_{D,h}^{(t)}$ may depend on t. In (10), we define

$$v|_\Gamma^{(R)} = \begin{cases} u_D, & \text{for } \Gamma \subset \Gamma_{D,h}, \\ v|_\Gamma^{(L)}, & \text{for } \Gamma \subset \Gamma_{N,h}. \end{cases}$$

This choice is often used in practical computations, where we do not know u in advance and therefore cannot use the approach from section 3.4.1. On the other hand, it may happen that Γ_D, as defined in (12), may contain a subset on which u_D is not defined in the original continuous problem. This problem does not arise, if we have u_D defined on the whole $\partial\Omega$, even though some parts of the boundary will belong to Γ_N, e.g. when u_D represents some 'far-field' state. The situation is trivial, if we prescribe a Dirichlet boundary condition on the whole $\partial\Omega$, i.e. $\Gamma_N = \emptyset$.

Definition 1. *We say that $u_h \in C^1([0,T]; S_h)$ is a DGFE solution of the convection-diffusion problem (2) - (5), if $u_h(0) = u_h^0$, an S_h approximation of the initial condition u^0, and for all $t \in (0,T)$*

$$\frac{d}{dt}\big(u_h(t), \varphi_h\big) + b_h\big(u_h(t), \varphi_h\big) + \varepsilon J_h\big(u_h(t), \varphi_h\big) + \varepsilon a_h\big(u_h(t), \varphi_h\big) = l_h\big(\varphi_h\big)(t), \quad \forall \varphi_h \in S_h. \tag{13}$$

Similarly as in [16] we can show that a sufficiently regular exact solution u of problem (2) satisfies

$$\frac{d}{dt}\big(u(t), \varphi_h\big) + b_h\big(u(t), \varphi_h\big) + \varepsilon J_h\big(u(t), \varphi_h\big) + \varepsilon a_h\big(u(t), \varphi_h\big) = l_h\big(\varphi_h\big)(t), \tag{14}$$

for all $\varphi_h \in S_h$ and for all $t \in (0,T)$, which implies the *Galerkin orthogonality property* of the error.

4 Some necessary results and assumptions

4.1 Regularity of the exact solution

We assume that the weak solution u is sufficiently regular, namely

$$u_t := \frac{\partial u}{\partial t} \in L^2\big(0,T; H^{p+1}(\Omega)\big), \quad u \in L^\infty(0,T; W^{1,\infty}(\Omega)), \tag{15}$$

where $p \geq 1$ denotes the given degree of approximation. Under these conditions, u satisfies equation (2) pointwise and $u \in C\big([0,T]; H^{p+1}(\Omega)\big)$.

4.2 Geometry of the mesh

We consider a system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$, $h_0 > 0$, of quasi-uniform triangulations of Ω with the following properties:

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

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

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

$$h_K \leq C_2\, d(\Gamma), \quad \forall K \in \mathcal{T}_h, \quad \forall \Gamma \subset \partial K,\ \Gamma \in \mathcal{F}_h, \quad \forall h \in (0, h_0).$$

(3) System $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ satisfies the *inverse assumption*: There exists a constant $C_3 > 0$ such that

$$h \leq C_3\, h_K, \quad \forall K \in \mathcal{T}_h, \quad \forall h \in (0, h_0).$$

4.3 Some auxiliary results

Throughout this work we denote by C a generic constant independent of h, t, ε. Now we can state two necessary results needed in the following analysis:

Lemma 3 (Multiplicative trace inequality)**.** *There exists a constant $C_M > 0$ independent of h, K such that for all $K \in \mathcal{T}_h$, $v \in H^1(K)$ and $h \in (0, h_0)$*

$$||v||^2_{L^2(\partial K)} \leq C_M\big(||v||_{L^2(K)}|v|_{H^1(K)} + h_K^{-1}||v||^2_{L^2(K)}\big).$$

Proof. Cf. [9] and [21] for a detailed proof. □

Lemma 4 (Inverse inequalities)**.** *There exists a constant $C_I > 0$ independent of h, K such that for all $K \in \mathcal{T}_h$ and $v \in P^p(K)$*

$$|v|_{H^1(K)} \leq C_I h_K^{-1}||v||_{L^2(K)},$$
$$\|v\|_{L^\infty(K)} \leq C_I h_K^{-d/2}\|v\|_{L^2(K)}.$$

Proof. Cf. e.g. [12]. □

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

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

Using this projection, we define the following quantities used in the error splitting:

$$\eta_h(t) := u(t) - \Pi_h u(t) \in H^{p+1}(\mathcal{T}_h), \qquad \xi_h(t) = \Pi_h u(t) - u_h(t) \in S_h$$

for $t \in (0, T)$. Then we can write the error e_h as $e_h(t) := u(t) - u_h(t) = \eta_h(t) + \xi_h(t)$. For simplicity, we shall usually drop the subscript h and the argument t.

Lemma 5. *There exists a constant $C > 0$ independent of h, K such that for all $h \in (0, h_0)$*

$$\begin{aligned}
&a)\ ||\eta_h(t)|| \le Ch^{p+1}|u(t)|_{H^{p+1}},\\
&b)\ |\eta_h(t)|_{H^1(\mathcal{T}_h)} \le Ch^{p}|u(t)|_{H^{p+1}},\\
&c)\ \Big|\Big|\frac{\partial \eta_h(t)}{\partial t}\Big|\Big| \le Ch^{p+1}\Big|\frac{\partial u(t)}{\partial t}\Big|_{H^{p+1}},\\
&d)\ ||\eta_h(t)||_\infty \le Ch^{p}|u(t)|_{H^{p+1}},\\
&e)\ ||\eta_h(t)||_\infty \le Ch|u(t)|_{W^{1,\infty}}.
\end{aligned}$$

Proof. The proof follows from standard approximation results found e.g. in [12]. □

4.4 Properties of the numerical flux

Definition 2. *For $v \in H^1(\mathcal{T}_h)$ we define the function $\alpha(v) = \alpha\big(v^{(L)}, v^{(R)}\big)$ on each $\Gamma \in \mathcal{F}_h$ by*

$$\alpha(v)\big|_\Gamma = \begin{cases} [v]^{-1}\big(H(v^{(L)}, v^{(R)}, \mathbf{n}) - \mathbf{f}(\langle v\rangle)\cdot\mathbf{n}\big), & \text{if } [v] \neq 0,\\ \big|\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n}\big|, & \text{if } [v] = 0. \end{cases}$$

Here $v^{(R)} : \partial\Omega \to \mathbb{R}$ is an arbitrarily defined but fixed function. .

Lemma 6. *(cf. [44]) There exists a constant $C \ge 0$, such that for all $v \in H^1(\mathcal{T}_h)$, we have $0 \le \alpha(v) \le C$ and*

$$\big|\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n}\big| \le 2\alpha(v) + C\big|[v]\big|. \tag{16}$$

Moreover, if $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$, we have the estimate

$$\mathbf{f}''(\langle v\rangle)\cdot\mathbf{n}[v] \le 8\alpha(v) + C[v]^2. \tag{17}$$

Proof. Since H is an E-flux, we have non-negativity of $\alpha(v)$. Assumptions (H1), (H2) imply the boundedness of $\alpha(v)$. Inequality (16) is trivially satisfied if $[v] = 0$. Otherwise, we have two cases:

(i) Let $\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} \ge 0$. Then

$$\alpha(v) = [v]^{-1}\big(H(v^{(L)}, v^{(R)}, \mathbf{n}) - \mathbf{f}(v^{(L)})\cdot\mathbf{n}\big) + [v]^{-1}\big(\mathbf{f}(v^{(L)})\cdot\mathbf{n} - \mathbf{f}(\langle v\rangle)\cdot\mathbf{n}\big). \tag{18}$$

The first term is nonnegative due to assumption (H4). In the second term we shall use a Taylor expansion in the point $\langle v\rangle$. We obtain

$$\alpha(v) \ge [v]^{-1}\Big(\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n}\big(v^{(L)} - \langle v\rangle\big) + \frac{1}{2}\mathbf{f}''_{v^{(L)},\langle v\rangle}\cdot\mathbf{n}\big(v^{(L)} - \langle v\rangle\big)^2\Big) \ge \frac{1}{2}\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} - C\big|[v]\big|,$$

where $\mathbf{f}''_{v^{(L)},\langle v\rangle} \le \|\mathbf{f}\|_{W^{2,\infty}(\mathbb{R})}$ is the Lagrange form of the remainder of the Taylor expansion, i.e. $\mathbf{f}''_{v^{(L)},\langle v\rangle}(x)$ has components $f_s''(\vartheta_s v^{(L)}(x) + (1-\vartheta_s)\langle v(x)\rangle)$ for some $\vartheta_s \in [0,1]$ and $s = 1, \cdots, d$.

(ii) Let $\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} < 0$. Then

$$\alpha(v) = [v]^{-1}\big(H(v^{(L)}, v^{(R)}, \mathbf{n}) - \mathbf{f}(v^{(R)})\cdot\mathbf{n}\big) + [v]^{-1}\big(\mathbf{f}(v^{(R)})\cdot\mathbf{n} - \mathbf{f}(\langle v\rangle)\cdot\mathbf{n}\big). \tag{19}$$

Again, the first term is nonnegative and we use the Taylor expansion in the second:

$$\alpha(v) \ge [v]^{-1}\Big(\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n}\big(v^{(R)} - \langle v\rangle\big) + \frac{1}{2}\mathbf{f}''_{v^{(R)},\langle v\rangle}\cdot\mathbf{n}\big(v^{(R)} - \langle v\rangle\big)^2\Big) \ge -\frac{1}{2}\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} - C\big|[v]\big|.$$

Here $\mathbf{f}''_{v^{(R)},\langle v\rangle}(x) := \mathbf{f}''(\vartheta v^{(R)}(x) + (1-\vartheta)\langle v(x)\rangle)$ for some $\vartheta \in [0,1]$.

As for (17), this is trivially satisfied if $[v] = 0$. Otherwise, we have two cases:

(i) Let $\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} \geq 0$. Then we can write $\alpha(v)$ as in (18). The first term is nonnegative and we apply a third order Taylor expansion to the second:

$$\begin{aligned}\alpha(v) &\geq [v]^{-1}\Big(\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n}\big(v^{(L)} - \langle v\rangle\big) + \frac{1}{2}\mathbf{f}''(\langle v\rangle)\cdot\mathbf{n}\big(v^{(L)} - \langle v\rangle\big)^2 + \frac{1}{2}\mathbf{f}'''_{v^{(L)},\langle v\rangle}\cdot\mathbf{n}\big(v^{(L)} - \langle v\rangle\big)^3\Big)\\ &= \frac{1}{2}\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} + \frac{1}{8}\mathbf{f}''(\langle v\rangle)\cdot\mathbf{n}[v] + \frac{1}{2}\mathbf{f}'''_{v^{(L)},\langle v\rangle}\cdot\mathbf{n}[v]^2 \geq \frac{1}{8}\mathbf{f}''(\langle v\rangle)\cdot\mathbf{n}[v] - C[v]^2.\end{aligned}$$

Again, $\mathbf{f}'''_{v^{(L)},\langle v\rangle} \leq \|\mathbf{f}\|_{W^{3,\infty}(\mathbb{R})}$ is the Lagrange form of the remainder of the Taylor expansion.

(ii) Let $\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} < 0$. Then from (19), by use of the Taylor expansion,

$$\begin{aligned}\alpha(v) &\geq [v]^{-1}\Big(\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n}\big(v^{(R)} - \langle v\rangle\big) + \frac{1}{2}\mathbf{f}''(\langle v\rangle)\cdot\mathbf{n}\big(v^{(R)} - \langle v\rangle\big)^2 + \frac{1}{2}\mathbf{f}'''_{v^{(R)},\langle v\rangle}\cdot\mathbf{n}\big(v^{(R)} - \langle v\rangle\big)^3\Big)\\ &= -\frac{1}{2}\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n} + \frac{1}{8}\mathbf{f}''(\langle v\rangle)\cdot\mathbf{n}[v] - \frac{1}{2}\mathbf{f}'''_{v^{(R)},\langle v\rangle}\cdot\mathbf{n}[v]^2 \geq \frac{1}{8}\mathbf{f}''(\langle v\rangle)\cdot\mathbf{n}[v] - C[v]^2.\end{aligned}$$

□

5 Analysis of the convective terms

Definition 3 (Bilinear form $\widetilde{J}_h$)**.** *Let $v, w \in H^1(\mathcal{T}_h)$. We define*

$$\widetilde{J}_h(v,w) := \int_{\mathcal{F}_h^I} \alpha(u_h)[v][w]\,\mathrm{d}S + \int_{\mathcal{F}_h^D} \alpha(u_h)v^{(L)}w^{(L)}\,\mathrm{d}S + \frac{1}{2}\int_{\mathcal{F}_h^N} \alpha(u)v^{(L)}w^{(L)}\,\mathrm{d}S. \quad (20)$$

Here we must interpret $\alpha(u_h) = \alpha(u_h^{(L)}, u_D)$ on $\mathcal{F}_h^D$ and $\alpha(u) = \alpha(u^{(L)}, u^{(L)})$ on $\mathcal{F}_h^N$.

Remark 3. *We define $\widetilde{J}_h$ by (20) if we use the approach to boundary conditions outlined in Section 3.4.1. If we proceed as in Section 3.4.2, we use instead*

$$\widetilde{J}_h(v,w) := \int_{\mathcal{F}_h^I} \alpha(u_h)[v][w]\,\mathrm{d}S + \int_{\mathcal{F}_h^D} \alpha(u_h)v^{(L)}w^{(L)}\,\mathrm{d}S + \frac{1}{2}\int_{\mathcal{F}_h^N} \alpha(u_h)v^{(L)}w^{(L)}\,\mathrm{d}S, \quad (21)$$

where we interpret $\alpha(u_h) = \alpha(u_h^{(L)}, u_h^{(L)})$ on $\mathcal{F}_h^N$.

Remark 4. *From $\alpha(\cdot) \geq 0$, it follows that $\big(\widetilde{J}_h(\cdot,\cdot)\big)^{1/2}$ is a seminorm on $H^1(\mathcal{T}_h)$.*

Lemma 7. *There exists a constant $C \geq 0$ independent of h, t, ε, such that*

$$b_h(u_h,\xi) - b_h(u,\xi) \leq C\Big(1 + \frac{\|e_h(t)\|_\infty^2}{h^2}\Big)\big(h^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2\big) - \frac{1}{2}\widetilde{J}_h(\xi,\xi). \quad (22)$$

Proof. We write

$$\begin{aligned}b_h(u_h,\xi) - b_h(u,\xi) &= \sum_{K\in\mathcal{T}_h}\int_K \big(\mathbf{f}(u) - \mathbf{f}(u_h)\big)\cdot\nabla\xi\,\mathrm{d}x - \int_{\mathcal{F}_h^I}\big(\mathbf{f}(u)\cdot\mathbf{n} - \mathbf{f}(\langle u_h\rangle)\cdot\mathbf{n}\big)[\xi]\,\mathrm{d}S\\ &- \int_{\mathcal{F}_h^I}\Big(\mathbf{f}(\langle u_h\rangle)\cdot\mathbf{n} - H(u_h^{(L)}, u_h^{(R)}, \mathbf{n})\Big)[\xi]\,\mathrm{d}S - \int_{\mathcal{F}_h^B}\Big(\mathbf{f}(u)\cdot\mathbf{n} - \mathbf{f}\big(\tfrac{1}{2}(u + u_h^{(L)})\big)\cdot\mathbf{n}\Big)\xi^{(L)}\,\mathrm{d}S\\ &- \int_{\mathcal{F}_h^B}\Big(\mathbf{f}\big(\tfrac{1}{2}(u + u_h^{(L)})\big)\cdot\mathbf{n} - H(u_h^{(L)}, u_h^{(R)}, \mathbf{n})\Big)\xi^{(L)}\,\mathrm{d}S.\end{aligned} \quad (23)$$

By the Taylor expansion of $\mathbf{f}$ with respect to u, we have

$$\begin{aligned}
\mathbf{f}(u)-\mathbf{f}(u_h) &= \mathbf{f}'(u)\xi+\mathbf{f}'(u)\eta-\frac{1}{2}\mathbf{f}''_{u,u_h}e_h^2,\\
\mathbf{f}(u)-\mathbf{f}(\langle u_h\rangle) &= \mathbf{f}'(u)\langle\xi\rangle+\mathbf{f}'(u)\langle\eta\rangle-\frac{1}{2}\mathbf{f}''_{u,\langle u_h\rangle}\langle e_h\rangle^2,\\
\mathbf{f}(u)-\mathbf{f}\big(\tfrac{1}{2}(u+u_h^{(L)})\big) &= \frac{1}{2}\mathbf{f}'(u)\xi^{(L)}+\frac{1}{2}\mathbf{f}'(u)\eta^{(L)}-\frac{1}{8}\mathbf{f}''_{u,u_h^{(L)}}\big(e_h^{(L)}\big)^2,
\end{aligned}\tag{24}$$

where $\mathbf{f}''_{u,u_h}, \mathbf{f}''_{u,\langle u_h\rangle}$ and $\mathbf{f}''_{u,u_h^{(L)}}$ are the Lagrange forms of the remainder of the Taylor expansion, i.e. $\mathbf{f}''_{u,u_h}(x,t)$ has components $f_s''(\vartheta_s u(x,t)+(1-\vartheta_s)u_h(x,t))$ for some $\vartheta_s\in[0,1]$ and $s=1,\cdots,d$. Similarly, we define $\mathbf{f}''_{u,\langle u_h\rangle}$ and $\mathbf{f}''_{u,u_h^{(L)}}$.

Due to Green's theorem, we have

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}'(u)\cdot\nabla\xi\,\xi\,\mathrm{d}x=\\
&\quad-\frac{1}{2}\sum_{K\in\mathcal{T}_h}\int_K \operatorname{div}\big(\mathbf{f}'(u)\big)\xi^2\,\mathrm{d}x+\int_{\mathcal{F}_h^I}\mathbf{f}'(u)\cdot\mathbf{n}\langle\xi\rangle[\xi]\,\mathrm{d}S+\frac{1}{2}\int_{\mathcal{F}_h^B}\mathbf{f}'(u)\cdot\mathbf{n}(\xi^{(L)})^2\,\mathrm{d}S.
\end{aligned}$$

This and (23) – (24) implies

$$\begin{aligned}
&b_h(u_h,\xi)-b_h(u,\xi)=\\
&\underbrace{-\frac{1}{2}\sum_{K\in\mathcal{T}_h}\int_K \operatorname{div}\big(\mathbf{f}'(u)\big)\xi^2\,\mathrm{d}x}_{Y_1}+\underbrace{\int_{\mathcal{F}_h^I}\mathbf{f}'(u)\cdot\mathbf{n}\langle\xi\rangle[\xi]\,\mathrm{d}S}_{Y_2}+\underbrace{\frac{1}{2}\int_{\mathcal{F}_h^B}\mathbf{f}'(u)\cdot\mathbf{n}(\xi^{(L)})^2\,\mathrm{d}S}_{Y_3}\\
&+\underbrace{\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}'(u)\cdot\nabla\xi\,\eta\,\mathrm{d}x}_{Y_4}-\underbrace{\frac{1}{2}\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''_{u,u_h}\cdot\nabla\xi\,e_h^2\,\mathrm{d}x}_{Y_5}-\underbrace{\int_{\mathcal{F}_h^I}\mathbf{f}'(u)\cdot\mathbf{n}\langle\xi\rangle[\xi]\,\mathrm{d}S}_{Y_6}\\
&\underbrace{-\int_{\mathcal{F}_h^I}\mathbf{f}'(u)\cdot\mathbf{n}\langle\eta\rangle[\xi]\,\mathrm{d}S}_{Y_7}+\underbrace{\frac{1}{2}\int_{\mathcal{F}_h^I}\mathbf{f}''_{u,\langle u_h\rangle}\cdot\mathbf{n}\,\langle e_h\rangle^2[\xi]\,\mathrm{d}S}_{Y_8}\\
&\underbrace{-\int_{\mathcal{F}_h^I}\Big(\mathbf{f}(\langle u_h\rangle)\cdot\mathbf{n}-H(u_h^{(L)},u_h^{(R)},\mathbf{n})\Big)[\xi]\,\mathrm{d}S}_{Y_9}\\
&\underbrace{-\frac{1}{2}\int_{\mathcal{F}_h^B}\mathbf{f}'(u)\cdot\mathbf{n}\big(\xi^{(L)}\big)^2\,\mathrm{d}S}_{Y_{10}}-\underbrace{\frac{1}{2}\int_{\mathcal{F}_h^B}\mathbf{f}'(u)\cdot\mathbf{n}\,\eta^{(L)}\xi^{(L)}\,\mathrm{d}S}_{Y_{11}}+\underbrace{\frac{1}{8}\int_{\mathcal{F}_h^B}\mathbf{f}''_{u,u_h^{(L)}}\cdot\mathbf{n}\,\big(e_h^{(L)}\big)^2\xi^{(L)}\,\mathrm{d}S}_{Y_{12}}\\
&\underbrace{-\int_{\mathcal{F}_h^B}\Big(\mathbf{f}\big(\tfrac{1}{2}(u+u_h^{(L)})\big)\cdot\mathbf{n}-H(u_h^{(L)},u_h^{(R)},\mathbf{n})\Big)\xi^{(L)}\,\mathrm{d}S}_{Y_{13}}.
\end{aligned}\tag{25}$$

We shall estimate these terms individually.

(A) Term $\mathbf{Y_1}$: Due to the boundedness of $\mathbf{f}''$ and the regularity of u, we have

$$|Y_1|\le C\|\xi\|^2.$$

(B) Terms $\mathbf{Y_2}, \mathbf{Y_6}$: These terms cancel each other.

(C) Terms $\mathbf{Y_3}, \mathbf{Y_{10}}$: These terms cancel each other.
(D) Term $\mathbf{Y_4}$: We set $\overline{u}_K := \frac{1}{|K|}\int_K u \,\mathrm{d}x$. Standard approximation results, cf. [12], imply

$$\|u - \overline{u}_K\|_{L^\infty(K)} \le Ch_K |u|_{L^\infty(W^{1,\infty})} \le Ch_K.$$

Furthermore, due to the definition of η, we have $\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}'(\overline{u}_K)\cdot\nabla\xi\,\eta\,\mathrm{d}x = 0$, since $\mathbf{f}'(\overline{u}_K)\cdot\nabla\xi|_K \in P^p(K)$. Therefore, by the Lipschitz continuity of $\mathbf{f}'$,Lemma 5, a) and Young's inequality

$$\begin{aligned}|Y_4| &= \Big|\sum_{K\in\mathcal{T}_h}\int_K \big(\mathbf{f}'(u) - \mathbf{f}'(\overline{u}_K)\big)\cdot\nabla\xi\,\eta\,\mathrm{d}x\Big| \\ &\le \sum_{K\in\mathcal{T}_h} C\|u-\overline{u}_K\|_{L^\infty(K)} C_I h_K^{-1}\|\xi\|_{L^2(K)}\|\eta\|_{L^2(K)} \le Ch^{2p+2}|u(t)|^2_{H^{p+1}} + \|\xi\|^2.\end{aligned}$$

(E) Term $\mathbf{Y_5}$: We apply the inverse inequality, Lemma 5, a) and Young's inequality:

$$|Y_5| \le C\|e_h\|_\infty \|e_h\| C_I h^{-1}\|\xi\| \le Ch^{-2}\|e_h\|^2_\infty\big(Ch^{2p+2}|u(t)|^2_{H^{p+1}} + \|\xi\|^2\big) + \|\xi\|^2$$

(F) Terms $\mathbf{Y_7}, \mathbf{Y_{11}}$: First we estimate $\mathbf{f}'(u)\cdot\mathbf{n}$. Due to the Lipschitz continuity of $\mathbf{f}'$ and (16), we have

$$\begin{aligned}\text{On } \mathcal{F}_h^I: \quad |\mathbf{f}'(u)\cdot\mathbf{n}| &\le |\mathbf{f}'(u)\cdot\mathbf{n} - \mathbf{f}'(\langle u_h\rangle)\cdot\mathbf{n}| + |\mathbf{f}'(\langle u_h\rangle)\cdot\mathbf{n}| \\ &\le C|u - \langle u_h\rangle| + 2\alpha(u_h) + C\big|[u_h]\big| \le C\|e_h\|_\infty + 2\alpha(u_h). \\ \text{On } \mathcal{F}_h^D: \quad |\mathbf{f}'(u)\cdot\mathbf{n}| &\le \big|\mathbf{f}'(u)\cdot\mathbf{n} - \mathbf{f}'\big(\tfrac12(u+u_h^{(L)})\big)\cdot\mathbf{n}\big| + \big|\mathbf{f}'\big(\tfrac12(u+u_h^{(L)})\big)\cdot\mathbf{n}\big| \\ &\le C\|e_h\|_\infty + 2\alpha(u_h). \\ \text{On } \mathcal{F}_h^N: \quad |\mathbf{f}'(u)\cdot\mathbf{n}| &= \alpha(u).\end{aligned} \tag{26}$$

By applying these estimates in Y_7, Y_{11}, Young's inequality, the *multiplicative trace* and *inverse* inequalities and Lemma 5, a), b), we obtain

$$\begin{aligned}|Y_7 + Y_{11}| &\le \Big(\int_{\mathcal{F}_h^I}\langle\eta\rangle^2\,\mathrm{d}S\Big)^{1/2}\Big(\int_{\mathcal{F}_h^I} 2\big(C\|e_h\|^2_\infty + 2\alpha(u_h)^2\big)[\xi]^2\,\mathrm{d}S\Big)^{1/2} \\ &\quad + \Big(\int_{\mathcal{F}_h^D}\big(\eta^{(L)}\big)^2\,\mathrm{d}S\Big)^{1/2}\Big(\int_{\mathcal{F}_h^D} 2\big(C\|e_h\|^2_\infty + 2\alpha(u_h)^2\big)\big(\xi^{(L)}\big)^2\,\mathrm{d}S\Big)^{1/2} \\ &\quad + \Big(\int_{\mathcal{F}_h^N}\big(\eta^{(L)}\big)^2\,\mathrm{d}S\Big)^{1/2}\Big(\int_{\mathcal{F}_h^N} 2\alpha(u)^2\big(\xi^{(L)}\big)^2\,\mathrm{d}S\Big)^{1/2} \\ &\le Ch^{2p+1}|u(t)|^2_{H^{p+1}} + C\|e_h\|^2_\infty\sum_{K\in\mathcal{T}_h}\int_{\partial K}\xi^2\,\mathrm{d}S \\ &\quad + \frac14\int_{\mathcal{F}_h^I}\alpha(u_h)[\xi]^2\,\mathrm{d}S + \frac14\int_{\mathcal{F}_h^D}\alpha(u_h)\big(\xi^{(L)}\big)^2\,\mathrm{d}S + \frac18\int_{\mathcal{F}_h^N}\alpha(u)\big(\xi^{(L)}\big)^2\,\mathrm{d}S \\ &\le Ch^{2p+1}|u(t)|^2_{H^{p+1}} + Ch^{-1}\|e_h\|^2_\infty\|\xi\|^2 + \frac14\widetilde{J}_h(\xi,\xi).\end{aligned} \tag{27}$$

We cannot estimate the term $\widetilde{J}_h(\xi,\xi)$, it shall be compensated by a similar term in **(H)**.

(G) Terms $\mathbf{Y_8}, \mathbf{Y_{12}}$: By Young's inequality, the *multiplicative trace* and *inverse* inequalities and Lemma 5, a), b), we obtain

$$
\begin{aligned}
&|Y_8 + Y_{12}| \\
&\le C\|e_h\|_\infty \Big(\int_{\mathcal{F}_h^I} \big(e_h^{(L)}\big)^2 + \big(e_h^{(R)}\big)^2 + \big(\xi^{(L)}\big)^2 + \big(\xi^{(R)}\big)^2 \,\mathrm{d}S + \int_{\mathcal{F}_h^B} \big(e_h^{(L)}\big)^2 + \big(\xi^{(L)}\big)^2 \,\mathrm{d}S \Big) \\
&\le C\|e_h\|_\infty \sum_{K\in\mathcal{T}_h} \int_{\partial K} e_h^2\big|_K + \xi^2\big|_K \,\mathrm{d}S \\
&\le C\|e_h\|_\infty \sum_{K\in\mathcal{T}_h} \Big(\|e_h\|_{L^2(K)} |e_h|_{H^1(K)} + h_K^{-1}\|e_h\|_{L^2(K)}^2 + \|\xi\|_{L^2(K)} |\xi|_{H^1(K)} + h_K^{-1}\|\xi\|_{L^2(K)}^2 \Big) \\
&\le C\|e_h\|_\infty \big(Ch^{2p+1}|u(t)|_{H^{p+1}}^2 + Ch^{2p}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2 + C_I h^{-1}\|\xi\|^2 \big) \\
&\le Ch^{-1}\|e_h\|_\infty \big(Ch^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2 \big) \\
&\le C\big(1 + h^{-2}\|e_h\|_\infty^2\big)\big(Ch^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2 \big).
\end{aligned}
$$

In the last inequality we have used the fact that for $x := h^{-1}\|e_h\|_\infty$, we have $x \le 1 + x^2$.

(H) Terms $\mathbf{Y_9}, \mathbf{Y_{13}}$: First we treat the nonlinearities arising in these terms. By the definition of α and boundary conditions, we have

$$
\begin{aligned}
\text{On } \mathcal{F}_h^I: \quad & \mathbf{f}(\langle u_h\rangle)\cdot\mathbf{n} - H(u_h^{(L)}, u_h^{(R)}, \mathbf{n}) = -\alpha(u_h)[u_h] \\
& = \alpha(u_h)[u - u_h] = \alpha(u_h)[\eta] + \alpha(u_h)[\xi] \\
\text{On } \mathcal{F}_h^D: \quad & \mathbf{f}\big(\tfrac{1}{2}(u + u_h^{(L)})\big)\cdot\mathbf{n} - H(u_h^{(L)}, u_h^{(R)}, \mathbf{n}) \\
& = -\alpha(u_h)\big(u_h^{(L)} - u\big) = \alpha(u_h)\eta^{(L)} + \alpha(u_h)\xi^{(L)} \\
\text{On } \mathcal{F}_h^N: \quad & \mathbf{f}\big(\tfrac{1}{2}(u + u_h^{(L)})\big)\cdot\mathbf{n} - H(u_h^{(L)}, u_h^{(R)}, \mathbf{n}) = \mathbf{f}\big(\tfrac{1}{2}(u + u_h^{(L)})\big)\cdot\mathbf{n} - \mathbf{f}\big(u_h^{(L)}\big)\cdot\mathbf{n}.
\end{aligned}
\tag{28}
$$

On the Neumann part of the boundary, we shall employ a Taylor expansion at point u:

$$
\begin{aligned}
\mathbf{f}\big(\tfrac{1}{2}(u + u_h^{(L)})\big) - \mathbf{f}\big(u_h^{(L)}\big) &= \mathbf{f}\big(\tfrac{1}{2}(u + u_h^{(L)})\big) - \mathbf{f}(u) + \mathbf{f}(u) - \mathbf{f}\big(u_h^{(L)}\big) \\
&= -\mathbf{f}'(u)\tfrac{1}{2}e_h^{(L)} + \tfrac{1}{2}\mathbf{f}_1''.\big(\tfrac{1}{2}e_h^{(L)}\big)^2 + \mathbf{f}'(u)e_h^{(L)} - \tfrac{1}{2}\mathbf{f}_2''.\big(e_h^{(L)}\big)^2 \\
&= \tfrac{1}{2}\mathbf{f}'(u)\eta^{(L)} + \tfrac{1}{2}\mathbf{f}'(u)\xi^{(L)} + \tfrac{1}{8}\mathbf{f}_1''.\big(e_h^{(L)}\big)^2 - \tfrac{1}{2}\mathbf{f}_2''.\big(e_h^{(L)}\big)^2,
\end{aligned}
\tag{29}
$$

where $\mathbf{f}_1'', \mathbf{f}_2''$ are the Lagrange forms of the Taylor expansion remainder. By substituting (28), (29) into the definition of Y_9, Y_{13}, we have

$$
\begin{aligned}
Y_9 + Y_{13} &= -\int_{\mathcal{F}_h^I} \alpha(u_h)[\eta][\xi] + \alpha(u_h)[\xi]^2 \,\mathrm{d}S - \int_{\mathcal{F}_h^D} \alpha(u_h)\eta^{(L)}\xi^{(L)} + \alpha(u_h)\big(\xi^{(L)}\big)^2 \,\mathrm{d}S \\
&\quad - \int_{\mathcal{F}_h^N} \tfrac{1}{2}\mathbf{f}'(u)\cdot\mathbf{n}\,\eta^{(L)}\xi^{(L)} + \tfrac{1}{2}\mathbf{f}'(u)\cdot\mathbf{n}\big(\xi^{(L)}\big)^2 + \tfrac{1}{8}\mathbf{f}_1''\cdot\mathbf{n}\big(e_h^{(L)}\big)^2\xi^{(L)} - \tfrac{1}{2}\mathbf{f}_2''\cdot\mathbf{n}\big(e_h^{(L)}\big)^2\xi^{(L)} \,\mathrm{d}S \\
&= -\widetilde{J}_h(\xi,\xi) - \int_{\mathcal{F}_h^I} \alpha(u_h)[\eta][\xi] \,\mathrm{d}S - \int_{\mathcal{F}_h^D} \alpha(u_h)\eta^{(L)}\xi^{(L)} \,\mathrm{d}S - \int_{\mathcal{F}_h^N} \tfrac{1}{2}\mathbf{f}'(u)\cdot\mathbf{n}\,\eta^{(L)}\xi^{(L)} \,\mathrm{d}S \\
&\quad - \int_{\mathcal{F}_h^N} \tfrac{1}{8}\mathbf{f}_1''\cdot\mathbf{n}\big(e_h^{(L)}\big)^2\xi^{(L)} - \tfrac{1}{2}\mathbf{f}_2''\cdot\mathbf{n}\big(e_h^{(L)}\big)^2\xi^{(L)} \,\mathrm{d}S.
\end{aligned}
$$

Apart from the first term, all terms in the last estimate may be treated similarly as in **(F)** and **(G)**, respectively. Thus

$$
\begin{aligned}
Y_9 + Y_{13} &\le -\widetilde{J}_h(\xi,\xi) + C\big(1 + h^{-2}\|e_h\|_\infty^2\big)\big(Ch^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2\big) + \tfrac{1}{4}\widetilde{J}_h(\xi,\xi) \\
&= -\tfrac{3}{4}\widetilde{J}_h(\xi,\xi) + C\big(1 + h^{-2}\|e_h\|_\infty^2\big)\big(Ch^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2\big).
\end{aligned}
\tag{30}
$$

As stated in **(F)**, the term $-\frac{3}{4}\widetilde{J}_h(\xi,\xi)$ in (30) is used to dominate the term $\frac{1}{4}\widetilde{J}_h(\xi,\xi)$ in (27).

The proof is completed by collecting all the estimates in **(A)**-**(H)**. □

Remark 5. *In the proof, we have treated boundary conditions as in Section 3.4.1. If we wish to define boundary conditions as in 3.4.2, taking into account Remark 3, we only need two modifications:*

(i) In (26), on $\mathcal{F}_h^N$ we estimate $|\mathbf{f}'(u)\cdot\mathbf{n}| \le |\mathbf{f}'(u_h)\cdot\mathbf{n}| + |\mathbf{f}'(u)\cdot\mathbf{n} - \mathbf{f}'(u_h)\cdot\mathbf{n}| \le \alpha(u_h) + \|e_h\|_\infty$ and proceed similarly as on $\mathcal{F}_h^D$.

(ii) We use a Taylor expansion at u_h instead of u in (29).

We can improve estimate (22), if we suppose $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$ and $\Gamma_N = \emptyset$. Namely, we obtain a factor of $h^{-1}\|e_h\|_\infty^2$ instead of $h^{-2}\|e_h\|_\infty^2$. This improved estimate is useful in proving the resulting estimates for lower order polynomials and with a less restrictive CFL condition, cf. Remarks 8 and 13.

Lemma 8. *Let $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$ and $\Gamma_N = \emptyset$. There exists a constant $C \ge 0$ independent of h, t, ε, such that*

$$b_h(u_h,\xi) - b_h(u,\xi) \le C\Big(1 + \frac{\|e_h(t)\|_\infty^2}{h}\Big)\big(h^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi\|^2\big) - \frac{1}{6}\widetilde{J}_h(\xi,\xi). \tag{31}$$

Proof. As in the proof of the preceding lemma, we estimate individual terms in (25). We note that if $\Gamma_N = \emptyset$, we only need to estimate terms Y_5, Y_8 and Y_{12}, the remaining terms already satisfy the improved estimate (31). We shall treat these suboptimal terms in more carefully than in the proof of Lemma 7:

(A) Term $\mathbf{Y_5}$: We write

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''_{u,u_h}\cdot\nabla\xi\, e_h^2\,\mathrm{d}x &= \sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''(u)\cdot\nabla\xi\,\xi^2\,\mathrm{d}x + \sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''_{u,u_h}\cdot\nabla\xi\, e_h^2 - \mathbf{f}''(u)\cdot\nabla\xi\,\xi^2\,\mathrm{d}x\\ &= \sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''(u)\cdot\nabla\xi\,\xi^2\,\mathrm{d}x + \sum_{K\in\mathcal{T}_h}\int_K \big(\mathbf{f}''_{u,u_h} - \mathbf{f}''(u)\big)\cdot\nabla\xi\,\xi^2 + 2\mathbf{f}''_{u,u_h}\cdot\nabla\xi\,\xi\eta + \mathbf{f}''_{u,u_h}\cdot\nabla\xi\,\eta^2\,\mathrm{d}x\end{aligned} \tag{32}$$

The first right-hand side integral in (32) will be estimated later along with similar terms in **(D)**.

As for the second right-hand side integral in (32), we estimate its individual summands:

- By definition of the Lagrange form of the remainder of the Taylor expansion, $\mathbf{f}''_{u,u_h}(x,t)$ has components $f_s''(\vartheta_s u(x,t) + (1-\vartheta_s)u_h(x,t))$ for some $\vartheta_s \in [0,1]$ and $s = 1,\cdots,d$. Hence, by the Lipschitz continuity of $\mathbf{f}''$, we have $|\mathbf{f}''_{u,u_h} - \mathbf{f}''(u)| \le C|u-u_h| = C|e_h|$. Therefore, by the inverse inequality and Lemma 5, (e),

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\int_K \big(\mathbf{f}''_{u,u_h} - \mathbf{f}''(u)\big)\cdot\nabla\xi\,\xi^2\,\mathrm{d}x &\le C\|e_h\|_\infty\|\xi\|_\infty C_I h^{-1}\|\xi\|^2\\ &\le C\|e_h\|_\infty(\|e_h\|_\infty + \|\eta\|_\infty)C_I h^{-1}\|\xi\|^2\\ &\le Ch^{-1}\|e_h\|_\infty^2\|\xi\|^2 + C\|e_h\|_\infty h|u(t)|_{W^{1,\infty}}h^{-1}\|\xi\|^2\\ &\le Ch^{-1}\|e_h\|_\infty^2\|\xi\|^2 + C(1+h^{-1}\|e_h\|_\infty^2)\|\xi\|^2,\end{aligned} \tag{33}$$

  since $x \le (1+h_0)(1+h^{-1}x^2)$ for all $x \ge 0$. Here we set $x := \|e_h\|_\infty$.

- Due to the inverse inequality, we have be Lemma 5, (e)

$$\sum_{K\in\mathcal{T}_h}\int_K 2\mathbf{f}''_{u,u_h}\cdot\nabla\xi\,\xi\eta\,\mathrm{d}x \le C\|\eta\|_\infty C_I h^{-1}\|\xi\|^2 \le Ch|u(t)|_{W^{1,\infty}}h^{-1}\|\xi\|^2 \le C\|\xi\|^2. \tag{34}$$

- By the inverse and Young's inequalities, Lemma 5, (e), we have

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''_{u,u_h}\cdot\nabla\xi\,\eta^2\,\mathrm{d}x \le C\|\eta\|_\infty\|\eta\|C_I h^{-1}\|\xi\|\\ &\le Ch|u(t)|_{W^{1,\infty}}h^{p+1}|u(t)|_{H^{p+1}}h^{-1}\|\xi\| \le Ch^{2p+2}|u(t)|^2_{H^{p+1}}+\|\xi\|^2.\end{aligned} \tag{35}$$

(B) Term $\mathbf{Y_8}$: We write

$$\begin{aligned}&\int_{\mathcal{F}_h^I}\mathbf{f}''_{u,\langle u_h\rangle}\cdot\mathbf{n}\langle e_h\rangle^2[\xi]\,\mathrm{d}S\\ &=\int_{\mathcal{F}_h^I}\mathbf{f}''(u)\cdot\mathbf{n}\langle\xi\rangle^2[\xi]\,\mathrm{d}S+\int_{\mathcal{F}_h^I}\mathbf{f}''_{u,\langle u_h\rangle}\cdot\mathbf{n}\langle e_h\rangle^2[\xi]-\mathbf{f}''(u)\cdot\mathbf{n}\langle\xi\rangle^2[\xi]\,\mathrm{d}S\\ &=\int_{\mathcal{F}_h^I}\mathbf{f}''(u)\cdot\mathbf{n}\langle\xi\rangle^2[\xi]\,\mathrm{d}S\\ &\quad+\int_{\mathcal{F}_h^I}\big(\mathbf{f}''_{u,\langle u_h\rangle}-\mathbf{f}''(u)\big)\cdot\mathbf{n}\langle\xi\rangle^2[\xi]+2\mathbf{f}''_{u,\langle u_h\rangle}\cdot\mathbf{n}\langle\eta\rangle\langle\xi\rangle[\xi]+\mathbf{f}''_{u,\langle u_h\rangle}\cdot\mathbf{n}\langle\eta\rangle^2[\xi]\,\mathrm{d}S\end{aligned} \tag{36}$$

The first right-hand side integral in (36) will be estimated later along with similar terms in **(D)**.

As for the second right-hand side integral in (36), its individual summands have the same structure as those in (32), thus their estimates are essentially similar:

- By the Lipschitz continuity of $\mathbf{f}''$, we have $|\mathbf{f}''_{u,\langle u_h\rangle}-\mathbf{f}''(u)|\le C|e_h|$. Therefore, by Lemma 3,

$$\int_{\mathcal{F}_h^I}\big(\mathbf{f}''_{u,\langle u_h\rangle}-\mathbf{f}''(u)\big)\cdot\mathbf{n}\langle\xi\rangle^2[\xi]\,\mathrm{d}S\le C\|e_h\|_\infty\|\xi\|_\infty h^{-1}\|\xi\|^2,$$

  which can be further estimated as in (33)

- Due to Lemmas 3 and 5, (e)

$$\int_{\mathcal{F}_h^I}2\mathbf{f}''_{u,\langle u_h\rangle}\cdot\mathbf{n}\langle\eta\rangle\langle\xi\rangle[\xi]\,\mathrm{d}S\le C\|\eta\|_\infty h^{-1}\|\xi\|^2\le C\|\xi\|^2.$$

- By Lemmas 3 and 5, (e), we have

$$\int_{\mathcal{F}_h^I}\mathbf{f}''_{u,\langle u_h\rangle}\cdot\mathbf{n}\langle\eta\rangle^2[\xi]\,\mathrm{d}S\le C\|\eta\|_\infty h^{-1}\|\eta\|\|\xi\|\le Ch^{2p+2}|u(t)|^2_{H^{p+1}}+\|\xi\|^2.$$

(C) Term $\mathbf{Y_{12}}$: For simplicity of notation, we shall remove the superscript $^{(L)}$ at ξ, e_h and η. We write

$$\begin{aligned}&\int_{\mathcal{F}_h^B}\mathbf{f}''_{u,u_h^{(L)}}\cdot\mathbf{n}e_h^2\xi\,\mathrm{d}S=\int_{\mathcal{F}_h^B}\mathbf{f}''(u)\cdot\mathbf{n}\xi^3\,\mathrm{d}S+\int_{\mathcal{F}_h^B}\mathbf{f}''_{u,u_h^{(L)}}\cdot\mathbf{n}e_h^2\xi-\mathbf{f}''(u)\cdot\mathbf{n}\xi^3\,\mathrm{d}S\\ &=\int_{\mathcal{F}_h^B}\mathbf{f}''(u)\cdot\mathbf{n}\xi^3\,\mathrm{d}S+\int_{\mathcal{F}_h^B}\big(\mathbf{f}''_{u,u_h^{(L)}}-\mathbf{f}''(u)\big)\cdot\mathbf{n}\xi^3+2\mathbf{f}''_{u,u_h^{(L)}}\cdot\mathbf{n}\eta\xi^2+\mathbf{f}''_{u,u_h^{(L)}}\cdot\mathbf{n}\eta^2\xi\,\mathrm{d}S\end{aligned} \tag{37}$$

The first right-hand side integral in (37) will be estimated later along with similar terms in **(D)**. Individual summands in the second right-hand side integral in (37) can be estimated as in **(B)**.

(D) Remaining terms from $\mathbf{Y_5}, \mathbf{Y_8}, \mathbf{Y_{12}}$: Again, we shall omit the superscript $^{(L)}$ in quantities on Γ_B. First, we have due to Green's theorem

$$\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''(u)\cdot\nabla\xi\,\xi^2\,\mathrm{d}x = \frac{1}{3}\sum_{K\in\mathcal{T}_h}\int_{\partial K}\mathbf{f}''(u)\cdot\mathbf{n}_K\xi|_K^3\,\mathrm{d}S - \frac{1}{3}\sum_{K\in\mathcal{T}_h}\int_K \operatorname{div}\big(\mathbf{f}''(u)\big)\xi^3\,\mathrm{d}x$$
$$= \frac{1}{3}\int_{\mathcal{F}_h^I}\mathbf{f}''(u)\cdot\mathbf{n}\big((\xi^{(L)})^3-(\xi^{(R)})^3\big)\,\mathrm{d}S + \frac{1}{3}\int_{\mathcal{F}_h^B}\mathbf{f}''(u)\cdot\mathbf{n}\xi^3\,\mathrm{d}S - \frac{1}{3}\sum_{K\in\mathcal{T}_h}\int_K \operatorname{div}\big(\mathbf{f}''(u)\big)\xi^3\,\mathrm{d}x.$$

Thus, we may estimate the remaining terms of Y_5, Y_5, Y_{12} as

$$-\frac{1}{2}\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}''(u)\cdot\nabla\xi\,\xi^2\,\mathrm{d}x + \frac{1}{2}\int_{\mathcal{F}_h^I}\mathbf{f}''(u)\cdot\mathbf{n}\langle\xi\rangle^2[\xi]\,\mathrm{d}S + \frac{1}{8}\int_{\mathcal{F}_h^B}\mathbf{f}''(u)\cdot\mathbf{n}\,\xi^3\,\mathrm{d}S$$
$$= \underbrace{\int_{\mathcal{F}_h^I}\mathbf{f}''(u)\cdot\mathbf{n}\Big(-\tfrac{1}{6}\big((\xi^{(L)})^3-(\xi^{(R)})^3\big)+\tfrac{1}{2}\langle\xi\rangle^2[\xi]\Big)\,\mathrm{d}S}_{Z_1}$$
$$+\underbrace{\int_{\mathcal{F}_h^B}\mathbf{f}''(u)\cdot\mathbf{n}\big(-\tfrac{1}{6}\xi^3+\tfrac{1}{8}\xi^3\big)\,\mathrm{d}S}_{Z_2} + \underbrace{\frac{1}{6}\sum_{K\in\mathcal{T}_h}\int_K \operatorname{div}\big(\mathbf{f}''(u)\big)\xi^3\,\mathrm{d}x}_{Z_3}.$$

- $\mathbf{Z_1}$: Using the identity $-\frac{1}{6}\big((\xi^{(L)})^3-(\xi^{(R)})^3\big)+\frac{1}{2}\langle\xi\rangle^2[\xi] = -\frac{1}{24}[\xi]^3$ and the Taylor expansion we have

$$\begin{aligned}-\mathbf{f}''(u)\cdot\mathbf{n}[\xi] &= -\mathbf{f}''\big(\langle u_h\rangle\big)\cdot\mathbf{n}[\xi] - \mathbf{f}'''_{u,\langle u_h\rangle}\cdot\mathbf{n}\big(u-\langle u_h\rangle\big)[\xi]\\ &= \mathbf{f}''\big(\langle u_h\rangle\big)\cdot\mathbf{n}[\eta] + \mathbf{f}''\big(\langle u_h\rangle\big)\cdot\mathbf{n}[u_h] - \mathbf{f}'''_{u,\langle u_h\rangle}\cdot\mathbf{n}\big(\langle u-u_h\rangle\big)[\xi].\end{aligned}$$

  Therefore, due to (17) and Lemma 5, d), we have

$$\begin{aligned}|\mathbf{f}''(u)\cdot\mathbf{n}[\xi]| &\le \big|\mathbf{f}''\big(\langle u_h\rangle\big)\big|\,2\|\eta\|_\infty + 8\alpha(u_h) + C[u_h]^2 + \big|\mathbf{f}'''_{u,\langle u_h\rangle}\big|\|e_h\|_\infty\big|[e_h]-[\eta]\big|\\ &\le C\|\eta\|_\infty + 8\alpha(u_h) + C[e_h]^2 + C\|e_h\|_\infty\big(2\|e_h\|_\infty + 2\|\eta\|_\infty\big)\\ &\le Ch^p + 8\alpha(u_h) + C\|e_h\|_\infty^2.\end{aligned}$$

  Therefore, we may conclude using Lemma 3

$$|Z_1| \le \frac{1}{3}\int_{\mathcal{F}_h^I}\alpha(u_h)[\xi]^2\,\mathrm{d}S + \big(Ch^{p-1}+Ch^{-1}\|e_h\|_\infty^2\big)\|\xi\|^2$$

- $\mathbf{Z_2}$: We use the identity $-\frac{1}{6}\xi^3+\frac{1}{8}\xi^3 = -\frac{1}{24}\xi^3$ and by the Taylor expansion we have

$$\begin{aligned}-\mathbf{f}''(u)\cdot\mathbf{n}\xi &= -\mathbf{f}''\big(\tfrac{1}{2}(u+u_h)\big)\cdot\mathbf{n}\,\xi - \mathbf{f}_1'''\cdot\mathbf{n}\big(u-\tfrac{1}{2}(u+u_h)\big)\xi\\ &= \mathbf{f}''\big(\tfrac{1}{2}(u+u_h)\big)\cdot\mathbf{n}\,\eta + \mathbf{f}''\big(\tfrac{1}{2}(u+u_h)\big)\cdot\mathbf{n}(u-u_h) - \mathbf{f}_1'''\cdot\mathbf{n}\tfrac{1}{2}e_h\xi.\end{aligned}$$

  Therefore

$$Z_2 = \frac{1}{24}\int_{\mathcal{F}_h^B}\mathbf{f}''\big(\tfrac{1}{2}(u+u_h)\big)\cdot\mathbf{n}\,\eta\xi^2\,\mathrm{d}S + \frac{1}{24}\int_{\mathcal{F}_h^B}\mathbf{f}''\big(\tfrac{1}{2}(u+u_h)\big)\cdot\mathbf{n}\,(u-u_h)\xi^2\,\mathrm{d}S$$
$$-\frac{1}{48}\int_{\mathcal{F}_h^B}\mathbf{f}_1'''\cdot\mathbf{n}e_h\xi^3\,\mathrm{d}S = Z_2^{(a)}+Z_2^{(b)}+Z_2^{(c)}.$$

We estimate using Lemmas 3 and 5, d)

$$|Z_2^{(a)}| \le C\|\eta\|_\infty \int_{\mathcal{F}_h^B} |\xi|^2 \,\mathrm{d}S \le Ch^{p-1}\|\xi\|^2 \le C\|\xi\|^2.$$

Now, we estimate using (17) and Lemma 3

$$|Z_2^{(b)}| \le \frac{1}{24}\int_{\mathcal{F}_h^B} 8\alpha(u_h)\xi^2 + C(u_h-u)^2\xi^2 \,\mathrm{d}S \le \frac{1}{3}\int_{\mathcal{F}_h^B} \alpha(u_h)\xi^2 \,\mathrm{d}S + Ch^{-1}\|e_h\|_\infty^2\|\xi\|^2.$$

Finally, due to Lemma 3

$$\begin{aligned}
&|Z_2^{(c)}| \le C\int_{\mathcal{F}_h^B} |e_h||\xi|^3 \,\mathrm{d}S \le C\|e_h\|_\infty\|\xi\|_\infty h^{-1}\|\xi\|^2 \\
&\le Ch^{-1}\|e_h\|_\infty^2\|\xi\|^2 + Ch^{-1}\|e_h\|_\infty\|\eta\|_\infty\|\xi\|^2 \le Ch^{-1}\|e_h\|_\infty^2\|\xi\|^2 + Ch^{p-1}\|e_h\|_\infty\|\xi\|^2 \\
&\le Ch^{-1}\|e_h\|_\infty^2\|\xi\|^2 + C(1+h^{-1}\|e_h\|_\infty^2)\|\xi\|^2,
\end{aligned}$$

where we have used the fact that $x \le (1+h_0)(1+h^{-1}x^2)$ for all $x \ge 0$. Here we set $x := \|e_h\|_\infty$. This completes the estimate of Z_2.

- $\mathbf{Z_3}$: We estimate

$$\begin{aligned}
|Z_3| &\le C\|\xi\|_\infty\|\xi\|^2 \le (C+\|\xi\|_\infty^2)\|\xi\|^2 \\
&\le (C+2\|e_h\|_\infty^2 + 2\|\eta\|_\infty^2)\|\xi\|^2 \le C(1+h^{-1}\|e_h\|_\infty^2)\|\xi\|^2.
\end{aligned}$$

The proof is completed by collecting all the estimates in **(A)-(D)**. □

6 Further properties of the convection and diffusion forms

Let us define the bilinear form

$$A_h(v,w) := a_h(v,w) + J_h(v,w), \quad \forall v,w \in H^2(cT_h),$$

and the following (energy) norm in $H^1(\mathcal{T}_h)$:

$$\|w\|_{DG} = \Big(\frac{1}{2}\big(|w|_{H^1(\mathcal{T}_h)}^2 + J_h(w,w)\big)\Big)^{1/2}.$$

Lemma 9 (Ellipticity and boundedness of A_h)**.** *Let the constant C_W from (9) satisfy*

$$C_W \begin{cases} \ge 4C_M(1+C_I) & \text{for } \Theta = 1, \text{ i.e. the symmetric variant,} \\ \ge 2C_M(1+C_I) & \text{for } \Theta = 0, \text{ i.e. the incomplete variant,} \\ > 0 & \text{for } \Theta = -1, \text{ i.e. the nonsymmetric variant,} \end{cases} \tag{38}$$

where C_M, C_I are constants from Lemmas 3 and 4, respectively. Then the form A_h is elliptic*, i.e.*

$$\|v\|_{DG}^2 \le A_h(v,v), \quad \forall v \in S_h,$$

and bounded*, i.e.*

$$A_h(v,w) \le \|v\|_{DG}\|w\|_{DG}, \quad \forall v,w \in S_h.$$

Moreover, we have the estimate

$$A_h\big(\eta_h(t),\xi_h(t)\big) \le Ch^p|u(t)|_{H^{p+1}}\|\xi_h(t)\|_{DG}.$$

Proof. Cf. [3] for an overview and [22] for a thorough discussion. □

Remark 6. *Using Lemmas 3 and 4 it is straightforward to prove that for all* $v_h \in S_h$

$$\|v_h\|_{DG} \le Ch^{-1}\|v_h\|.$$

Lemma 10 (Consistency error of b_h)**.** *There exists a constant* $L_{b_h} > 0$ *independent of* u, u_h, h, ε*, such that for all* $v, \bar{v}, w \in S_h$*.*

$$\begin{aligned}&|b_h(v,w) - b_h(\bar{v},w)|\\ &\quad\le L_{b_h}\Big(\|v-\bar{v}\|\|w\|_{DG} + \big(\|v-\bar{v}\|\|v-\bar{v}\|_{DG} + \|v-\bar{v}\|^2\big)^{1/2}\big(\|w\|\|w\|_{DG} + \|w\|^2\big)^{1/2}\Big).\end{aligned}$$

Proof. By the definition of b_h, we have

$$b_h(v,\varphi) = -\sum_{K\in\mathcal{T}_h}\int_K \mathbf{f}(v)\cdot\nabla\varphi\,\mathrm{d}x + \int_{\mathcal{F}_h^I\cup\mathcal{F}_h^D} H(v^{(L)}, v^{(R)}, \mathbf{n})[\varphi]\,\mathrm{d}S + \int_{\mathcal{F}_h^N}\mathbf{f}(v^{(L)})\cdot\mathbf{n}\,\varphi^{(L)}\,\mathrm{d}S.$$

The first two integrals may be treated directly using the multiplicative trace inequality as in [20]. As for the integral over $\mathcal{F}_h^N$, we estimate

$$\Big|\int_{\mathcal{F}_h^N}\big(\mathbf{f}(v^{(L)})\cdot\mathbf{n} - \mathbf{f}(\bar{v}^{(L)})\cdot\mathbf{n}\big)\,w^{(L)}\,\mathrm{d}S\Big| \le C\|v-\bar{v}\|_{L^2(\Gamma_N)}\|w\|_{L^2(\Gamma_N)}. \tag{39}$$

These terms cannot be estimated from above directly, since $\|\cdot\|_{DG}$ contains only integrals over Γ_D. We must proceed as in [17] with the use of the *global multiplicative trace inequality*, which, for our purposes, we write in the form

$$\|v\|^2_{L^2(\partial\Omega)} \le C'_M\big(\|v\|_{DG}\|v\| + \|v\|^2\big), \quad \forall v \in S_h, \tag{40}$$

where C'_M is a constant independent of h. Applying (40) to (39) gives the desired result. □

7 Error analysis for the method of lines

We proceed in a standard way. We subtract (13) from (14) and set $\varphi_h := \xi_h(t) \in S_h$. Since

$$\Big(\frac{\partial\xi_h(t)}{\partial t}, \xi_h(t)\Big) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi_h(t)\|^2,$$

we get

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi_h(t)\|^2 + \varepsilon A_h\big(\xi_h(t),\xi_h(t)\big)\\ &\qquad= -\varepsilon A_h\big(\eta_h(t),\xi_h(t)\big) + b_h\big(u_h(t),\xi_h(t)\big) - b_h\big(u(t),\xi_h(t)\big) - \Big(\frac{\partial\eta_h(t)}{\partial t}, \xi_h(t)\Big).\end{aligned} \tag{41}$$

For the last right-hand side term, by the Cauchy and Young's inequalities and Lemma 5, we have

$$|(\eta_t,\xi)| \le \|\eta_t\|\,\|\xi\| \le \tfrac{1}{2}\left(\|\eta_t\|^2 + \|\xi\|^2\right) \le \tfrac{1}{2}\left(C\,h^{2(p+1)}|u_t(t)|^2_{H^{p+1}} + \|\xi\|^2\right).$$

As for the right-hand side diffusion terms, we use Lemma 9 and Young's inequality:

$$A_h(\eta,\xi) \le C\,h^{2p}|u(t)|^2_{H^{p+1}} + \tfrac{1}{2}\|\xi\|^2_{DG}.$$

Finally, we use the ellipticity of A_h and Lemma 7 to obtain from (41)

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\|\xi_h(t)\|^2+\varepsilon\|\xi_h(t)\|_{DG}^2+\widetilde{J}_h\big(\xi_h(t),\xi_h(t)\big)\\&\quad\le C\Big(1+\frac{\|e_h(t)\|_\infty^2}{h^2}\Big)\big(h^{2p+1}|u(t)|_{H^{p+1}}^2+\varepsilon h^{2p}|u(t)|_{H^{p+1}}^2+h^{2p+2}|u_t(t)|_{H^{p+1}}^2+\|\xi_h(t)\|^2\big).\end{aligned}$$

Integrating from 0 to $t\in[0,T]$ yields

$$\begin{aligned}&\|\xi_h(t)\|^2+\int_0^t\varepsilon\|\xi_h(\vartheta)\|_{DG}^2+\widetilde{J}_h\big(\xi_h(\vartheta),\xi_h(\vartheta)\big)\,\mathrm{d}\vartheta\\&\le C\int_0^t\Big(1+\frac{\|e_h(\vartheta)\|_\infty^2}{h^2}\Big)\Big(\big(h^{2p+1}+\varepsilon h^{2p}\big)|u(\vartheta)|_{H^{p+1}}^2+h^{2p+2}|u_t(\vartheta)|_{H^{p+1}}^2+\|\xi_h(\vartheta)\|^2\Big)\,\mathrm{d}\vartheta,\end{aligned}\tag{42}$$

where the constant $C\ge 0$ is independent of h,t,ε. For simplicity, we assume that $\xi_h(0)=0$, i.e. $u_h^0=\Pi_h u^0$. Otherwise we assume e.g. $\|\xi_h(0)\|^2\le Ch^{2p+1}|u^0|_{H^{p+1}}^2$ and include this term in estimate (42).

7.1 Estimates based on continuous mathematical induction

We notice that if we knew *apriori* that $\|e_h\|_\infty=O(h)$ then the unpleasant term $\|e_h\|_\infty^2h^{-2}$ in (42) would be $O(1)$. Thus we could simply apply the standard Gronwall inequality to obtain the desired error estimates. We state this formally:

Lemma 11. *Let $t\in[0,T]$ and $p\ge d/2$. If*

$$\|e_h(\vartheta)\|\le h^{1+d/2},\quad\forall\vartheta\in[0,t],\tag{43}$$

then we have the estimate

$$\max_{\vartheta\in[0,t]}\|e_h(\vartheta)\|^2+\int_0^t\varepsilon\|e_h(\vartheta)\|_{DG}^2+\widetilde{J}_h\big(e_h(\vartheta),e_h(\vartheta)\big)\,\mathrm{d}\vartheta\le C_T^2\big(h^{2p+1}+\varepsilon h^{2p}\big),\tag{44}$$

where C_T is a constant independent of h,t and ε.

Proof. Due to the inverse inequality, Lemma 5 and assumption (43) we have

$$\begin{aligned}\|e_h(\vartheta)\|_\infty&\le\|\eta_h(\vartheta)\|_\infty+\|\xi_h(\vartheta)\|_\infty\le Ch|u(\vartheta)|_{W^{1,\infty}}+C_Ih^{-d/2}\|\xi_h(\vartheta)\|\\&\le Ch+C_Ih^{-d/2}\|e_h(\vartheta)\|+C_Ih^{-d/2}\|\eta_h(\vartheta)\|\le Ch+Ch^{p+1-d/2}|u(\vartheta)|_{H^{p+1}(\mathcal{T}_h)}\le Ch,\end{aligned}\tag{45}$$

where the constant C is independent of h,ϑ,ε. If we use this estimate in (42), we obtain

$$\begin{aligned}&\|\xi_h(t)\|^2+\int_0^t\varepsilon\|\xi_h(\vartheta)\|_{DG}^2+\widetilde{J}_h\big(\xi_h(\vartheta),\xi_h(\vartheta)\big)\,\mathrm{d}\vartheta\\&\quad\le C\Big(\big(h^{2p+1}+\varepsilon h^{2p}\big)|u|_{L^2(H^{p+1})}^2+h^{2p+2}|u_t|_{L^2(H^{p+1})}^2\Big)+C\int_0^t\|\xi_h(\vartheta)\|^2\,\mathrm{d}\vartheta\\&\quad\le\tilde{C}\big(h^{2p+1}+\varepsilon h^{2p}\big)+C\int_0^t\|\xi_h(\vartheta)\|^2\,\mathrm{d}\vartheta,\end{aligned}\tag{46}$$

where the constants $\tilde{C},C$ are independent of h,t,ε. Gronwall's inequality applied to (46) states that there exists a constant $\widetilde{C}_T$, independent of h,t,ε, but depending exponentially on T, such that

$$\max_{\vartheta\in[0,t]}\|\xi_h(\vartheta)\|^2+\int_0^t\varepsilon\|\xi_h(\vartheta)\|_{DG}^2+\widetilde{J}_h\big(\xi_h(t),\xi_h(t)\big)\,\mathrm{d}\vartheta\le\widetilde{C}_T\big(h^{2p+1}+\varepsilon h^{2p}\big).$$

By Lemma 5, we have the estimate

$$\max_{\vartheta\in[0,t]} \|\eta_h(\vartheta)\|^2 + \int_0^t \varepsilon\|\eta_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(\eta_h(t),\eta_h(t)\big)\,\mathrm{d}\vartheta \le \bar{C}_T\big(h^{2p+1}+\varepsilon h^{2p}\big). \tag{47}$$

Therefore, by the triangle inequality we obtain estimate (44) with some constant C_T. □

Now it remains to get rid of the *apriori* assumption $\|e_h\|_\infty = O(h)$. In [44] this is done for an explicit scheme using mathematical induction. If e_h^n denotes the error of the explicit DG scheme at the n-th time node t_n, we get for the initial condition at least $\|e_h^0\| = O(h^{p+1/2})$. Then it is easy to prove the following induction step:

$$\|e_h^n\| = O(h^{p+1/2}) \quad\Longrightarrow\quad \|e_h^{n+1}\|_\infty = O(h) \quad\Longrightarrow\quad \|e_h^{n+1}\| = O(h^{p+1/2}). \tag{48}$$

For the method of lines we have no discrete structure with respect to time and hence cannot use mathematical induction straightforwardly. However, we can divide $[0,T]$ into a finite number of sufficiently small intervals $[t_n,t_{n+1}]$ on which "e_h *does not change too much*" and use induction with respect to n. This is close to the philosophy of the so-called *continuous mathematical induction* introduced in [11]. This states that if $\varphi(t)$ is a propositional function depending on $t\in[0,T]$ such that

(*i*) $\varphi(0)$ is true,

(*ii*) $\exists\delta_0>0:\ \varphi(t)$ implies $\varphi(t+\delta),\ \forall t\in[0,T]\ \forall\delta\in[0,\delta_0] : t+\delta\in[0,T]$.

Then $\varphi(t)$ holds for all $t\in[0,T]$. Due to the simple nature of this concept, we shall not formulate the proof of the main theorem as a continuous mathematical induction argument, but rather prove it directly using the aforementioned partition of $[0,T]$ and continuity of e_h with respect to time.

Remark 7. *Due to the regularity assumptions, the functions $u(\cdot), u_h(\cdot)$ are continuous mappings from $[0,T]$ to $L^2(\Omega)$. Since $[0,T]$ is a compact set, $e_h(\cdot)$ is a* uniformly continuous *function from $[0,T]$ to $L^2(\Omega)$. By definition,*

$$\forall\bar\epsilon>0\ \exists\delta>0:\ s,\bar s\in[0,T], |s-\bar s|\le\delta \quad\Longrightarrow\quad \|e_h(s)-e_h(\bar s)\|\le\bar\epsilon.$$

Theorem 12 (Main theorem)**.** *Let $p>1+d/2$. Let $h_1>0$ be such that $C_T(h_1^{p+1/2}+\sqrt{\varepsilon}h_1^p)=\frac{1}{2}h_1^{1+d/2}$, where C_T is the constant from Lemma 11. Then for all $h\in(0,h_1]$ we have the estimate*

$$\max_{\vartheta\in[0,T]} \|e_h(\vartheta)\|^2 + \int_0^T \varepsilon\|e_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(e_h(\vartheta),e_h(\vartheta)\big)\,\mathrm{d}\vartheta \le C_T^2\big(h^{2p+1}+\varepsilon h^{2p}\big). \tag{49}$$

Proof. We have $p>1+d/2$, therefore h_1 is uniquely determined and $C_T(h^{p+1/2}+\sqrt{\varepsilon}h^p)\le\frac{1}{2}h^{1+d/2}$ for all $h\in(0,h_1]$. We fix an arbitrary $h\in(0,h_1]$. By Remark 7, there exists $\delta>0$, such that if $s,\bar s\in[0,T], |s-\bar s|\le\delta$, then $\|e_h(s)-e_h(\bar s)\|\le\frac{1}{2}h^{1+d/2}$.

We define $t_i=i\delta$, $i=0,1,\dots$ and set $N:=\max\{i=0,1,\dots;t_i<T\}$, $t_{N+1}:=T$. This defines a partition $0=t_0<t_1<\cdots<t_{N+1}=T$ of the interval $[0,T]$ into $N+1$ intervals of length (at most) δ.

We shall now prove by induction that for all $n=1,\dots,N+1$

$$\max_{\vartheta\in[0,t_n]} \|e_h(\vartheta)\|^2 + \int_0^{t_n} \varepsilon\|e_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(e_h(\vartheta),e_h(\vartheta)\big)\,\mathrm{d}\vartheta \le C_T^2\big(h^{2p+1}+\varepsilon h^{2p}\big). \tag{50}$$

Inequality (49) is thus obtained by taking $n := N + 1$ in (50).
(i) $n = 1$: We know that $\|e_h(0)\| = \|\eta_h(0)\| \le C_T(h^{p+1/2} + \sqrt{\varepsilon}h^p) \le \frac{1}{2}h^{1+d/2}$. By uniform continuity, we have for all $s \in [0, t_1]$

$$\|e_h(s)\| \le \|e_h(0)\| + \|e_h(s) - e_h(0)\| \le \tfrac{1}{2}h^{1+d/2} + \tfrac{1}{2}h^{1+d/2} = h^{1+d/2}.$$

Therefore, by Lemma 11 we obtain estimate (50) on $[0, t_1]$, i.e. for $n = 1$.
(ii) Induction step: We assume that (50) holds for a general $n < N + 1$. Therefore $\|e_h(t_n)\| \le C_T(h^{p+1/2} + \sqrt{\varepsilon}h^p) \le \frac{1}{2}h^{1+d/2}$. By uniform continuity, we have that for all $s \in [t_n, t_{n+1}]$

$$\|e_h(s)\| \le \|e_h(t_n)\| + \|e_h(s) - e_h(t_n)\| \le \tfrac{1}{2}h^{1+d/2} + \tfrac{1}{2}h^{1+d/2} = h^{1+d/2}.$$

This and the induction assumption imply that $\|e_h(s)\| \le h^{1+d/2}$ for $s \in [0, t_n] \cup [t_n, t_{n+1}] = [0, t_{n+1}]$. Therefore, by Lemma 11, we obtain estimate (50) on $[0, t_{n+1}]$, i.e. for '$n := n + 1$'. □

Remark 8. *If we assume* $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$ *and* $\Gamma_N = \emptyset$*, then we may use the improved estimate of Lemma 8 which gives the more favorable factor* h^{-1} *instead of* h^{-2} *in the estimate of the convective terms. Hence in Theorem 12 we get the improved assumption* $p > (1 + d)/2$*. Furthermore, if* $\varepsilon = 0$ *we need to assume only* $p + 1/2 > (1 + d)/2$*, i.e.* $p > d/2$*.*

Remark 9. *The derived estimates are suboptimal in the* $L^\infty(L^2)$*-norm, where we would expect an* $O(h^{p+1})$ *convergence rate, however this estimate is valid for all* $\varepsilon \ge 0$*. For* $\varepsilon > 0$ *also a DG energy norm is included in the estimates, which is estimated optimally. However, this estimate degenerates for* $\varepsilon \to 0$ *and does not hold for* $\varepsilon = 0$*. For this reason, we view the* $L^\infty(L^2)$*-estimate as primary, even though it is suboptimal.*

7.2 Estimates based on a nonlinear Gronwall's inequality

For the method of lines we can use a more direct approach to prove Theorem 12 than in Section 7.1, an appropriate *nonlinear Gronwall-type lemma*. As we prove in Lemma 14, this is not possible in the case of an implicit scheme.

Lemma 13 (Nonlinear Gronwall's inequality)**.** *Let* $A(t) \ge 0$*, for all* $t \in [0, T]$*, and* $\alpha, \beta_1, \beta_2, T > 0$ *are constants. Let* $u \in C\big([0, T]; [0, \infty)\big)$ *such that*

$$0 \le u(t) + A(t) \le \alpha + \int_0^t \big(\beta_1 u(\vartheta) + \beta_2 u^2(\vartheta)\big)\,\mathrm{d}\vartheta, \quad \forall t \in [0, T].$$

If the coefficients satisfy

$$2\alpha\beta_2 e^{\beta_1 T} \le \beta_1, \tag{51}$$

then

$$u(t) + A(t) \le 2\alpha e^{\beta_1 t}, \quad \forall t \in [0, T]. \tag{52}$$

Proof. First, we assume that $u(t) > 0$. Defining $F(t) := \alpha + \int_0^t \big(\beta_1 u(\vartheta) + \beta_2 u^2(\vartheta)\big)\,\mathrm{d}\vartheta$, we have $0 < \alpha \le F$, $F \in C^1([0, T], R)$ and F is strictly increasing, since $F' > 0$.

Since $u(t) \le u(t) + A(t) \le F(t)$ for all $t \in [0, T]$, we obtain by integration

$$F'(t) \le \beta_1 u(t) + \beta_2 u^2(t) \le \beta_1 F(t) + \beta_2 F^2(t) \quad \Rightarrow \quad \int_0^t \frac{F'(\vartheta)}{F(\vartheta) + \gamma F^2(\vartheta)}\,\mathrm{d}\vartheta \le \int_0^t \beta_1\,\mathrm{d}\vartheta,$$

where $\gamma = \frac{\beta_2}{\beta_1}$. Since $F' > 0$, F is a bijection and we can evaluate the left-hand side integral by substitution, while the right-hand side is simply $\beta_1 t$. Thus

$$\ln \frac{F(t)(1+\gamma F(0))}{F(0)(1+\gamma F(t))} \le \beta_1 t \quad \Rightarrow \quad \frac{F(t)}{1+\gamma F(t)} \le \frac{F(0)}{1+\gamma F(0)} e^{\beta_1 t} \le \alpha e^{\beta_1 t}.$$

From this inequality, we may express $F(t)$, therefore

$$u(t) + A(t) \le F(t) \le \frac{1}{1-\gamma\alpha e^{\beta_1 t}} \alpha e^{\beta_1 t} \le 2\alpha e^{\beta_1 t},$$

due to condition (51). Thus we have proven (52) for the case $u(t) > 0$.

Generally, if $u(t) \ge 0$, we define $\tilde{u}(t) := u(t) + \delta > 0$ for some small $\delta > 0$. Then $\tilde{u}(t)$ satisfies

$$0 < \tilde{u}(t) + A(t) \le (\alpha + \delta) + \int_0^t \big(\beta_1 \tilde{u}(\vartheta) + \beta_2 \tilde{u}^2(\vartheta)\big)\,\mathrm{d}\vartheta, \quad \forall t \in [0,T].$$

Therefore, by (52)

$$u(t) + A(t) \le \tilde{u}(t) + A(t) \le 2(\alpha+\delta) e^{\beta_1 t},$$

which holds for all sufficiently small $\delta > 0$. By taking $\delta \to 0$, we obtain (52). □

Alternative proof of Theorem 12. Due to the inverse inequality and Lemma 5, we have

$$\begin{aligned} h^{-2}\|e_h(\vartheta)\|_\infty^2 &\le 2h^{-2}\|\eta_h(\vartheta)\|_\infty^2 + 2h^{-2}\|\xi_h(\vartheta)\|_\infty^2 \le \\ &Ch^{-2}h^2|u(\vartheta)|_{W^{1,\infty}}^2 + C_I h^{-2-d}\|\xi_h(\vartheta)\|^2 \le C + Ch^{-2-d}\|\xi_h(\vartheta)\|^2, \end{aligned}$$

where the constant C is independent of $h, \vartheta, t, \varepsilon$. If we use this estimate in (42), we obtain

$$\begin{aligned} &\|\xi_h(t)\|^2 + \int_0^t \varepsilon\|\xi_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(\xi_h(\vartheta), \xi_h(\vartheta)\big)\,\mathrm{d}\vartheta \qquad (53)\\ &\le \tilde{C}\big(h^{2p+1} + \varepsilon h^{2p}\big) + C\int_0^t \big(1 + h^{2p-1-d} + \varepsilon h^{2p-2-d}\big)\|\xi_h(\vartheta)\|^2 + h^{-2-d}\|\xi_h(\vartheta)\|^4\,\mathrm{d}\vartheta, \end{aligned}$$

where the constants $\tilde{C}, C$ are independent of h, t, ε. Now, we shall apply Lemma 13 to (53) by setting

$$\begin{aligned} u(t) &:= \|\xi_h(t)\|^2, \\ A(t) &:= \int_0^t \varepsilon\|\xi_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(\xi_h(\vartheta), \xi_h(\vartheta)\big)\,\mathrm{d}\vartheta, \\ \alpha &:= \tilde{C}\big(h^{2p+1} + \varepsilon h^{2p}\big), \\ \beta_1 &:= C\big(1 + h^{2p-1-d} + \varepsilon h^{2p-2-d}\big) \le \bar{C}, \\ \beta_2 &:= Ch^{-2-d}. \end{aligned}$$

Condition (51) can be written as

$$2\tilde{C}\big(h^{2p+1} + \varepsilon h^{2p}\big)Ch^{-2-d}e^{\bar{C}T} \le \bar{C} \quad \Longleftrightarrow \quad h^{2p-1-d} + \varepsilon h^{2p-2-d} \le \tfrac{1}{2}\tilde{C}^{-1}C^{-1}\bar{C}e^{-\bar{C}T},$$

which is satisfied for sufficiently small h if $p > 1 + d/2$. By (52), we have

$$\|\xi_h(t)\|^2 + \int_0^t \varepsilon\|\xi_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(\xi_h(\vartheta), \xi_h(\vartheta)\big)\,\mathrm{d}\vartheta \le \tilde{C}_T\big(h^{2p+1} + \varepsilon h^{2p}\big),$$

for some $\tilde{C}_T$ independent of h, t, ε. The proof is completed by taking (47) and the triangle inequality.

8 Error estimates for a fully implicit scheme

In this section, we shall introduce and analyze the DG scheme with a standard first order implicit time discretization. Here we cannot use the approach of [44] for the explicit scheme, since in Lemma 14 we prove that the first implication in the induction step (48) does not hold. Essentially the error equation along with the estimates of individual terms does not contain sufficient information about the problem to prove the desired estimates.

In Section 7.1 the key ingredient was the continuity of e_h with respect to time, which guarantees that the error cannot suddenly blow up. We then use a continuous mathematical induction argument. However, for the implicit scheme we have a discrete temporal structure, hence no continuity. To overcome this obstacle, we introduce an appropriate continuation of the discrete solution and error with respect to time. This is constructed using an auxiliary problem, essentially a modification of the discrete implicit problem. This allows us to derive error estimates for the continuated solution and consequently for the original implicit scheme

We consider a partition $0 = t_0 < t_1 < \cdots < t_{N+1} = T$ of the time interval $[0, T]$ and set $\tau_n = t_{n+1} - t_n$ for $n = 0, \cdots, N$. The exact solution $u(t_n)$ will be approximated by $u_h^n \in S_h$.

Definition 4. *We say that $\{u_h^n\}_{n=0}^{N+1} \subset S_h$ is an implicit DGFE solution of the convection-diffusion problem (2) - (5), if $u_h^0 = \Pi_h u^0 \in S_h$ and for $n = 0, \cdots, N$*

$$\Big(\frac{u_h^{n+1} - u_h^n}{\tau_n}, \varphi_h\Big) + b_h\big(u_h^{n+1}, \varphi_h\big) + \varepsilon A_h\big(u_h^{n+1}, \varphi_h\big) = l_h\big(\varphi_h\big)(t_{n+1}), \quad \forall \varphi_h \in S_h. \tag{54}$$

Similarly as in Section 4, we define $\eta_h^n = u(t_n) - \Pi_h u(t_n) \in H^{p+1}(\mathcal{T}_h)$ and $\xi_h^n = \Pi_h u(t_n) - u_h^n \in S_h$. Then we can write the error e_h^n as $e_h^n := u(t_n) - u_h^n = \eta_h^n + \xi_h^n$.

To obtain error estimates for the implicit scheme, we would now subtract the equation for the exact and approximate solution

$$\begin{aligned}\big(e_h^{n+1} - e_h^n, \varphi_h\big) + \tau_n\big(b_h(u(t_{n+1}), \varphi_h) - b_h(u_h^{n+1}, \varphi_h)\big) + \tau_n \varepsilon A_h(e_h^{n+1}, \varphi_h)\\ = \big(u(t_{n+1}) - u(t_n) - \tau_n u_t(t_{n+1}), \varphi_h\big).\end{aligned}$$

In standard approaches, we set $\varphi_h := \xi_h^{n+1}$ and apply the derived estimates of b_h, A_h and Lemma 17:

$$\begin{aligned}&\|\xi_h^{n+1}\|^2 + \|\xi_h^{n+1} - \xi_h^n\|^2 + \tau_n \varepsilon \|\xi_h^{n+1}\|_{DG}^2 + \tau_n \widetilde{J}_h\big(\xi_h^{n+1}, \xi_h^{n+1}\big) \qquad (55)\\ &\le \|\xi_h^n\|^2 + C\tau_n\Big(1 + \frac{\|e_h^{n+1}\|_\infty^2}{h^2}\Big)\big(h^{2p+1} + \|\xi_h^{n+1}\|^2\big) + C\tau_n\big(\varepsilon h^{2p} + h^{2p+2} + \tau_n^2 + \|\xi_h^{n+1}\|^2\big).\end{aligned}$$

At this point, we would apply induction or some (nonlinear) discrete Gronwall lemma to obtain the desired error estimates as in Section 7.2. However, it is simple to see that (55) does not imply the desired error estimates without any additional assumptions and we need to proceed more carefully in our analysis.

Lemma 14. *Inequalities (55) taken for $n = 0, \cdots, N$, do not imply the desired error estimate (73).*

Proof. We take (55) on the first time level $n = 0$:

$$\begin{aligned}&\|\xi_h^1\|^2 + \|\xi_h^1 - \xi_h^0\|^2 + \tau_0 \varepsilon \|\xi_h^1\|_{DG}^2 + \tau_0 \widetilde{J}_h\big(\xi_h^1, \xi_h^1\big) \qquad (56)\\ &\le \|\xi_h^0\|^2 + C\tau_0\Big(1 + \frac{\|e_h^1\|_\infty^2}{h^2}\Big)\big(h^{2p+1} + \|\xi_h^1\|^2\big) + C\tau_0\big(\varepsilon h^{2p} + h^{2p+2} + \tau_0^2 + \|\xi_h^1\|^2\big).\end{aligned}$$

Let us denote $X := \|e_h^1\|$. We shall prove that for all h, the right-hand side of (56) "*grows faster with respect to* X" than the left-hand side as $X \to \infty$. Therefore (56) is satisfied not only for small $X = O(h^{p+1/2} + \sqrt{\varepsilon}h^p + \tau)$, but also for arbitrary X sufficiently large. Hence, (56) does not imply the error estimate (73) even for $n = 0$.

We proceed as follows: let X be sufficiently large. Then we have due to Lemmas 3, 4 and 5

$$\begin{aligned}
\|\xi_h^1\| &\le \|e_h^1\| + \|\eta_h^1\| \le X + Ch^{p+1} \le 2X, \quad \text{for } X \text{ sufficiently large,}\\
\|\xi_h^1\| &\ge \|e_h^1\| - \|\eta_h^1\| \ge X - Ch^{p+1} \ge X/2, \quad \text{for } X \text{ sufficiently large,}\\
\|\xi_h^1 - \xi_h^0\| &\le \|\xi_h^1\| + \|\xi_h^0\| \le 2X + Ch^{p+1} \le 3X,\\
\|\xi_h^1\|_{DG} &\le C_1 h^{-1}\|\xi_h^1\| \le C_1 h^{-1} 2X, \quad \text{due to Remark 6,}\\
\widetilde{J}_h\big(\xi_h^1, \xi_h^1\big) &\le C_2 h^{-1}\|\xi_h^1\|^2 \le C_2 h^{-1} 4X^2,\\
\|e_h^1\|_\infty &\ge X|\Omega|^{-1/2},
\end{aligned}$$

where the last inequality follows directly from Hölders inequality.

Applying these estimates, we can estimate the left-hand side of (56) as

$$\begin{aligned}
LHS &= \|\xi_h^1\|^2 + \|\xi_h^1 - \xi_h^0\|^2 + \tau_0\varepsilon\|\xi_h^1\|_{DG}^2 + \tau_0\widetilde{J}_h\big(\xi_h^1, \xi_h^1\big)\\
&\le 4X^2 + 9X^2 + \tau_0\varepsilon C_1^2 h^{-2} 4X^2 + \tau_0 C_2 h^{-1} 4X^2.
\end{aligned}$$

On the other hand, we get for the right-hand side of (56)

$$RHS \ge C\tau_0\Big(1 + \frac{\|e_h^1\|_\infty^2}{h^2}\Big)\|\xi_h^1\|^2 \ge C\tau_0\Big(1 + \frac{X^2}{|\Omega|h^2}\Big)X^2/4.$$

We want to determine, for what X is (56) satisfied, i.e. when is $LHS \le RHS$. This happens e.g. if

$$LHS \le 4X^2 + 9X^2 + \tau_0\varepsilon C_1^2 h^{-2} 4X^2 + \tau_0 C_2 h^{-1} 4X^2 \le C\tau_0\Big(1 + \frac{X^2}{|\Omega|h^2}\Big)X^2/4 \le RHS,$$

i.e. when X satisfies

$$C\tau_0\Big(1 + \frac{X^2}{|\Omega|h^2}\Big)X^2/4 - 4X^2 - 9X^2 - \tau_0\varepsilon C_1^2 h^{-2} 4X^2 - \tau_0 C_2 h^{-1} 4X^2 \ge 0. \tag{57}$$

However, the leading term X^4 in (57) has a positive coefficient $C\tau_0|\Omega|^{-1}h^{-2}/4$. Hence inequality (57) – and therefore inequality (56) – is satisfied for all X sufficiently large. □

8.1 Auxiliary problem and continuation of the discrete solution

Problem (54) represents a nonlinear equation on each time level t^{n+1} for the unknown function u_h^{n+1}. First, we prove that u_h^{n+1} exists is uniquely determined and depends continuously on τ_n. To this end we define a general abstract formulation of problem (54):

Definition 5. *(Auxiliary problem) Let* $t \in [0,T], \tau \in [0,T]$ *and* $U_h \in S_h$. *We seek* $u_\tau \in S_h$ *such that*

$$\big(u_\tau - U_h, \varphi_h\big) + \tau b_h\big(u_\tau, \varphi_h\big) + \tau\varepsilon A_h\big(u_\tau, \varphi_h\big) = \tau l_h\big(\varphi_h\big)(t+\tau), \quad \forall \varphi_h \in S_h. \tag{58}$$

Remark 10. *If we take $\tau := \tau_n, U_h := u_h^n, t := t_n$ and define $u_h^{n+1} := u_\tau$, the auxiliary problem (58) reduces to equation (54), which defines the approximate solution u_h^{n+1}. On the other hand, if we take $\tau := 0$ the solution of (58) is $u_\tau = u_h^n$. In between these two cases u_τ changes continuously with τ. For that we need to assume the right-hand side of (58) behaves "continuously with respect to time".*

Lemma 15. *Let $g \in C([0,T]; L^2(\Omega))$, $g_N \in C([0,T]; L^2(\Gamma_N))$ and $u_D \in C([0,T]; L^2(\Gamma_D))$. Then for all $h \in (0, h_0)$ there exists a function $\lambda_h \in C([0,T]; S_h)$ such that*

$$l_h(\varphi_h)(t) = \big(\lambda_h(t), \varphi_h\big), \quad \forall \varphi_h \in S_h. \tag{59}$$

Proof. We equip S_h with the $L^2(\Omega)$-scalar product. Then $l_h(\cdot)(t)$ is a linear functional on the finite-dimensional Hilbert space S_h. Hence, $l_h(\cdot)(t)$ is a continuous functional and by the Riesz representation theorem there exists $\lambda_h(t)$ such that (59) holds. It remains to show the continuity of $\lambda_h(t)$ with respect to t. Let $s, t \in [0,T]$, we estimate

$$\begin{aligned}
&\big\|\lambda_h(t) - \lambda_h(s)\big\| = \sup_{\varphi_h \in S_h} \tfrac{1}{\|\varphi_h\|}\big(\lambda_h(t) - \lambda_h(s), \varphi_h\big) = \sup_{\varphi_h \in S_h} \tfrac{1}{\|\varphi_h\|}\big(l_h(\varphi_h)(t) - l_h(\varphi_h)(s)\big) \\
&\le \sup_{\varphi_h \in S_h} \tfrac{1}{\|\varphi_h\|}\big(\|g(t) - g(s)\|\|\varphi_h\| + \|g_N(t) - g_N(s)\|_{L^2(\Gamma_N)}\|\varphi_h\|_{L^2(\Gamma_N)} \\
&\quad + \varepsilon|\Theta|\|u_D(t) - u_D(s)\|_{L^2(\Gamma_D)}\|\nabla\varphi_h\|_{L^2(\Gamma_D)} + \varepsilon C h^{-1}\|u_D(t) - u_D(s)\|_{L^2(\Gamma_D)}\|\varphi_h\|_{L^2(\Gamma_D)}\big).
\end{aligned}$$

Since $\varphi_h \in S_h$, we may use the multiplicative trace and inverse inequalities. Thus

$$\begin{aligned}
&\big\|\lambda_h(t) - \lambda_h(s)\big\| \le \sup_{\varphi_h \in S_h} \tfrac{1}{\|\varphi_h\|}\|\varphi_h\|\big(\|g(t) - g(s)\| + \|g_N(t) - g_N(s)\|_{L^2(\Gamma_N)} C h^{-1} \\
&\quad + \varepsilon|\Theta|\|u_D(t) - u_D(s)\|_{L^2(\Gamma_D)} C h^{-2} + \varepsilon C h^{-2}\|u_D(t) - u_D(s)\|_{L^2(\Gamma_D)}\big) \longrightarrow 0, \quad \text{as } s \to t.
\end{aligned}$$

□

Now we are ready to prove the basic result used in the construction of a continuation of the discrete solution, the continuity of u_τ with respect to τ. We shall assume the assumptions of Lemma 15 are satisfied from now on.

Lemma 16. *There exist constants $C_1, C_2 > 0$ independent of h, τ, t, ε, such that the following holds. Let $t \in [0,T], h \in (0, h_0), U_h \in S_h$ and $\tau \in [0, \tau_0)$, where $\tau_0 = \max\{C_1\varepsilon, C_2 h\}$. Then u_τ, the solution of (58), exists, is uniquely determined, $\|u_\tau\|$ is uniformly bounded with respect to $\tau \in [0, \tau_0)$ and $\|u_\tau\|$ depends continuously on τ.*

Proof. (**i**) **Existence**: We shall use the nonlinear Lax-Milgram theorem, cf. [43]. First, we define the forms $B_\tau : S_h \times S_h \to \mathbb{R}, L_\tau : S_h \to \mathbb{R}$:

$$\begin{aligned}
B_\tau(u, v) &:= (u, v) + \tau b_h(u, v) + \tau\varepsilon A_h(u, v), \\
L_\tau(v) &:= (U_h, v) + \tau l_h(v)(t + \tau).
\end{aligned}$$

Then problem (58) can be written as $B_\tau(u_\tau, \varphi_h) = L_\tau(\varphi_h)$ for all $\varphi_h \in S_h$. If we equip S_h with the $L^2(\Omega)$-norm, then $L_\tau(\cdot)$ is a linear functional on the finite-dimensional space S_h, hence L_τ is continuous and uniformly bounded with respect to τ, i.e. $\|L_\tau\|_{\mathcal{L}(S_h, \mathbb{R})} \le \|U_h\| + \tau\|\lambda_h(t)\| \le \|U_h\| + T\|\lambda_h\|_{L^\infty(S_h)} < +\infty$, since $\tau \le T$. In order to apply the nonlinear Lax-Milgram theorem, it remains to prove monotonicity and Lipschitz continuity of B_τ in the space S_h.

Monotonicity: For all $u, v \in S_h$, we have due to Lemmas 9 and 10, by Young's inequality

$$\begin{aligned} &B_\tau(u, u-v) - B_\tau(v, u-v) \\ &\quad \geq \|u-v\|^2 + \tau\varepsilon\|u-v\|_{DG}^2 - \tau L_{b_h}\|u-v\|_{DG}\|u-v\| - \tau L_{b_h}\|u-v\|^2 \\ &\quad \geq (1-\tau L_{b_h})\|u-v\|^2 + \tau\varepsilon\|u-v\|_{DG}^2 - \tau\varepsilon\|u-v\|_{DG}^2 - \tfrac{\tau}{4\varepsilon}L_{b_h}^2\|u-v\|^2 \\ &\quad = (1-\tau_0 L_{b_h} - \tfrac{\tau_0}{4\varepsilon}L_{b_h}^2)\|u-v\|^2. \end{aligned}$$

On the other hand, we may estimate by the inverse inequality and multiplicative trace inequalities $\tau L_{b_h}\|u-v\|_{DG} \leq \tau L_{b_h} C h^{-1}\|u-v\|$. Therefore

$$\begin{aligned} &B_\tau(u, u-v) - B_\tau(v, u-v) \\ &\quad \geq \|u-v\|^2 + \tau\varepsilon\|u-v\|_{DG}^2 - \tau L_{b_h} C h^{-1}\|u-v\|^2 - \tau L_{b_h}\|u-v\|^2 \\ &\quad \geq (1-\tau_0 L_{b_h} - \tau_0 L_{b_h} C h^{-1})\|u-v\|^2. \end{aligned}$$

Thus we have $L^2(\Omega)$-monotonicity of B_τ, e.g. if $\tau_0 < \frac{1}{2}L_{b_h}^{-1}$ and either $\tau_0 < 2\varepsilon L_{b_h}^{-2}$ or $\tau_0 < \frac{1}{2}L_{b_h}^{-1}C^{-1}h$.

Lipschitz continuity: For all $u, v, w \in S_h$, we have due to Lemmas 9, 10 and Remark 6

$$\begin{aligned} &\big|B_\tau(u,u-v) - B_\tau(v, u-v)\big| \\ &\qquad \leq \|u-v\|^2 + \tau\varepsilon\|u-v\|_{DG}^2 + \tau L_{b_h}\|u-v\|_{DG}\|u-v\| + \tau L_{b_h}\|u-v\|^2 \\ &\qquad \leq (1 + \tau_0\varepsilon C h^{-2} + \tau_0 L_{b_h} C h^{-1} + \tau_0 L_{b_h}^2)\|u-v\|^2. \end{aligned}$$

By the nonlinear Lax-Milgram theorem, we obtain the existence and uniqueness of $u_\tau \in S_h$, the solution of (58). Moreover, $\|u_\tau\|$ is uniformly bounded for $\tau \in [0, \tau_0)$, since $\|u_\tau\| \leq C\|L_\tau\|_{\mathcal{L}(L^2(\Omega),\mathbb{R})} \leq C\|U_h\| + \tau_0\|l(t)\|_{\mathcal{L}(S_h,\mathbb{R})}$.

We note that by taking $v := 0$ in the monotonicity estimates of B_τ, we may prove S_h-coercivity, i.e. there exists some $\alpha > 0$ such that

$$B_\tau(u,u) = B_\tau(u, u-0) - B_\tau(0, u-0) \geq \alpha\|u\|^2.$$

(**ii**) **Continuity with respect to** τ: Let $\tau, \bar\tau \in (0, \tau_0)$. We subtract (58) for τ and $\bar\tau$, obtaining

$$B_\tau(u_\tau, \varphi_h) - B_{\bar\tau}(u_{\bar\tau}, \varphi_h) = L_\tau(\varphi_h) - L_{\bar\tau}(\varphi_h). \tag{60}$$

First, we estimate the right-hand side of (60) using the representation formula (59):

$$\begin{aligned} L_\tau(\varphi_h) - L_{\bar\tau}(\varphi_h) &= \tau\big(\lambda_h(t+\tau) - \lambda_h(t+\bar\tau), \varphi_h\big) + (\tau-\bar\tau)\big(\lambda_h(t+\bar\tau), \varphi_h\big) \\ &\leq T\|\lambda_h(t+\tau) - \lambda_h(t+\bar\tau)\|\|\varphi_h\| + |\tau-\bar\tau|\|\lambda_h\|_{L^\infty(S_h)}\|\varphi_h\|. \end{aligned} \tag{61}$$

Setting $\varphi_h := u_\tau - u_{\bar\tau}$ in (60) and rearranging gives us

$$\begin{aligned} &B_\tau(u_\tau, u_\tau - u_{\bar\tau}) - B_\tau(u_{\bar\tau}, u_\tau - u_{\bar\tau}) \\ &\quad = B_{\bar\tau}(u_{\bar\tau}, u_\tau - u_{\bar\tau}) - B_\tau(u_{\bar\tau}, u_\tau - u_{\bar\tau}) + L_\tau(u_\tau - u_{\bar\tau}) - L_{\bar\tau}(u_\tau - u_{\bar\tau}). \end{aligned}$$

Taking into account the monotonicity of B_τ on the left-hand side and estimating the right-hand side by (61), Lemma 9 and 10 and the inverse inequality, we obtain

$$\begin{aligned} &\alpha\|u_\tau - u_{\bar\tau}\|^2 \\ &\leq |\tau-\bar\tau|\big(C(\varepsilon,h)\|u_{\bar\tau}\| + \|\lambda_h\|_{L^\infty(S_h)}\big)\|u_\tau - u_{\bar\tau}\| + T\|\lambda_h(t+\tau) - \lambda_h(t+\bar\tau)\|\|u_\tau - u_{\bar\tau}\| \\ &\implies \quad \|u_\tau - u_{\bar\tau}\| \leq \tfrac{C(\varepsilon,h)}{\alpha}|\tau-\bar\tau|\big(C(\varepsilon,h)\|u_{\bar\tau}\| + \|\lambda_h\|_{L^\infty(S_h)}\big) + \tfrac{T}{\alpha}\|\lambda_h(t+\tau) - \lambda_h(t+\bar\tau)\|, \end{aligned}$$

where $\alpha > 0$ is the monotonicity constant of B_τ and $C(\varepsilon, h)$ is a constant depending on ε, h. It follows directly that $\|u_{\bar\tau}\| \to \|u_\tau\|$ as $\bar\tau \to \tau$, since by the inverted triangle inequality

$$\big|\|u_\tau\| - \|u_{\bar\tau}\|\big| \le \|u_\tau - u_{\bar\tau}\| \longrightarrow 0, \quad \text{as } \bar\tau \to \tau.$$

We have proved the continuity of $\|u_\tau\|$ with respect to $\tau \in (0, \tau_0)$. It remains to prove the continuity of $\|u_\tau\|$ at $\tau = 0$. This is straightforward, since for $\tau = 0$, we have $u_\tau = U_h$. We test (58) with $\varphi_h := u_\tau - U_h$

$$\begin{aligned}\|u_\tau - U_h\|^2 &\le \tau |l_h\big(u_\tau - U_h\big)(t)| + \tau |b_h\big(u_\tau, u_\tau - U_h\big)| + \tau\varepsilon |A_h\big(u_\tau, u_\tau - U_h\big)| \\ &\le \tau\big(\|\lambda_h\|_{L^\infty(S_h)} + C(\varepsilon, h)\|u_\tau\|\big)\|u_\tau - U_h\|.\end{aligned}$$

This implies that $\|u_\tau - U_h\| \to 0$, and therefore $\|u_\tau\| \to \|U_h\|$, as $\tau \to 0$. □

As we have seen in Remark 10, by taking $U_h := u_h^n$ in (58) we obtain $u_\tau = u_h^{n+1}$ for $\tau := \tau_n$ and $u_\tau = u_h^n$ for $\tau := 0$. For general $\tau \in [0, \tau_n]$, u_τ depends continuously on the parameter τ. This allows us to construct a function $\tilde u_h \in C([0,T]; S_h)$ which "interpolates" the values $\{u_h^n\}_{n=0}^N$ and which is constructed using essentially the implicit problem itself.

Definition 6 (Continuated discrete solution)**.** *Let $\tilde u_h : [0,T] \to S_h$ be defined as follows: For $s \in [t_n, t_{n+1}]$ we set $\tilde u_h(s) := u_\tau$, the solution of the auxiliary problem (58) with $\tau := s - t_n$, $t := t_{n+1}$ and $U_h := u_h^n$. Furthermore, we define $\tilde e_h := u - \tilde u_h$ and $\tilde\xi_h := \Pi_h u - \tilde u_h$.*

Remark 11. *Under the assumptions of Lemma 16, $\tilde u_h$ is uniquely determined, $\tilde u_h \in C([0,T]; S_h)$ and $\tilde e_h \in C([0,T]; L^2(\Omega))$, due to the regularity (15). Also, $\tilde u_h(t_n) = u_h^n$, $\tilde e_h(t_n) = e_h^n$ and $\tilde\xi_h(t_n) = \xi_h^n$, for all $n = 0, \cdots, N$. Therefore, estimates of $\tilde e_h(\cdot)$ imply estimates of e_h^n. Finally, we note that $\tilde e_h = \eta_h + \tilde\xi_h$.*

8.2 Estimates based on continuous mathematical induction.

Since $\tilde u_h$ is constructed using the auxiliary problem (58), which is essentially the original implicit scheme (54) with special data, we can derive error estimates for $\tilde u_h$ in a standard manner. We start by proving a discrete analogy of Lemma 11. However, first we need some standard results concerning the approximation of time derivatives.

Lemma 17. *Let $u_{tt} = \frac{\partial^2 u}{\partial t^2} \in L^2(0,T; L^2(\Omega))$. Let $s \in [t_n, t_{n+1}]$ and $\varphi_h \in S_h$. Then*

$$\begin{aligned}\big|\big(u(s) - u(t_n) - (s - t_n)u_t(s), \varphi_h\big)\big| &\le C\tau^2 \|u_{tt}\|_{L^\infty(L^2)}\|\varphi_h\|, \\ \big|\big(\eta_h(s) - \eta_h(t_n), \varphi_h\big)\big| &\le C\tau h^{p+1}\|u_t\|_{L^\infty(H^{p+1})}\|\varphi_h\|.\end{aligned}$$

Proof. Cf. [18]. □

Now we shall prove a discrete analogy of Lemma 11. For simplicity we assume a uniform partition of $[0,T]$, i.e. $\tau_n := \tau$ for all $n = 0, \cdots, N$.

Lemma 18. *Let $p \ge d/2$. Let $s \in (t_n, t_{n+1}]$ for some $n \in \{0, \cdots, N\}$. If*

$$\|\tilde e_h(s)\| \le h^{1+d/2} \quad \text{and} \quad \|\tilde e_h(t_k)\| \le h^{1+d/2},\ \forall k = 0, \cdots, n, \tag{62}$$

then we have the estimate

$$\begin{aligned}&\max_{t \in \{t_0, \cdots, t_n, s\}} \|\tilde e_h(t)\|^2 + \sum_{k=1}^{n} \tau\Big(\varepsilon\|\tilde e_h(t_k)\|_{DG}^2 + \widetilde J_h\big(\tilde e_h(t_k), \tilde e_h(t_k)\big)\Big) \\ &\quad + (s - t_n)\Big(\varepsilon\|\tilde e_h(s)\|_{DG}^2 + \widetilde J_h\big(\tilde e_h(s), \tilde e_h(s)\big)\Big) \le C_T^2\big(h^{2p+1} + \varepsilon h^{2p} + \tau^2\big),\end{aligned} \tag{63}$$

where the constant C_T is independent of $s, n, h, \tau, \varepsilon$.

Proof. Since $\tilde{e}_h = u - \tilde{u}_h$ and $\tilde{u}_h$ is defined by the Auxiliary problem (58) with $\tau = s - t_n$, $U_h = u_h^n$, in order to obtain an equation for $\tilde{e}_h$, we subtract (58) from (14). Furthermore, in (14) we introduce the time difference instead of the time derivative u_t. Thus $\tilde{e}_h(s)$ satisfies

$$\begin{aligned}\big(\tilde{e}_h(s)-\tilde{e}_h(t_n),\varphi_h\big) + (s-t_n)\big(b_h(u(s),\varphi_h) - b_h(\tilde{u}_h(s),\varphi_h)\big) + (s-t_n)\varepsilon A_h(\tilde{e}_h(s),\varphi_h)\\ = \big(u(s)-u(t_n)-(s-t_n)u_t(s),\varphi_h\big).\end{aligned} \tag{64}$$

We set $\varphi_h := \tilde{\xi}_h(s)$ and use the fact that $2(a-b,a) = \|a\|^2 - \|b\|^2 + \|a-b\|^2$. Furthermore, we estimate the convective terms using Lemma 7 and the diffusion terms using Lemma 9. For the remaining terms, we use Lemma 17. Thus we obtain the inequality

$$\begin{aligned}&\|\tilde{\xi}_h(s)\|^2 - \|\tilde{\xi}_h(t_n)\|^2 + \|\tilde{\xi}_h(s)-\tilde{\xi}_h(t_n)\|^2 + (s-t_n)\varepsilon\|\tilde{\xi}_h(s)\|_{DG}^2 + (s-t_n)\widetilde{J}_h\big(\tilde{\xi}_h(s),\tilde{\xi}_h(s)\big)\\ &\le C\tau\Big(1+\frac{\|\tilde{e}_h(s)\|_\infty^2}{h^2}\Big)\Big(\big(h^{2p+1}+\varepsilon h^{2p}\big)|u|^2_{L^\infty(H^{p+1})} +\\ &\quad + h^{2p+2}\|u_t\|^2_{L^\infty(H^{p+1})} + \tau^2\|u_{tt}\|^2_{L^\infty(L^2)} + \|\tilde{\xi}_h(s)\|^2\Big).\end{aligned} \tag{65}$$

As in (45), we may show that from (62), it follows that $\|\tilde{e}_h(s)\|_\infty \le Ch$. Thus (65) reduces to

$$\begin{aligned}\|\tilde{\xi}_h(s)\|^2 + (s-t_n)\varepsilon\|\tilde{\xi}_h(s)\|_{DG}^2 + (s-t_n)\widetilde{J}_h\big(\tilde{\xi}_h(s),\tilde{\xi}_h(s)\big)\\ \le \|\tilde{\xi}_h(t_n)\|^2 + C\tau\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2+\|\tilde{\xi}_h(s)\|^2\big),\end{aligned} \tag{66}$$

which may be written as

$$\|\tilde{\xi}_h(s)\|^2 \le \tfrac{1}{1-C\tau}\|\tilde{\xi}_h(t_n)\|^2 + \tfrac{C\tau}{1-C\tau}\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2\big). \tag{67}$$

Similarly as $\tilde{e}_h(s)$ satisfies (64), $\tilde{e}_h(t_k)$ satisfies the following equation for all $k = 0, \cdots, n-1$:

$$\begin{aligned}\big(\tilde{e}_h(t_{k+1})-\tilde{e}_h(t_k),\varphi_h\big) + \tau\big(b_h(u(t_{k+1}),\varphi_h) - b_h(\tilde{u}_h(t_{k+1}),\varphi_h)\big) + \tau\varepsilon A_h(\tilde{e}_h(t_{k+1}),\varphi_h)\\ = \big(u(t_{k+1})-u(t_k)-\tau u_t(t_{k+1}),\varphi_h\big).\end{aligned} \tag{68}$$

We set $\varphi_h := \tilde{\xi}_h(t_{k+1})$ and proceed similarly as in estimates (65)-(66) to obtain

$$\begin{aligned}\|\tilde{\xi}_h(t_{k+1})\|^2 + \tau\varepsilon\|\tilde{\xi}_h(t_{k+1})\|_{DG}^2 + \tau\widetilde{J}_h\big(\tilde{\xi}_h(t_{k+1}),\tilde{\xi}_h(t_{k+1})\big)\\ \le \|\tilde{\xi}_h(t_k)\|^2 + C\tau\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2+\|\tilde{\xi}_h(t_{k+1})\|^2\big),\end{aligned} \tag{69}$$

which we simplify to

$$\|\tilde{\xi}_h(t_{k+1})\|^2 \le \tfrac{1}{1-C\tau}\|\tilde{\xi}_h(t_k)\|^2 + \tfrac{C\tau}{1-C\tau}\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2\big). \tag{70}$$

Assuming $C\tau \le 1/2$, we may define

$$A := h^{2p+1}+\varepsilon h^{2p}+\tau^2, \quad B := \tfrac{1}{1-C\tau} = 1 + \tfrac{C\tau}{1-C\tau} \le 1 + 2C\tau \le e^{2C\tau}.$$

Taking into account (67) and (70), we have by induction

$$\begin{aligned}\|\tilde{\xi}_h(s)\|^2 &\le B^{n+1}\|\tilde{\xi}_h(0)\|^2 + \tfrac{B^{n+1}-1}{B-1}BC\tau A \\ &\le e^{(n+1)2C\tau}BA \le e^{2CT}\tfrac{1}{2}A = \widetilde{C}_T\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2\big),\end{aligned} \tag{71}$$

since $\tilde{\xi}_h(0)=0$. Similarly, we obtain by induction from (70)

$$\|\tilde{\xi}_h(t_k)\|^2 \le \widetilde{C}_T\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2\big), \quad \forall k=0,\cdots,n. \tag{72}$$

Since $\tilde{e}_h = \tilde{\xi}_h + \eta_h$, the triangle inequality and (71), (72) gives us the desired estimate of the first left-hand side term in (63).

As for the remaining left-hand side terms in (63), we sum (66) and (69) for all $k=0,\cdots,n-1$. After applying estimates (71) and (72), we obtain

$$\begin{aligned}&\sum_{k=1}^{n}\tau\Big(\varepsilon\|\tilde{\xi}_h(t_k)\|_{DG}^2 + \widetilde{J}_h\big(\tilde{\xi}_h(t_k),\tilde{\xi}_h(t_k)\big)\Big) + (s-t_n)\Big(\varepsilon\|\tilde{\xi}_h(s)\|_{DG}^2 + \widetilde{J}_h\big(\tilde{\xi}_h(s),\tilde{\xi}_h(s)\big)\Big) \\ &\quad\le \|\tilde{\xi}_h(0)\|^2 + (n+1)C\tau\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2 + \widetilde{C}_T(h^{2p+1}+\varepsilon h^{2p}+\tau^2)\big) \\ &\quad\le CT(1+\widetilde{C}_T)(h^{2p+1}+\varepsilon h^{2p}+\tau^2).\end{aligned}$$

Again, we apply the triangle inequality and suitable estimates for η_h to obtain (63). □

Remark 12. *The functions $u(\cdot), \tilde{u}_h(\cdot)$ are continuous mappings from $[0,T]$ to $L^2(\Omega)$. Therefore, $\tilde{e}_h(\cdot)$ is a* uniformly continuous *function from $[0,T]$ to $L^2(\Omega)$. By definition,*

$$\forall\epsilon>0\ \exists\delta>0:\ s,\bar{s}\in[0,t], |s-\bar{s}|\le\delta \quad\Longrightarrow\quad \|\tilde{e}_h(s)-\tilde{e}_h(\bar{s})\|\le\epsilon.$$

Theorem 19 (Main theorem – implicit version)**.** *Let $p>1+d/2$. Let $h_1,\tau_1>0$ be such that $C_T(h_1^{p+1/2}+\sqrt{\varepsilon}h_1^p+\tau_1)=\frac{1}{2}h_1^{1+d/2}$, where C_T is the constant from Lemma 18 and let $\tau_1<\tau_0$, where τ_0 is defined in Lemma 16. Then for all $h\in(0,h_1), \tau\in(0,\tau_1)$ we have the estimate*

$$\max_{n\in\{0,\cdots,N+1\}}\|e_h^n\|^2 + \sum_{n=1}^{N+1}\tau\Big(\varepsilon\|e_h^n\|_{DG}^2 + \widetilde{J}_h\big(e_h^n,e_h^n\big)\Big) \le C_T^2\big(h^{2p+1}+\varepsilon h^{2p}+\tau^2\big). \tag{73}$$

Proof. We have $p>1+d/2$, therefore h_1,τ_1 exist and $C_T(h^{p+1/2}+\sqrt{\varepsilon}h^p+\tau)\le\frac{1}{2}h^{1+d/2}$ for all $h\in(0,h_1], \tau\in(0,\tau_1]$. We fix an arbitrary $h\in(0,h_1]$ and $\tau\in(0,\tau_1]$. By Remark 12, there exists $\delta>0$ such that if $s,\bar{s}\in[0,T], |s-\bar{s}|\le\delta$ then $\|\tilde{e}_h(s)-\tilde{e}_h(\bar{s})\|\le\frac{1}{2}h^{1+d/2}$.

We define $s_i=i\delta$, $i=0,1,\ldots$ and set $M:=\max\{i=0,1,\ldots; s_i<T\}$, $s_{M+1}:=T$. This defines a partition $0=s_0<s_1<\cdots<s_{M+1}=T$ of the interval $[0,T]$ into $M+1$ intervals of length (at most) δ.

We shall now prove by induction that for all $i=1,\ldots,M+1$

$$\max_{\vartheta\in[0,s_i]}\|\tilde{e}_h(\vartheta)\| \le C_T\big(h^{p+1/2}+\sqrt{\varepsilon}h^p+\tau\big) \le h^{1+d/2}. \tag{74}$$

Inequality (73) is obtained by taking $i:=M+1$ in (74) and applying Lemma 18 with $s:=t_{N+1}=T$.

(i) $i=1$: We know that $\|\tilde{e}_h(0)\| = \|\eta_h(0)\| \le C_T(h^{p+1/2}+\sqrt{\varepsilon}h^p+\tau)\le\frac{1}{2}h^{1+d/2}$. By uniform continuity, we have for all $s\in[0,s_1]$

$$\|\tilde{e}_h(s)\| \le \|\tilde{e}_h(0)\| + \|\tilde{e}_h(s)-\tilde{e}_h(0)\| \le \tfrac{1}{2}h^{1+d/2}+\tfrac{1}{2}h^{1+d/2} = h^{1+d/2}.$$

Therefore, by Lemma 18 we obtain estimate (74) on $[0, s_1]$, i.e. for $i = 1$.
(ii) Induction step: We assume that (74) holds for a general $i < M + 1$. Therefore, $\|\tilde{e}_h(s_i)\| \leq C_T(h^{p+1/2} + \sqrt{\varepsilon}h^p + \tau) \leq \frac{1}{2}h^{1+d/2}$. By uniform continuity, we have that for all $s \in [s_i, s_{i+1}]$

$$\|\tilde{e}_h(s)\| \leq \|\tilde{e}_h(s_i)\| + \|\tilde{e}_h(s) - \tilde{e}_h(s_i)\| \leq \tfrac{1}{2}h^{1+d/2} + \tfrac{1}{2}h^{1+d/2} = h^{1+d/2}.$$

This and the induction assumption imply that $\|\tilde{e}_h(\vartheta)\| \leq h^{1+d/2}$ for $\vartheta \in [0, s_i] \cup [s_i, s_{i+1}] = [0, s_{i+1}]$. Therefore, by Lemma 18 we obtain estimate (74) on $[0, s_{i+1}]$, i.e. for '$i := i + 1$'. □

Remark 13. *We conclude with several remarks.*

- *The* CFL *condition required in Theorem 19 effectively imposes $\tau = O(h^{1+d/2})$. This is rather restrictive from the perspective of an implicit scheme. This condition arises due to the key step in our analysis, where we require $C_T(h^{p+1/2}+\sqrt{\varepsilon}h^p+\tau) \leq \frac{1}{2}h^{1+d/2}$. If we assumed higher regularity of $\mathbf{f}$ and used Lemma 8, we would obtain the* CFL *condition $\tau = O(h^{(1+d)/2})$ along with the less restrictive order condition $p > (1 + d)/2$.*

- *Such a restrictive* CFL *condition is purely an artefact of the proof due to the nonlinearity of the problem. For linear problems, we may expect the standard $\tau = O(h)$ condition in explicit schemes (for third-order RungeKutta schemes in [10]) as well as in the space-time DG scheme ([38]).*

- *For the purely convective case $\varepsilon = 0$, we require only $C_T(h^{p+1/2} + \tau) \leq \frac{1}{2}h^{1+d/2}$, which leads to the order condition $p > (1 + d)/2$. Using Lemma 8, this improves to $p > d/2$.*

- *We have treated the simplest first order temporal discretization. Were we to analyze a scheme of kth order in time, we would require that $C_T(h^{p+1/2}+\sqrt{\varepsilon}h^p+\tau^k) \leq \frac{1}{2}h^{1+d/2}$, which leads to the less restrictive CFL condition $\tau = O(h^{(1+d/2)/k})$ or $\tau = O(h^{(1+d)/(2k)})$, using Lemma 8.*

- *Essentially, we are still limited by the CFL condition $\tau = O(max\{\varepsilon, h\})$ of Lemma 16 to guarantee existence and continuity of the continuated discrete solution.*

9 From global to local Lipschitz continuity of f

Up to now, we have assumed $\mathbf{f} \in (C_b^2(\mathbb{R}))^d$, i.e. *global* Lipschitz continuity of $\mathbf{f}$ and $\mathbf{f}'$. We have stated in Remark 1 that in [44], these global assumptions are replaced by local ones by modifying $\mathbf{f}$ away from the set of values of u. This procedure does not change the exact solution of the continuous problem, however, one obtains a new discrete problem for which we cannot guarantee *apriori* that it has the same solution as the unmodified version. In this section we show how to obtain error estimates for locally Lipschitz continuous $\mathbf{f}$ – without modifying the scheme – using continuous mathematical induction.

9.1 Basic notation and results

For simplicity we shall derive estimates for the method of lines, the implicit scheme can be treated similarly. Let $\mathbf{f} \in (C^2(\mathbb{R}))^d$ and let $u_h \in C^1([0,T]; S_h)$ be the DG solution of problem (2)–(5) as defined in Definition 1.

Definition 7. *Let $v : Q_T \to \mathbb{R}$ be a function. We define the* range *of v as $\mathcal{R}(v) := \overline{\{v(x,t); (x,t) \in Q_T\}}$.*

Definition 8. *Let $A \subset \mathbb{R}$ be a closed set. We define the* local generalized Lipschitz constant *of $\mathbf{f}$ on A as $L_{\mathbf{f}}(A) := \|\mathbf{f}\|_{(W^{2,\infty}(A))^d}$.*

Remark 14. *If $A \subset \mathbb{R}$ is bounded then $L_{\mathbf{f}}(A) < \infty$. Due to the regularity assumptions (15), we have $u \in C(\overline{Q_T})$ and thus $L_{\mathbf{f}}(\mathcal{R}(u)) < \infty$ is a constant depending only on u.*

Definition 9. *Let $h \in (0, h_0), t \in [0.T]$. We define the* admissible set $\mathcal{U}_h^{ad}(t) := \{v \in S_h; \|u(t) - v\| \le h^{1+d/2}\}$.

In Section 7.1, we have used mathematical induction with respect to t to ensure that (43) holds, i.e. $u_h(\vartheta) \in \mathcal{U}_h^{ad}(\vartheta)$ for $\vartheta \in [0,t]$. As a byproduct, we got the desired error estimates. Here we shall do something similar using the fact that functions in $\mathcal{U}_h^{ad}(t)$ have values in some compact interval and thus we may use Lipschitz continuity of $\mathbf{f}$ for such functions.

Lemma 20. *Let $p+1 \ge d/2$. There exists a constant $R \ge 0$ independent of h such that for all $h \in (0, h_0)$*

a) $\mathcal{R}(u) \subseteq [-R, R]$,
b) $\mathcal{R}(v) \subseteq [-R, R]$, *for all v such that $v(t) \in \mathcal{U}_h^{ad}(t)$ for all $t \in [0,T]$.*

Proof. Inclusion *a)* is trivial due to Remark 14. As for *b)*, let $t \in [0,T], v(t) \in \mathcal{U}_h^{ad}(t)$. Then due to Lemmas 4 and 5

$$
\begin{aligned}
&\|v(t) - u(t)\|_\infty \\
&\quad \le \|v(t) - \Pi_h u(t)\|_\infty + \|\Pi_h u(t) - u(t)\|_\infty \le C_I h^{-d/2} \|v(t) - \Pi_h u(t)\| + Ch|u(t)|_{W^{1,\infty}} \\
&\quad \le C_I h^{-d/2} \|v(t) - u(t)\| + C_I h^{-d/2} \|u(t) - \Pi_h u(t)\| + Ch_0 |u|_{L^\infty(W^{1,\infty})} \\
&\quad \le C_I h^{-d/2} h^{1+d/2} + C_I h^{-d/2} C h^{p+1} |u(t)|_{H^{p+1}} + Ch_0 |u|_{L^\infty(W^{1,\infty})} \\
&\quad \le C_I h_0 + C_I h_0^{p+1-d/2} |u|_{L^\infty(H^{p+1})} + Ch_0 |u|_{L^\infty(W^{1,\infty})} =: R_1.
\end{aligned}
$$

Therefore due to *a)*

$$\|v(t)\|_\infty \le \|v(t) - u(t)\|_\infty + \|u(t)\|_\infty \le R,$$

with some constant R independent of h. □

Definition 10. *We set $L := L_{\mathbf{f}}\big([-R, R]\big) < \infty$, where R is the constant from Lemma 20.*

Now, we shall state the fundamental properties of the convective terms which we need in our analysis. These results can be proved similarly as in the case of *global* properties of $\mathbf{f}$, since in our analysis all the arguments of $\mathbf{f}$, e.g. u, u_h, have values in $[-R, R]$. On this compact set of values, we may use Lipschitz continuity and boundedness of $\mathbf{f}$

and its derivatives. However, to proceed, we also need assumptions on the numerical flux H which mimic those of $\mathbf{f}$:

Assumption $(\mathbf{H1})_{\mathbf{loc}}$: $H(v,w,\mathbf{n})$ is defined in $\mathbb{R}^2 \times B_1$, where $B_1 = \{\mathbf{n} \in \mathbb{R}^2; |\mathbf{n}| = 1\}$, and is *Lipschitz-continuous* on $[-R,R]$ with respect to v, w:

$$|H(v,w,\mathbf{n}) - H(v^*,w^*,\mathbf{n})| \le C_L(|v-v^*|+|w-w^*|), \quad \forall v,\, w,\, v^*,\, w^* \in [-R,R],\ \mathbf{n} \in B_1.$$

In the following, we shall assume that H satisfies conditions $(\mathrm{H1})_{\mathrm{loc}}$, (H2), (H3) and (H4). We note that $(\mathrm{H1})_{\mathrm{loc}}$ is satisfied e.g. if H is *locally* Lipschitz continuous with respect to the first two arguments.

Lemma 21. *There exists a constant $C \ge 0$ such that for all $v \in H^1(\mathcal{T}_h)$ which satisfy $\overline{\{v(x); x \in \Omega\}} \subseteq [-R,R]$, we have $0 \le \alpha(v) \le C$ and*

$$\big|\mathbf{f}'(\langle v\rangle)\cdot\mathbf{n}\big| \le 2\alpha(v) + C\big|[v]\big|.$$

Proof. The proof is identical to that of Lemma 6, one only needs to realize that it is possible to use L instead of $\|\mathbf{f}\|_{W^{2,\infty}(\mathbb{R})}$. This is due to the fact that in the proof of Lemma 6 one only estimates remainders in Taylor expansions, e.g. $\mathbf{f}''_{v^{(L)},\langle v\rangle}(x)$ with components $f_s''(\vartheta_s v^{(L)}(x) + (1-\vartheta_s)\langle v(x)\rangle)$ for some $\vartheta_s \in [0,1]$ and $s = 1,\cdots,d$. Since v has values in the interval $[-R,R]$, we also have $\vartheta_s v^{(L)}(x) + (1-\vartheta_s)\langle v(x)\rangle \in [-R,R]$ and thus $|\mathbf{f}''_{v^{(L)},\langle v\rangle}| \le L$. □

Now, we shall prove the "local" analogy of Lemma 7. Again, we usually omit the argument t for simplicity.

Lemma 22. *Let $p \ge d/2$. There exists a constant $C \ge 0$ independent of h, t, u_h, such that if $u_h(t) \in \mathcal{U}_h^{ad}(t)$, then*

$$b_h\big(u_h,\xi\big) - b_h\big(u,\xi\big) \le C\big(h^{2p+1}|u(t)|^2_{H^{p+1}} + \|\xi\|^2\big) - \frac{1}{2}\widetilde{J}_h(\xi,\xi). \tag{75}$$

Proof. The proof is identical to that of Lemma 7, only one uses L instead of $\|\mathbf{f}\|_{W^{2,\infty}(\mathbb{R})}$. This is due to the fact that in the proof of Lemma 7 all arguments of $\mathbf{f}, \mathbf{f}'$ and $\mathbf{f}''$ lie in the interval $[-R,R]$. Specifically, one uses Lipschitz continuity and boundedness for arguments such as $u, \overline{u}_K, u_h, \langle u_h\rangle$ or $\frac{1}{2}(u + u_h^{(L)})$, which all have values in $[-R,R]$, since $u(t), u_h(t) \in \mathcal{U}_h^{ad}(t)$.

Furthermore, one estimates remainders in Taylor expansions, such as $\mathbf{f}''_{u,u_h}, \mathbf{f}''_{u,\langle u_h\rangle}$ and $\mathbf{f}''_{u,u_h^{(L)}}$, which by definition is $\mathbf{f}''$ evaluated at some convex combination of the subscript arguments, e.g. $\vartheta_s u(x,t) + (1-\vartheta_s)u_h(x,t)$ for some $\vartheta_s \in [0,1]$. Hence, all these arguments also have values in $[-R,R]$ and we can bound e.g. $|\mathbf{f}''_{u,u_h}| \le L$. Finally, throughout the proof one uses the properties of $\alpha(u_h(t))$ and $\alpha(u(t))$ from Lemma 21. Both $u_h(t)$ and $u(t)$ satisfy the assumptions of this Lemma.

In such a way we obtain the same estimate as in Lemma 7,

$$b_h(u_h,\xi) - b_h(u,\xi) \le C\Big(1 + \frac{\|e_h\|^2_\infty}{h^2}\Big)\big(h^{2p+1}|u(t)|^2_{H^{p+1}} + \|\xi\|^2\big) - \frac{1}{2}\widetilde{J}_h(\xi,\xi). \tag{76}$$

Now we use the fact that $u_h(t) \in \mathcal{U}_h^{ad}(t)$, i.e. $\|e_h(t)\| \le h^{1+d/2}$. Since $p \ge d/2$, this implies due to (45), that $\|e_h(t)\|_\infty \le Ch$ for some constant C independent of $h, t, u_h(t)$. Thus we may estimate $h^{-2}\|e_h\|^2_\infty \le C$ and (76) reduces to (75). □

9.2 Estimates for the method of lines

Basically, we proceed as in Section 7. We subtract (13) from (14), set $\varphi_h := \xi_h(t) \in S_h$ and apply standard estimates to the evolutionary and diffusion terms. As for the convective terms, we notice that if we know *apriori* that $u_h(\vartheta) \in \mathcal{U}_h^{ad}(\vartheta)$ for $\vartheta \in [0,t]$, then we may apply Lemma 22 on this interval. By applying Gronwall's lemma, we obtain error estimates on $[0,t]$. Finally, we use continuous mathematical induction to go from $t=0$ to $t=T$.

Lemma 23. *Let $t \in [0,T]$ and $p \geq d/2$. If $u_h(\vartheta) \in \mathcal{U}_h^{ad}(\vartheta)$ for all $\vartheta \in [0,t]$, then we have the estimate*

$$\max_{\vartheta\in[0,t]} \|e_h(\vartheta)\|^2 + \int_0^t \varepsilon\|e_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(e_h(\vartheta), e_h(\vartheta)\big)\,\mathrm{d}\vartheta \leq C_T^2\big(h^{2p+1} + \varepsilon h^{2p}\big), \tag{77}$$

where C_T is a constant independent of h, t and ε.

Proof. We subtract (13) from (14) and set $\varphi_h := \xi_h(t) \in S_h$. We get

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi_h(t)\|^2 + \varepsilon A_h\big(\xi_h(t), \xi_h(t)\big) \\ &\qquad = -\varepsilon A_h\big(\eta_h(t), \xi_h(t)\big) + b_h\big(u_h(t), \xi_h(t)\big) - b_h\big(u(t), \xi_h(t)\big) - \Big(\frac{\partial \eta_h(t)}{\partial t}, \xi_h(t)\Big). \end{aligned} \tag{78}$$

Now, we apply the ellipticity and boundedness of A_h and Lemma 22 for the convective terms. Finally, we obtain from (78)

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\|\xi_h(t)\|^2 + \varepsilon\|\xi_h(t)\|_{DG}^2 + \widetilde{J}_h\big(\xi_h(t), \xi_h(t)\big) \\ &\qquad \leq C\big(h^{2p+1}|u(t)|_{H^{p+1}}^2 + \varepsilon h^{2p}|u(t)|_{H^{p+1}}^2 + h^{2p+2}|u_t(t)|_{H^{p+1}}^2 + \|\xi_h(t)\|^2\big). \end{aligned}$$

Integration from 0 to $t \in [0,T]$ yields

$$\|\xi_h(t)\|^2 + \int_0^t \varepsilon\|\xi_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(\xi_h(\vartheta), \xi_h(\vartheta)\big)\,\mathrm{d}\vartheta \leq C\big(h^{2p+1} + \varepsilon h^{2p}\big) + C\int_0^t \|\xi_h(\vartheta)\|^2\,\mathrm{d}\vartheta, \tag{79}$$

where the constant C is independent of h, t, ε. Gronwall's inequality applied to (79) gives us a constant $\widetilde{C}_T$, independent of h, t, ε, such that

$$\max_{\vartheta\in[0,t]} \|\xi_h(\vartheta)\|^2 + \int_0^t \varepsilon\|\xi_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(\xi_h(t), \xi_h(t)\big)\,\mathrm{d}\vartheta \leq \widetilde{C}_T\big(h^{2p+1} + \varepsilon h^{2p}\big).$$

Using a similar inequality for η_h, we obtain estimate (77) for some constant C_T. □

Now it remains to get rid of the *apriori* assumption $u_h(\vartheta) \in \mathcal{U}_h^{ad}(\vartheta)$ on $[0,t]$. As in Section 7.1, we shall use the uniform continuity of $e_h(\cdot)$ as a function from $[0,t]$ to $L^2(\Omega)$, cf. Remark 7.

Theorem 24 (Main theorem – local version)**.** *Let $p > 1 + d/2$. Then there exists $h_1 > 0$ such that $C_T(h_1^{p+1/2} + \sqrt{\varepsilon}h_1^p) = \frac{1}{2}h_1^{1+d/2}$. Furthermore, for all $h \in (0, h_1]$ we have the estimate*

$$\max_{\vartheta\in[0,T]} \|e_h(\vartheta)\|^2 + \int_0^T \varepsilon\|e_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(e_h(\vartheta), e_h(\vartheta)\big)\,\mathrm{d}\vartheta \leq C_T^2\big(h^{2p+1} + \varepsilon h^{2p}\big). \tag{80}$$

Proof. We have $p > 1 + d/2$, therefore h_1 is uniquely determined and $C_T(h^{p+1/2} + \sqrt{\varepsilon}h^p) \le \frac{1}{2}h^{1+d/2}$ for all $h \in (0, h_1]$. We fix an arbitrary $h \in (0, h_1]$. By Remark 7, there exists $\delta > 0$, such that if $s, \bar{s} \in [0, T], |s - \bar{s}| \le \delta$, then $\|e_h(s) - e_h(\bar{s})\| \le \frac{1}{2}h^{1+d/2}$.

We define $t_i = i\delta$, $i = 0, 1, \dots$ and set $N := \max\{i = 0, 1, \dots; t_i < T\}$, $t_{N+1} := T$. This defines a partition $0 = t_0 < t_1 < \cdots < t_{N+1} = T$ of the interval $[0, T]$ into $N + 1$ intervals of length (at most) δ.

We shall now prove by induction that for all $n = 1, \dots, N + 1$

$$\max_{\vartheta\in[0,t_n]} \|e_h(\vartheta)\|^2 + \int_0^{t_n} \varepsilon\|e_h(\vartheta)\|_{DG}^2 + \widetilde{J}_h\big(e_h(\vartheta), e_h(\vartheta)\big)\,\mathrm{d}\vartheta \le C_T^2\big(h^{2p+1} + \varepsilon h^{2p}\big). \quad (81)$$

Inequality (80) is thus obtained by taking $n := N + 1$ in (81).
(i) $n = 1$: We know that $\|e_h(0)\| = \|\eta_h(0)\| \le C_T(h^{p+1/2} + \sqrt{\varepsilon}h^p) \le \frac{1}{2}h^{1+d/2}$. By uniform continuity, we have for all $s \in [0, t_1]$

$$\|e_h(s)\| \le \|e_h(0)\| + \|e_h(s) - e_h(0)\| \le \tfrac{1}{2}h^{1+d/2} + \tfrac{1}{2}h^{1+d/2} = h^{1+d/2}.$$

Therefore, $u_h(s) \in \mathcal{U}_h^{ad}(s)$ for all $s \in [0, t_1]$ and by Lemma 23 we obtain estimate (81) on $[0, t_1]$.
(ii) Induction step: We assume that (81) holds for a general $n < N + 1$. Therefore $\|e_h(t_n)\| \le C_T(h^{p+1/2} + \sqrt{\varepsilon}h^p) \le \frac{1}{2}h^{1+d/2}$. By uniform continuity, we have that for all $s \in [t_n, t_{n+1}]$

$$\|e_h(s)\| \le \|e_h(t_n)\| + \|e_h(s) - e_h(t_n)\| \le \tfrac{1}{2}h^{1+d/2} + \tfrac{1}{2}h^{1+d/2} = h^{1+d/2}.$$

This and the induction assumption imply that $\|e_h(s)\| \le h^{1+d/2}$ for $s \in [0, t_n] \cup [t_n, t_{n+1}] = [0, t_{n+1}]$. Therefore, $u_h(s) \in \mathcal{U}_h^{ad}(s)$ for all $s \in [0, t_{n+1}]$ and by Lemma 23, we obtain estimate (81) on $[0, t_{n+1}]$. □

Remark 15. *If we assume* $\mathbf{f} \in (C^3(\mathbb{R}))^d$ *then we may derive a "local" version of estimate (17) and Lemma 8 similarly as we proved Lemmas 21 and 22. In this case, we would define the admissible set as* $\mathcal{U}_h^{ad}(t) := \{v \in S_h; \|u(t) - v\| \le h^{(1+d)/2}\}$ *and in Theorem 24 we would get the improved assumption* $p > (1 + d)/2$, *or* $p > d/2$ *if* $\varepsilon = 0$.

Remark 16. *Assuming only* $\mathbf{f} \in (C^2(\mathbb{R}))^d$, *we may obtain error estimates for the implicit scheme simply by ensuring that the continuated error* $\tilde{e}_h(t) \in \mathcal{U}_h^{ad}(t)$ *on* $[0, T]$. *As in Section 9.2, we first prove a "local" version of Lemma 18 under the assumption* $\tilde{e}_h(\vartheta) \in \mathcal{U}_h^{ad}(\vartheta)$ *for all* $\vartheta \in [0, t]$. *Then we use continuous mathematical induction to go from* $t = 0$ *to* $t = T$ *as in Theorem 24.*

10 Conclusion

We have presented an analysis of the discontinuous Galerkin finite element method for a nonlinear singularly perturbed convection-diffusion problem on quasi-uniform triangulations. Building on results from [44], which dealt with an explicit time discretization, we proved apriori $L^\infty(L^2)$ error estimates independent of the diffusion coefficient $\varepsilon \ge 0$ for the method of lines and a fully implicit scheme. The derived estimates are suboptimal in the $L^\infty(L^2)$-norm, but for $\varepsilon > 0$ also a DG energy norm is included in the estimates, which is estimated optimally, but degenerates for $\varepsilon \to 0$.

- We have extended the key estimate of the convective term from [44] to the case of mixed Dirichlet-Neumann conditions if $\mathbf{f} \in (C_b^2(\mathbb{R}))^d$. An improved estimate for $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$ is obtained for Dirichlet boundary conditions.

- Using these estimates and the apriori assumption $\|e_h(t)\|_\infty = O(h)$, we prove that if $p > 1 + d/2$, then $\|e_h\|_{L^\infty(L^2)} \leq C_T(h^{p+1/2} + \varepsilon h^p)$ with C_T independent of the diffusion coefficient ε. Using continuous mathematical induction we eliminate the apriori assumption.

- We show that for the method of lines the same estimate can be obtained directly using a nonlinear Gronwall lemma.

- For a fully implicit scheme we show that the error equation and the considered estimates of its individual terms do not imply the desired error estimate. Hence, more information about the discrete problem is needed to proceed with the analysis.

- Using an appropriate auxiliary problem derived from the implicit scheme, we introduce a suitable continuation $\tilde{u}_h$ with respect to time of the discrete solution u_h^n. Using continuous mathematical induction we prove error estimates for $\tilde{u}_h$, which imply estimates for u_h^n.

- For the first order implicit scheme we have that if $p > 1+d/2$, then $\sup_{n=0,\cdots,N} \|e_h^n\| \leq C_T(h^{p+1/2} + \varepsilon h^p + \tau)$ with C_T independent of ε. This is proved under the rather restrictive CFL-like condition $\tau = O(h^{1+d/2})$.

- For $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$, we obtain the derived estimates under the assumption $p > (1 + d)/2$ and the less restrictive CFL condition $\tau = O(h^{(1+d)/2})$. For $\varepsilon = 0$ this improves to $p > d/2$.

- Finally, we extend the obtained results to the *locally Lipschitz* case $\mathbf{f} \in (C^2(\mathbb{R}))^d$ and $\mathbf{f} \in (C^3(\mathbb{R}))^d$ using continuous mathematical induction directly, without the need to modify the original problem as in [44].

There are several open problems connected with the presented analysis:

- Eliminating the order condition $p > 1 + d/2$ and $p > (1 + d)/2$, respectively.

- Eliminating the unnatural CFL condition $\tau = O(h^{1+d/2})$ and $\tau = O(h^{(1+d)/2})$, respectively.

- Obtaining optimal error estimates of the order $O(h^{p+1/2} + \varepsilon h^{p+1})$ using some form of the Aubin-Nitsche technique.

- The extension to nonlinear diffusion terms using e.g. the estimates derived in [32] for a diffusion of the form $-\mathrm{div}\big(\beta(u)\nabla u\big)$.

Acknowledgements This work is a part of the research project P201/11/P414 of the Czech Science Foundation. The author is a junior researcher in the University Centre for Mathematical Modelling, Applied Analysis and Computational Mathematics (Math MAC).

References

[1] Arnold, D. N.: An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, **19**, 742–760 (1982).

[2] Arnold, D. N., Brezzi, F., Cockburn, B., Marini, D.: Discontinuos Galerkin methods for elliptic problems, in *Discontinuous Galerkin methods. Theory, Computation and Applications. Lecture Notes in Computational Science and Engineering 11* (Eds. B.Cockburn et al.), Springer, Berlin, 89–101 (2000).

[3] Arnold, D. N., Brezzi, F., Cockburn B., Marini, D.: Unified analysis of discontinuous Galerkin methods for elliptic problems, *SIAM J. Numer. Anal.*, **39**, 1749–1779 (2001).

[4] Barth, T., Ohlberger, M.: *Finite Volume Methods: Foundation and Analysis, Encyclopedia of Computational Mechanics*, volume 1, John Wiley & Sons, Chichester, New York, Brisbane, 439–474 (2004).

[5] Babuška, I., Baumann, C. E., Oden, J. T. A discontinuous *hp* finite element method for diffusion problems, 1-D analysis, *Comput. Math. Appl.*, **37**, 103–122 (1999).

[6] Bassi, F., Rebay, S.: A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations, *J. Comput. Phys.*, **131**, 267–279 (1997).

[7] Bassi, F., Rebay, S.: High-order accurate discontinuous finite element solution of the 2D Euler equations, *J. Comput. Phys.*, **138**, 251–285 (1997).

[8] Baumann, C. E., Oden, J. T.: A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations, *Int. J. Numer. Methods* Fluids, **31**, 79–95 (1999).

[9] Brenner, S.C., Scott L.R.: *The mathematical theory of finite element methods,* Springer Verlag, New York (1994).

[10] Burman, E., Ern, A., Fernández, M.: Explicit RungeKutta Schemes and Finite Elements with Symmetric Stabilization for First-Order Linear PDE Systems, *SIAM J. Numer. Anal.*, **48** (6), 2019–2042 (2010).

[11] Chao, Y. R.: A note on "Continuous mathematical induction", *Bull. Amer. Math. Soc.*, **26** (1), 17–18 (1919).

[12] Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam (1979).

[13] Cockburn, B., Shu, C. W.: TVB Runge–Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II. General framework, *Math. Comp.*, **52**, 411–435 (1989).

[14] Dawson, C., Aizinger, V.: A discontinuous Galerkin method for three-dimensional shallow water equations, *J. Sci. Comput.*, **22-23**, 245–267 (2005).

[15] Dolejší, V., Feistauer, M.: A semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow, *J. Comput. Phys.*, **198**, 727–746 (2004).

[16] Dolejší, V., Feistauer, M.: Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems, *Numer. Funct. Anal. Optimiz.,* **26**, 349–383 (2005).

[17] Dolejší, V., Feistauer, M., Havle, O.: Discontinuous Galerkin method for nonlinear convection-diffusion problems with mixed dirichlet-neumann boundary conditions, *Appl. Math.*, **55**(5), 353–372 (2010).

[18] Dolejší, V., Feistauer, M., Hozman, J.: Analysis of semi-implicit DGFEM for nonlinear convection–diffusion problems on nonconforming meshes, *Comput. Methods Appl. Mech. Engrg.,* **197**, 2813–2827 (2007).

[19] Dolejší, V., Feistauer, M., Kučera, V.: On the Discontinuous Galerkin method for the Simulation of Compressible Flow with wide range of Mach numbers , *Computing and Visualization in Science*, **10**, 17–27(2007).

[20] Dolejší, V., Feistauer, M., Kučera, V., Sobotíková, V.: L-infinity(L-2)-error estimates for the DGFEM applied to convection-diffusion problems on nonconforming meshes, *J. Numer. Math.,* **17** (1), 45–65 (2009).

[21] Dolejší, V., Feistauer, M., Cchwab, C.: A finite volume discontinuous Galerkin scheme for nonlinear convectiondiffusion problems, *Calcolo,* **39**, 1–40 (2002).

[22] Feistauer, M.: A remark to the DGFEM for nonlinear convection-diffusion problems applied on nonconforming meshes, *Numerical Mathematics and Advanced Applications*, Proc. of ENUMATH 2007, Kunisch, K., Of, G., Steinbach, O., Springer, Heidelberg, 323–330 (2008).

[23] Feistauer, M., Kučera, V.: On a robust discontinuous Galerkin technique for the solution of compressible flow, *J. Comput. Phys.,* **224**, 208–221 (2007).

[24] Hartmann, R., Houston, P.: Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations, *Technical Report 2001-42 (SFB 359)*, IWR Heidelberg (2001).

[25] Houston, P., Perugia, I., Schötzau, D.: Mixed discontinuous Galerkin approximation of the Maxwell operator, Technical Report 2002/45, *University of Leicester, Department of Mathematics*, (SIAM J. Numer. Anal., **42**, 434–459 (2002).

[26] Houston, P., Schwab, C., Süli, E.: Discontinuous hp-finite element methods for advection-diffusion-reaction problems, *SIAM J. Numer. Anal.*, **39** (6), 2133–2163 (2002).

[27] Hu, C., Shu, C. W.: A discontinuos Galerkin finite element method for Hamilton-Jacobi equations, *SIAM J. Sci. Comput.*, **21**, 666–690 (1999).

[28] Jaffre, J., Johnson, C., Szepessy, A.: Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws, *Math. Models Methods Appl. Sci.*, **5**, 367–386 (1995).

[29] Johnson, C., Pitkäranta, J.: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation, *Math. Comp.*, **46**, 1–26 (1986).

[30] Karakashian, O., Makridakis, C.: A space-time finite element method for the nonlinear Schrödinger equation: the discontinuous Galerkin method, *Math. Comput.*, **67**, 479–499 (1998).

[31] Kufner, A., John O., Fučík, S.: *Function Spaces*, Academia, Prague (1977).

[32] Kučera, V.: Optimal $L^\infty(L^2)$-error Estimates for the DG Method Applied to Nonlinear Convection-Diffusion Problems with Nonlinear Diffusion, *Numerical Functional Analysis and Optimization*, **31**(3), 285-312 (2010).

[33] Le Saint, P., Raviart, P.–A.: On a finite element method for solving the neutron transport equation, in Mathematical Aspects of Finite Elements in Partial Differential Equations (Ed. C. de Boor), Academic Press, 89–145 (1974).

[34] Osher, S.: Riemann solvers, the entropy condition, and difference approximations, *SIAM. J. Numer. Anal.*, **21**, 217–235 (1984).

[35] Reed, W. H., Hill, T. R.: Triangular mesh methods for the neutron transport equation, *Technical Report LA–UR–73–479*, Los Alamos Scientific Laboratory (1973).

[36] Schötzau, D., Schwab, C., Toselli, A.: Mixed hp-DGFEM for incompressible flows, *SIAM J. Numer. Anal.*, **40**, 2171–2194 (2003).

[37] Sun, S., Wheeler, M.F.: $L^2(H^1)$-norm a posteriori error estimation for discontinuous Galerkin approximations of reactive transport problems, *J. Sci. Comput.*, **22-23**, 501–530 (2005).

[38] Feistauer, M., Hájek, J., Švadlenka, K.: Space-time discontinuos Galerkin method for solving nonstationary convection-diffusion-reaction problems, *Appl. Math.* **52**(3), 197–233 (2007).

[39] Feistauer, M., Švadlenka, K.: Discontinuous Galerkin method of lines for solving nonstationary singularly perturbed linear problems, *J. Numer. Math.* 12, 97–118(2004).

[40] Toselli, A.: HP discontinuous Galerkin approximations for the Stokes problem, *Math. Models Methods Appl. Sci.*, **12**, 1565–1597 (2002).

[41] van der Vegt, J.J.W, Van der Ven, H.: Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flow, part I. General formulation, *J. Comput. Phys.*, **182**, 546–585 (2002).

[42] Wheeler, M. F.: An elliptic collocation-finite element method with interior penalties, *SIAM J. Numer. Anal.*, **15**, 152–161 (1978).

[43] Zeidler, E.: *Nonlinear Functional Analysis and Its Applications II/B: Nonlinear Monotone Operators*, Springer (1986).

[44] Zhang, Q., Shu, C.-W.: Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws, *SIAM J. Numer. Anal.*, **42**(2), 641–666 (2004).

Simulation of compressible viscous flow in time-dependent domains

Jan Česenek[1], Miloslav Feistauer[1], Jaromír Horáček[2], Václav Kučera[1], and Jaroslava Prokopová[1]

[1]*Charles University Prague, Faculty of Mathematics and Physics,, Sokolovská 83, 186 75 Praha 8, Czech Republic*
[2]*Institute of Thermomechanics, Academy of Sciences of the Czech Republic,, Dolejškova 5, 182 00 Praha 8, Czech Republic*

Published in March 2013,
Applied Mathematics and Computation

Abstract

The paper is concerned with the simulation of viscous compressible flow in time dependent domains. The dependence on time of the domain occupied by the fluid is taken into account with the aid of the ALE (Arbitrary Lagrangian-Eulerian) formulation of the compressible Navier-Stokes equations. They are discretized by the discontinuous Galerkin finite element method using piecewise polynomial discontinuous approximations. The time discretization is based on a semi-implicit linearized scheme, which leads to the solution of a linear algebraic system on each time level. A suitable treatment of boundary conditions and shock capturing are used, allowing the solution of flow with a wide range of Mach numbers. The applicability of the developed method is demonstrated by computational results obtained for compressible viscous flow in a channel with moving walls and flow induced airfoil vibrations.

Keywords: compressible Navier-Stokes equations; time dependent domain; ALE method, discontinuous Galerkin method; semi-implicit time discretization; boundary conditions, shock indicator; artificial viscosity; flow in a channel with moving walls; fluid-structure interaction, flow induced airfoil vibrations.

1 Introduction

The interaction of fluid flow with vibrating bodies plays a significant role in many areas of science and technology. We mention, for example, development of airplanes (vibrations of wings) or turbines (blade vibrations), some problems from civil engineering (interaction of wind with constructions as bridges, TV towers or cooling towers of power stations), car industry (vibration of various elements of a carosery), but also in medicine (hemodynamics or flow in the glottis with vibrating vocal folds). In a number of these examples the moving medium is a gas, i.e. compressible flow. For low Mach number flows incompressible models are used (as e.g. in [1], [14]), but in some cases compressibility plays an important role.

The solution of fluid-structure interaction requires the coupling of the solution of equations describing the fluid flow with equations describing the structural behaviour.

Due to the deformation and/or vibrations of structures, the computational domain is time dependent. There exist several techniques of the solution of incompressible flow in time dependent domains. See, e.g. [1], [14] and references therein. The numerical simulation of compressible flow is much more difficult, particularly in time dependent domains. It is necessary to overcome difficulties caused by nonlinear convection dominating over diffusion, which leads to boundary layers and wakes for large Reynolds numbers and shock waves and contact discontinuities for high Mach numbers and instabilities caused by acoustic effects for low Mach numbers.

It appears that a suitable numerical method for the solution of compressible flow is the discontinuous Galerkin finite element method (DGFEM). It employs piecewise polynomial approximations without any requirement on the continuity on interfaces between neighbouring elements. The DGFEM was used for the numerical simulation of the compressible Euler equations, for example, by Bassi and Rebay in [2], where the space DG discretization was combined with explicit Runge-Kutta time discretization. In [3] Baumann and Oden describe an *hp* version of the space DG discretization with explicit time stepping to compressible flow. Van der Vegt and van der Ven apply space-time discontinuous Galerkin method to the solution of the Euler equations in [15], where the discrete problem is solved with the aid of a multigrid accelerated pseudo-time-integration. The papers [6] and [9] are concerned with a semi-implicit DGFEM technique for the solution of inviscid compressible flow, which is unconditionally stable. In [11], this method was extended so that the resulting scheme is robust with respect to the magnitude of the Mach number. The paper [5] is concerned with discontinuous Galerkin method for viscous compressible flow.

The goal of our research is the numerical simulation of interaction of compressible flow with structures as, e.g. flow induced airfoil vibrations or the flow past a vibrating elastic wall. We are concerned with the generalization of the method from [11], [9] and [5] to the solution of compressible inviscid and viscous flow in time dependent domains. The main ingredients of the method is the discontinuous Galerkin space semidiscretization of the governing equations written in the ALE (arbitrary Lagrangian-Eulerian, cf. [14]) form, semi-implicit time discretization, suitable treatment of boundary conditions and the shock capturing avoiding Gibbs phenomenon at discontinuities. Numerical experiments prove the applicability of the method.

2 Formulation of the problem

We shall be concerned with the numerical solution of compressible flow in a bounded domain $\Omega_t \subset \mathbb{R}^2$ depending on time $t \in [0, T]$. Let the boundary of Ω_t consist of three different parts: $\partial\Omega_t = \Gamma_I \cup \Gamma_O \cup \Gamma_{W_t}$, where Γ_I is the inlet, Γ_O is the outlet and Γ_{W_t} denotes impermeable walls that may move in dependence on time.

The system describing compressible flow, consisting of the continuity equation, the Navier-Stokes equations and the energy equation, can be written in the form

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = \sum_{s=1}^{2} \frac{\partial \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_s}, \tag{1}$$

where

$$\begin{aligned}
\boldsymbol{w} &= (w_1,\dots,w_4)^{\mathcal{T}} = (\rho,\rho v_1,\rho v_2,E)^{\mathcal{T}} \in I\!R^4, \\
\boldsymbol{w} &= \boldsymbol{w}(x,t),\ x\in\Omega_t,\ t\in(0,T), \\
\boldsymbol{f}_i(\boldsymbol{w}) &= (f_{i1},\cdots,f_{i4})^{\mathcal{T}} = (\rho v_i,\rho v_1 v_i+\delta_{1i}\,p,\rho v_2 v_i+\delta_{2i}\,p,(E+p)v_i)^{\mathcal{T}}, \\
\boldsymbol{R}_i(\boldsymbol{w},\nabla\boldsymbol{w}) &= (R_{i1},\dots,R_{i4})^{\mathcal{T}} = \left(0,\tau^V_{i1},\tau^V_{i2},\tau^V_{i1}\,v_1+\tau^V_{i2}\,v_2+k\partial\theta/\partial x_i\right)^{\mathcal{T}}, \\
\tau^V_{ij} &= \lambda\,\mathrm{div}\boldsymbol{v}\,\delta_{ij}+2\mu\,d_{ij}(\boldsymbol{v}),\ d_{ij}(\boldsymbol{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right).
\end{aligned} \tag{2}$$

We use the following notation: ρ – density, p – pressure, E – total energy, $\boldsymbol{v}=(v_1,v_2)$ – velocity, θ – absolute temperature, $\gamma>1$ – Poisson adiabatic constant, $c_v>0$ – specific heat at constant volume, $\mu>0, \lambda=-2\mu/3$ – viscosity coefficients, k – heat conduction. The vector-valued function $\boldsymbol{w}$ is called state vector, the functions $\boldsymbol{f}_i$ are the so-called inviscid fluxes and $\boldsymbol{R}_i$ represent viscous terms.

The above system is completed by the thermodynamical relations

$$p=(\gamma-1)(E-\rho|\boldsymbol{v}|^2/2),\quad \theta=\Big(\frac{E}{\rho}-\frac{1}{2}|\boldsymbol{v}|^2\Big)/c_v. \tag{3}$$

The resulting system is equipped with the initial condition

$$\boldsymbol{w}(x,0)=\boldsymbol{w}^0(x),\quad x\in\Omega_0, \tag{4}$$

and the following boundary conditions:

$$\begin{aligned}
&\text{a)}\ \ \rho|_{\Gamma_I}=\rho_D,\quad \text{b)}\ \ \boldsymbol{v}|_{\Gamma_I}=\boldsymbol{v}_D=(v_{D1},v_{D2})^{\mathrm{T}}, \\
&\text{c)}\ \ \sum_{i,j=1}^{2}\tau^V_{ij}n_iv_j+k\frac{\partial\theta}{\partial n}=0\quad \text{on } \Gamma_I, \\
&\text{d)}\ \ \boldsymbol{v}|_{\Gamma_{W_t}}=\boldsymbol{z}_D=\text{velocity of a moving wall},\ \text{e)}\ \ \frac{\partial\theta}{\partial n}|_{\Gamma_{W_t}}=0 \text{ on } \Gamma_{W_t}, \\
&\text{f)}\ \ \sum_{i=1}^{2}\tau^V_{ij}n_i=0,\quad j=1,\,2,\quad \text{g)}\ \ \frac{\partial\theta}{\partial n}=0 \text{ on } \Gamma_O.
\end{aligned} \tag{5}$$

In order to treat the time dependence of the domain, we use the so-called *arbitrary Lagrangian-Eulerian* ALE technique, see e.g. [12]. We define a reference domain Ω_0 and introduce a regular one-to-one ALE mapping of Ω_0 onto Ω_t:

$$\mathcal{A}_t:\overline{\Omega}_0\longrightarrow\overline{\Omega}_t,\ \textit{i.e.}\ \boldsymbol{X}\in\overline{\Omega}_0\longmapsto\boldsymbol{x}=\boldsymbol{x}(\boldsymbol{X},t)=\mathcal{A}_t(\boldsymbol{X})\in\overline{\Omega}_t.$$

Here we use the notation $\boldsymbol{X}$ for points in $\overline{\Omega}_0$ and $\boldsymbol{x}$ for points in $\overline{\Omega}_t$.

Further, we define the domain velocity:

$$\begin{aligned}
\tilde{\boldsymbol{z}}(\boldsymbol{X},t)&=\frac{\partial}{\partial t}\mathcal{A}_t(\boldsymbol{X}),\quad t\in[0,T],\ \boldsymbol{X}\in\Omega_0, \\
\boldsymbol{z}(\boldsymbol{x},t)&=\tilde{\boldsymbol{z}}(\mathcal{A}^{-1}(\boldsymbol{x}),t),\quad t\in[0,T],\ \boldsymbol{x}\in\Omega_t
\end{aligned}$$

and the ALE derivative of a function $f=f(\boldsymbol{x},t)$ defined for $\boldsymbol{x}\in\Omega_t$ and $t\in[0,T]$:

$$\frac{D^A}{Dt}f(\boldsymbol{x},t)=\frac{\partial\tilde{f}}{\partial t}(\boldsymbol{X},t), \tag{6}$$

where

$$\tilde{f}(\boldsymbol{X},t) = f(\mathcal{A}_t(\boldsymbol{X}),t),\ \boldsymbol{X}\in\Omega_0,\ \boldsymbol{x}=\mathcal{A}_t(\boldsymbol{X}).$$

As a direct consequence of the chain rule we get the relation

$$\frac{D^A f}{Dt} = \frac{\partial f}{\partial t} + \operatorname{div}(\boldsymbol{z}f) - f\operatorname{div}\boldsymbol{z}.$$

This leads to the ALE formulation of the Navier-Stokes equations

$$\frac{D^A \boldsymbol{w}}{Dt} + \sum_{s=1}^{2}\frac{\partial \boldsymbol{g}_s(\boldsymbol{w})}{\partial x_s} + \boldsymbol{w}\operatorname{div}\boldsymbol{z} = \sum_{s=1}^{2}\frac{\partial \boldsymbol{R}_s(\boldsymbol{w},\nabla\boldsymbol{w})}{\partial x_s}, \tag{7}$$

where

$$\boldsymbol{g}_s(\boldsymbol{w}) := \boldsymbol{f}_s(\boldsymbol{w}) - z_s\boldsymbol{w},\quad s=1,2,$$

are the ALE modified inviscid fluxes.

We see that in the ALE formulation of the Navier-Stokes equations the time derivative $\partial\boldsymbol{w}/\partial t$ is replaced by the ALE derivative $D^A\boldsymbol{w}/Dt$, the inviscid fluxes $\boldsymbol{f}_s$ are replaced by the ALE modified inviscid fluxes $\boldsymbol{g}_s$ and a new additional "reaction" term $\boldsymbol{w}\operatorname{div}\boldsymbol{z}$ appears.

3 Discrete problem

3.1 Discontinuous Galerkin space semidiscretization

For the space semidiscretization we use the discontinuous Galerkin finite element method. We construct a polygonal approximation Ω_{ht} of the domain Ω_t. By $\mathcal{T}_{ht}$ we denote a partition of the closure $\overline{\Omega}_{ht}$ of the domain Ω_{ht} into a finite number of closed triangles K with mutually disjoint interiors such that $\overline{\Omega}_{ht} = \bigcup_{K\in\mathcal{T}_{ht}} K$.

By $\mathcal{F}_{ht}$ we denote the system of all faces of all elements $K\in\mathcal{T}_{ht}$. Further, we introduce the set of all interior faces $\mathcal{F}_{ht}^I = \{\Gamma\in\mathcal{F}_{ht};\ \Gamma\subset\Omega_t\}$, the set of all boundary faces $\mathcal{F}_{ht}^B = \{\Gamma\in\mathcal{F}_{ht};\ \Gamma\subset\partial\Omega_{ht}\}$ and the set of all "Dirichlet" boundary faces $\mathcal{F}_{ht}^D = \{\Gamma\in\mathcal{F}_{ht}^B;$ a Dirichlet condition is prescribed on $\Gamma\}$. Each $\Gamma\in\mathcal{F}_{ht}$ is associated with a unit normal vector $\boldsymbol{n}_\Gamma$ to Γ. For $\Gamma\in\mathcal{F}_{ht}^B$ the normal $\boldsymbol{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega_{ht}$. We set $d(\Gamma)$ = length of $\Gamma\in\mathcal{F}_{ht}$.

For each $\Gamma\in\mathcal{F}_{ht}^I$ there exist two neighbouring elements $K_\Gamma^{(L)}, K_\Gamma^{(R)}\in\mathcal{T}_h$ such that $\Gamma\subset\partial K_\Gamma^{(R)}\cap\partial K_\Gamma^{(L)}$. We use the convention that $K_\Gamma^{(R)}$ lies in the direction of $\boldsymbol{n}_\Gamma$ and $K_\Gamma^{(L)}$ lies in the opposite direction to $\boldsymbol{n}_\Gamma$. The elements $K_\Gamma^{(L)}$, $K_\Gamma^{(R)}$ are called neighbours. If $\Gamma\in\mathcal{F}_{ht}^B$, then the element adjacent to Γ will be denoted by $K_\Gamma^{(L)}$.

The approximate solution will be sought in the space of piecewise polynomial functions

$$\boldsymbol{S}_{ht} = [S_{ht}]^4,\quad \text{with}\ \ S_{ht} = \{v; v|_K\in P_r(K)\ \forall K\in\mathcal{T}_{ht}\}, \tag{8}$$

where $r\geq 0$ is an integer and $P_r(K)$ denotes the space of all polynomials on K of degree $\leq r$. A function $\boldsymbol{\varphi}\in\boldsymbol{S}_{ht}$ is, in general, discontinuous on interfaces $\Gamma\in\mathcal{F}_{ht}^I$. By $\boldsymbol{\varphi}_\Gamma^{(L)}$ and $\boldsymbol{\varphi}_\Gamma^{(R)}$ we denote the values of $\boldsymbol{\varphi}$ on Γ considered from the interior and the exterior of $K_\Gamma^{(L)}$, respectively, and set

$$\langle\boldsymbol{\varphi}\rangle_\Gamma = (\boldsymbol{\varphi}_\Gamma^{(L)} + \boldsymbol{\varphi}_\Gamma^{(R)})/2,\quad [\boldsymbol{\varphi}]_\Gamma = \boldsymbol{\varphi}_\Gamma^{(L)} - \boldsymbol{\varphi}_\Gamma^{(R)}. \tag{9}$$

The discrete problem is derived in the following way: We multiply system (7) by a test function $\boldsymbol{\varphi}_h \in \boldsymbol{S}_{ht}$, integrate over $K \in \mathcal{T}_{ht}$, apply Green's theorem, sum over all elements $K \in \mathcal{T}_{ht}$, use the concept of the numerical flux and introduce suitable terms mutually vanishing for a regular exact solution. In this way we get the following identity:

$$\sum_{K\in\mathcal{T}_{ht}} \int_K \frac{D^A \boldsymbol{w}}{Dt} \cdot \boldsymbol{\varphi}_h \, dx + b_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) + a_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) + J_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) + d_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) = \ell_h(\boldsymbol{w}, \boldsymbol{\varphi}_h).$$

Here

$$b_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) = -\sum_{K\in\mathcal{T}_{ht}} \int_K \sum_{s=1}^{2} \boldsymbol{g}_s(\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \, dx \tag{10}$$
$$+ \sum_{\Gamma\in\mathcal{F}_{ht}^I} \int_\Gamma \mathbf{H}_g(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_\Gamma^{(R)}, \boldsymbol{n}_\Gamma) \cdot [\boldsymbol{\varphi}_h]_\Gamma \, dS + \sum_{\Gamma\in\mathcal{F}_{ht}^B} \int_\Gamma \mathbf{H}_g(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_\Gamma^{(R)}, \boldsymbol{n}_\Gamma) \cdot \boldsymbol{\varphi}_{h\Gamma}^{(L)} \, dS$$

is the convection form, defined with the aid of a numerical flux $\mathbf{H}_g$. We require that it is consistent with the fluxes $\boldsymbol{g}_s$:

$$\mathbf{H}_g(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{2} \boldsymbol{g}_s(\boldsymbol{w}) n_s \quad (\boldsymbol{n} = (n_1, n_2), |\boldsymbol{n}| = 1),$$

conservative:

$$\mathbf{H}_g(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{n}) = -\mathbf{H}_g(\boldsymbol{w}, \boldsymbol{u}, -\boldsymbol{n}),$$

and locally Lipschitz-continuous. The determination of the boundary state $\boldsymbol{w}_\Gamma^{(R)}$ in the case when $\Gamma \subset \partial\Omega_{ht}$ is described in Section 3.4.

Further, we define the viscous form

$$a_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) = \sum_{K\in\mathcal{T}_{ht}} \int_K \sum_{s=1}^{2} \boldsymbol{R}_s(\boldsymbol{w}, \nabla\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \, dx \tag{11}$$
$$- \sum_{\Gamma\in\mathcal{F}_{ht}^I} \int_\Gamma \sum_{s=1}^{2} \langle \boldsymbol{R}_s(\boldsymbol{w}, \nabla\boldsymbol{w}) \rangle_\Gamma (\boldsymbol{n}_\Gamma)_s \cdot [\boldsymbol{\varphi}_h]_\Gamma \, dS$$
$$- \sum_{\Gamma\in\mathcal{F}_{ht}^D} \int_\Gamma \sum_{s=1}^{2} \boldsymbol{R}_s(\boldsymbol{w}, \nabla\boldsymbol{w}) (\boldsymbol{n}_\Gamma)_s \cdot \boldsymbol{\varphi}_{h\Gamma}^{(L)} \, dS,$$

(we use the incomplete discretization of viscous terms - the so-called IIPG version), the interior and boundary penalty terms and the right-hand side form, respectively,

$$J_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_{ht}^I} \int_\Gamma \sigma [\boldsymbol{w}]_\Gamma \cdot [\boldsymbol{\varphi}_h]_\Gamma \, dS + \sum_{\Gamma\in\mathcal{F}_{ht}^D} \int_\Gamma \sigma \boldsymbol{w} \cdot \boldsymbol{\varphi}_{h\Gamma}^{(L)} \, dS, \tag{12}$$

$$\ell_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_{ht}^D} \int_\Gamma \sum_{s=1}^{2} \sigma \boldsymbol{w}_B \cdot \boldsymbol{\varphi}_{h\Gamma}^{(L)} \, dS. \tag{13}$$

Here $\sigma|_\Gamma = C_W \mu/d(\Gamma)$ and $C_W > 0$ is a sufficiently large constant. The reaction form reads

$$d_h(\boldsymbol{w}, \boldsymbol{\varphi}_h) = \sum_{K\in\mathcal{T}_{ht}} \int_K (\boldsymbol{w}\cdot\boldsymbol{\varphi}_h)\,\mathrm{div}\boldsymbol{z}\,dx. \tag{14}$$

The boundary state $\boldsymbol{w}_B$ is defined on the basis of the Dirichlet boundary conditions (5), a), b), d) and extrapolation:

$$\boldsymbol{w}_B = (\rho_D, \rho_D v_{D1}, \rho_D v_{D2}, c_v\rho_D\theta_\Gamma^{(L)} + \frac{1}{2}\rho_D|\boldsymbol{v}_D|^2) \quad \text{on } \Gamma_I, \tag{15}$$

$$\boldsymbol{w}_B = \boldsymbol{w}_\Gamma^{(L)} \quad \text{on } \Gamma_O, \tag{16}$$

$$\boldsymbol{w}_B = (\rho_\Gamma^{(L)}, \rho_\Gamma^{(L)} z_{D1}, \rho_\Gamma^{(L)} z_{D2}, c_v\rho_\Gamma^{(L)}\theta_\Gamma^{(L)} + \frac{1}{2}\rho_\Gamma^{(L)}|\boldsymbol{z}_D|^2) \quad \text{on } \Gamma_{W_t}. \tag{17}$$

The approximate solution is defined as $\boldsymbol{w}_h(t) \in \boldsymbol{S}_{ht}$ such that

$$\sum_{K\in\mathcal{T}_{ht}} \int_K \frac{D^A\boldsymbol{w}_h(t)}{Dt}\cdot\boldsymbol{\varphi}_h\,dx + b_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) + a_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) + J_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) + d_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) = \ell_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h)$$

holds for all $\boldsymbol{\varphi}_h \in \boldsymbol{S}_{ht}$, all $t \in (0,T)$ and $\boldsymbol{w}_h(0) = \boldsymbol{w}_h^0$ is an approximation of the initial state $\boldsymbol{w}^0$.

3.2 Time discretization

Let us construct a partition $0 = t_0 < t_1 < t_2 \ldots$ of the time interval $[0,T]$ and define the time step $\tau_k = t_{k+1} - t_k$. We use the approximations $\boldsymbol{w}_h(t_n) \approx \boldsymbol{w}_h^n \in \boldsymbol{S}_{ht_n}$, $\boldsymbol{z}(t_n) \approx \boldsymbol{z}^n$, $n = 0, 1, \ldots$ and introduce the function $\hat{\boldsymbol{w}}_h^k = \boldsymbol{w}_h^k \circ \mathcal{A}_{t_k} \circ \mathcal{A}_{t_{k+1}}^{-1}$, which is defined in the domain $\Omega_{ht_{k+1}}$. In order to approximate the ALE derivative at time t_{k+1}, we start from its definition and then use the backward difference:

$$\begin{aligned}\frac{D^A\boldsymbol{w}_h}{Dt}(x,t_{k+1}) &= \frac{\partial\tilde{\boldsymbol{w}}_h}{\partial t}(X,t_{k+1})\\ &\approx \frac{\tilde{\boldsymbol{w}}_h^{k+1}(X) - \tilde{\boldsymbol{w}}_h^k(X)}{\tau_k} = \frac{\boldsymbol{w}_h^{k+1}(x) - \hat{\boldsymbol{w}}_h^k(x)}{\tau_k}, \quad x = \mathcal{A}_{t_{k+1}}(X) \in \Omega_{ht_{k+1}}.\end{aligned}$$

By the symbol $(\cdot,\cdot)$ we shall denote the scalar product in $L^2(\Omega_{ht_{k+1}})$. A possible full discretization reads: We seek $\boldsymbol{w}_h^{k+1} \in \boldsymbol{S}_{ht_{k+1}}$ such that for all $\boldsymbol{\varphi}_h \in \boldsymbol{S}_{ht_{k+1}}$, $k = 0, 1, \ldots$,

$$\Big(\frac{\boldsymbol{w}_h^{k+1} - \hat{\boldsymbol{w}}_h^k}{\tau_k}, \boldsymbol{\varphi}_h\Big) + b_h(\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h) + a_h(\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h) + J_h(\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h) + d_h(\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h) = \ell_h(\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h). \tag{18}$$

Scheme (18) is of the first-order in time. In order to increase the accuracy in the time discretization, it is possible to use the second-order backward difference formula (BDF) for the approximation of the ALE derivative:

$$\frac{D^A\boldsymbol{w}_h}{Dt}(t_{k+1}) \approx \frac{2\tau_k + \tau_{k-1}}{\tau_k(\tau_k+\tau_{k-1})}\boldsymbol{w}_h^{k+1} - \frac{\tau_k+\tau_{k-1}}{\tau_k\tau_{k-1}}\hat{\boldsymbol{w}}_h^k + \frac{\tau_k}{\tau_{k-1}(\tau_k+\tau_{k-1})}\hat{\boldsymbol{w}}_h^{k-1}. \tag{19}$$

The problem for obtaining $\boldsymbol{w}_h^{k+1}$ from (18) is equivalent to a strongly nonlinear algebraic system and its solution is rather difficult.

Our goal is to develop a numerical scheme, which would be accurate and robust, with good stability properties and efficiently solvable. Therefore, we proceed similarly as in [6] and use a partial linearization of the forms b_h and a_h. This approach leads to a scheme that requires the solution of only one large sparse linear system on each time level. The efficiency and accuracy of this technique was analyzed in [6], [9] and [11] in the case of inviscid flow, where advantages of this method are demonstrated by numerical experiments. Here we show that this technique can be extended to the solution of viscous compressible flow in time-dependent domains with applications to fluid-structure interaction.

The linearization of the first term of the form b_h is based on the relations

$$\boldsymbol{g}_s(\boldsymbol{w}_h^{k+1}) = (\mathbb{A}_s(\boldsymbol{w}_h^{k+1}) - z_s^{k+1}\mathbb{I})\boldsymbol{w}_h^{k+1} \approx (\mathbb{A}_s(\overline{\boldsymbol{w}}_h^{k+1}) - z_s^{k+1}\mathbb{I})\boldsymbol{w}_h^{k+1},$$

where $\mathbb{A}_s(\mathbf{w})$ is the Jacobi matrix of $\boldsymbol{f}_s(\mathbf{w})$, cf. [10]. By $\overline{\boldsymbol{w}}_h^{k+1}$ we define the state obtained by the extrapolation:

$$\overline{\boldsymbol{w}}_h^{k+1} = \hat{\boldsymbol{w}}_h^k \ \text{ and } \overline{\boldsymbol{w}}_h^{k+1} = \frac{\tau_k + \tau_{k-1}}{\tau_{k-1}}\hat{\boldsymbol{w}}_h^k - \frac{\tau_k}{\tau_{k-1}}\hat{\boldsymbol{w}}_h^{k-1} \tag{20}$$

in the case of the first-order time discretization and second-order time discretization, respectively.

The second term of b_h is linearized with the aid of the Vijayasundaram numerical flux (cf. [16]) defined in the following way. Taking into account the definition of $\boldsymbol{g}_s$, we have

$$\frac{D\boldsymbol{g}_s(\boldsymbol{w})}{D\boldsymbol{w}} = \frac{D\boldsymbol{f}_s(\boldsymbol{w})}{D\boldsymbol{w}} - z_s\mathbb{I} = \mathbb{A}_s(\boldsymbol{w}) - z_s\mathbb{I}, \tag{21}$$

and can write

$$\mathbb{P}_g(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{2} \frac{D\boldsymbol{g}_s(\boldsymbol{w})}{D\boldsymbol{w}} n_s = \sum_{s=1}^{2} \left(\mathbb{A}_s(\boldsymbol{w})n_s - z_s n_s \mathbb{I}\right). \tag{22}$$

By [10], this matrix is diagonalizable. It means that there exists a nonsingular matrix $\mathbb{T} = \mathbb{T}(\boldsymbol{w}, \boldsymbol{n})$ such that

$$\mathbb{P}_g = \mathbb{T}\mathbb{\Lambda}\mathbb{T}^{-1}, \quad \mathbb{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_4), \tag{23}$$

where $\lambda_i = \lambda_i(\boldsymbol{w}, \boldsymbol{n})$ are eigenvalues of the matrix $\mathbb{P}_g$. Now we define the "positive" and "negative" parts of the matrix $\mathbb{P}_g$ by

$$\mathbb{P}_g^{\pm} = \mathbb{T}\mathbb{\Lambda}^{\pm}\mathbb{T}^{-1}, \quad \mathbb{\Lambda}^{\pm} = \mathrm{diag}(\lambda_1^{\pm}, \ldots, \lambda_4^{\pm}), \tag{24}$$

where $\lambda^+ = \max(\lambda, 0)$, $\lambda^- = \min(\lambda, 0)$. Using the above concepts, we introduce the modified Vijayasundaram numerical flux (cf. [16] or [10]) as

$$\boldsymbol{H}_g(\boldsymbol{w}_L, \boldsymbol{w}_R, \boldsymbol{n}) = \mathbb{P}_g^+\Big(\frac{\boldsymbol{w}_L + \boldsymbol{w}_R}{2}, \boldsymbol{n}\Big)\boldsymbol{w}_L + \mathbb{P}_g^-\Big(\frac{\boldsymbol{w}_L + \boldsymbol{w}_R}{2}, \boldsymbol{n}\Big)\boldsymbol{w}_R. \tag{25}$$

Using the above definition of the numerical flux, we introduce the approximations

$$\mathbf{H}_g(\boldsymbol{w}_{h\Gamma}^{k+1(L)}, \boldsymbol{w}_{h\Gamma}^{k+1(R)}, \boldsymbol{n}_\Gamma) \approx \mathbb{P}_g^+(\langle\overline{\boldsymbol{w}}_h^{k+1}\rangle_\Gamma, \boldsymbol{n}_\Gamma)\boldsymbol{w}_{h\Gamma}^{k+1(L)} + \mathbb{P}_g^-(\langle\overline{\boldsymbol{w}}_h^{k+1}\rangle_\Gamma, \boldsymbol{n}_\Gamma)\boldsymbol{w}_{h\Gamma}^{k+1(R)}.$$

for $\Gamma \in \mathcal{F}_{ht_{k+1}}^I$ and

$$\mathbf{H}_g(\boldsymbol{w}_{h\Gamma}^{k+1(L)}, \boldsymbol{w}_{h\Gamma}^{k+1(R)}, \boldsymbol{n}_\Gamma) \approx \mathbb{P}_g^+(\langle\overline{\boldsymbol{w}}_h^{k}\rangle_\Gamma, \boldsymbol{n}_\Gamma)\boldsymbol{w}_{h\Gamma}^{k+1(L)} + \mathbb{P}_g^-(\langle\overline{\boldsymbol{w}}_h^{k}\rangle_\Gamma, \boldsymbol{n}_\Gamma)\overline{\boldsymbol{w}}_{h\Gamma}^{k+1(R)}.$$

for $\Gamma \in \mathcal{F}^B_{ht_{k+1}}$. In this way we get the form

$$\begin{aligned}
&\hat{b}_h(\overline{\boldsymbol{w}}_h^{k+1}, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) \qquad (26)\\
&= -\sum_{K\in\mathcal{T}_{ht_{k+1}}} \int_K \sum_{s=1}^{2} (\mathbb{A}_s(\overline{\boldsymbol{w}}^{k+1}(x)) - z_s^{k+1}(x))\mathbb{I})\boldsymbol{w}^{k+1}(x))\cdot \frac{\partial \boldsymbol{\varphi}_h(x)}{\partial x_s} dx,\\
&+ \sum_{\Gamma\in\mathcal{F}^I_{ht_{k+1}}} \int_\Gamma \Big(\mathbb{P}_g^+\big(\langle \overline{\boldsymbol{w}}_h^{k+1}\rangle, \boldsymbol{n}_\Gamma\big)\boldsymbol{w}_h^{k+1(L)} + \mathbb{P}_g^-\big(\langle \overline{\boldsymbol{w}}_h^{k+1}\rangle, \boldsymbol{n}_\Gamma\big)\boldsymbol{w}_h^{k+1(R)}\Big)\cdot [\boldsymbol{\varphi}_h] dS\\
&+ \sum_{\Gamma\in\mathcal{F}^B_{ht_{k+1}}} \int_\Gamma \Big(\mathbb{P}_g^+\big(\langle \overline{\boldsymbol{w}}_h^{k+1}\rangle, \boldsymbol{n}_\Gamma\big)\boldsymbol{w}_h^{k+1(L)} + \mathbb{P}_g^-\big(\langle \overline{\boldsymbol{w}}_h^{k+1}\rangle, \boldsymbol{n}_\Gamma\big)\overline{\boldsymbol{w}}_h^{k+1(R)}\Big)\cdot \boldsymbol{\varphi}_h dS.
\end{aligned}$$

For the determination of the boundary state $\overline{\boldsymbol{w}}_h^{k+1(R)}$, if $\Gamma \subset \partial\Omega_{ht_{k+1}}$, we refer to Section 3.4.

The linearization of the form a_h is based on the fact that $\boldsymbol{R}_s(\boldsymbol{w}_h, \nabla \boldsymbol{w}_h)$ is linear in $\nabla \boldsymbol{w}$ and nonlinear in $\boldsymbol{w}$. We get the linearized viscous form

$$\begin{aligned}
\hat{a}_h(\overline{\boldsymbol{w}}_h^{k+1}, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) &= \sum_{K\in\mathcal{T}_{ht_{k+1}}} \int_K \sum_{s=1}^{2} \boldsymbol{R}_s(\overline{\boldsymbol{w}}_h^{k+1}, \nabla \boldsymbol{w}_h^{k+1})\cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s}\, dx \qquad (27)\\
&- \sum_{\Gamma\in\mathcal{F}^I_{ht_{k+1}}} \int_\Gamma \sum_{s=1}^{2} \big\langle \boldsymbol{R}_s(\overline{\boldsymbol{w}}_h^{k+1}, \nabla \boldsymbol{w}^{k+1})\big\rangle (\boldsymbol{n}_\Gamma)_s \cdot [\boldsymbol{\varphi}_h]\, dS\\
&- \sum_{\Gamma\in\mathcal{F}^D_{ht_{k+1}}} \int_\Gamma \sum_{s=1}^{2} \boldsymbol{R}_s(\overline{\boldsymbol{w}}_h^{k+1}, \nabla \boldsymbol{w}_h^{k+1})(\boldsymbol{n}_\Gamma)_s \cdot \boldsymbol{\varphi}_h\, dS.
\end{aligned}$$

3.3 Artificial viscosity

In high-speed gas flow with large Mach numbers, discontinuities - called shock waves or contact discontinuities - appear. In viscous high-speed flow these discontinuities may be smeared due to viscosity and heat conduction. Near shock waves and contact discontinuities, the so-called Gibbs phenomenon, manifested by nonphysical spurious overshoots and undershoots, usually occurs in the numerical solution. In order to avoid this undesirable phenomenon, it is necessary to apply a suitable limiting procedure. Here we use the approach proposed in [11] based on the discontinuity indicator

$$g^k(K) = \int_{\partial K} [\hat{\rho}_h^k]^2 \,\mathrm{d}S \big/ (h_K |K|^{3/4}), \quad K \in \mathcal{T}_{ht_{k+1}}, \qquad (28)$$

introduced in [7]. By $[\hat{\rho}_h^k]$ we denote the jump of the function $\hat{\rho}_h^k$ on the boundary ∂K and $|K|$ denotes the area of the element K. Then we define the discrete discontinuity indicator

$$G^k(K) = 0 \quad \text{if } g^k(K) < 1, \quad G^k(K) = 1 \quad \text{if } g^k(K) \geq 1, \quad K \in \mathcal{T}_{ht_{k+1}}, \qquad (29)$$

and the artificial viscosity forms

$$\hat\beta_h(\hat{\boldsymbol w}_h^k,\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)=\nu_1\sum_{K\in\mathcal T_{ht_{k+1}}} h_K G^k(K)\int_K\nabla\boldsymbol w_h^{k+1}\cdot\nabla\boldsymbol\varphi_h\,\mathrm d\boldsymbol x, \tag{30}$$

$$\hat J_h(\hat{\boldsymbol w}_h^k,\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)=\nu_2\sum_{\Gamma\in\mathcal F^I_{ht_{k+1}}}\frac12\big(G^k(K_\Gamma^{(L)})+G^k(K_\Gamma^{(R)})\big)\int_\Gamma[\boldsymbol w_h^{k+1}]\cdot[\boldsymbol\varphi_h]\,\mathrm d\mathcal S,$$

with parameters $\nu_1,\ \nu_2=O(1)$.

The resulting scheme has the following form: We seek $\boldsymbol w_h^{k+1}\in\boldsymbol S_{ht_{k+1}}$ such that for all $\boldsymbol\varphi_h\in\boldsymbol S_{ht_{k+1}},\ k=0,1,\ldots,$

$$\begin{aligned}&\Big(\frac{\boldsymbol w_h^{k+1}-\hat{\boldsymbol w}_h^k}{\tau_k},\boldsymbol\varphi_h\Big)+\hat b_h(\hat{\boldsymbol w}_h^k,\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)+\hat a_h(\hat{\boldsymbol w}_h^k,\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)+J_h(\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)\\&+d_h(\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)+\hat\beta_h(\hat{\boldsymbol w}_h^k,\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)+\hat J_h(\hat{\boldsymbol w}_h^k,\boldsymbol w_h^{k+1},\boldsymbol\varphi_h)=\ell(\overline{\boldsymbol w}_B^{k+1},\boldsymbol\varphi_h),\end{aligned} \tag{31}$$

in the case of the first-order time discretization. The second-order time discretization is obtained by replacing the expression $(\boldsymbol w_h^{k+1}-\hat{\boldsymbol w}_h^k,\boldsymbol\varphi_h)/\tau_k$ by the approximation (19) and replacing $\hat{\boldsymbol w}_h^k$ in the forms $\hat a_h$ and $\hat b_h$ by $\overline{\boldsymbol w}_h^{k+1}$.

This method successfully overcomes problems with the Gibbs phenomenon in the context of the semi-implicit scheme. It is important that the indicator $G^k(K)$ vanishes in regions, where the solution is regular and, therefore, the numerical solution does not contain any nonphysical entropy production in these regions. If the described artificial viscosity is not applied, then in the case of high-speed flow with shock waves and contact discontinuities the computational process collapses, because negative values of the approximation of the density and pressure appear.

3.4 Treatment of boundary states in the form $\hat b_h$

If $\Gamma\in\mathcal F^B_{ht_{k+1}}$, it is necessary to specify the boundary state $\overline{\boldsymbol w}_{h\Gamma}^{k+1(R)}$ appearing in the numerical flux $\boldsymbol H_g$ in the definition of the inviscid form $\hat b_h$. For simplicity we shall use the notation $\boldsymbol w^{(R)}$ for values of the function $\overline{\boldsymbol w}_{h\Gamma}^{k+1(R)}$ which should be determined at individual integration points on the face Γ. Similarly, $\boldsymbol w^{(L)}$ will denote the values of $\overline{\boldsymbol w}_{h\Gamma}^{k+1(L)}$ at the corresponding points.

On the inlet, which is assumed fixed, we proceed in the same way as in [11], Section 4. Using rotational invariance, we transform the Euler equations

$$\frac{\partial\boldsymbol w}{\partial t}+\sum_{s=1}^2\frac{\partial\boldsymbol f_s(\boldsymbol w)}{\partial x_s}=0$$

to the coordinates $\tilde x_1$, parallel with the normal direction $\boldsymbol n=\boldsymbol n_\Gamma$ to $\partial\Omega$, and $\tilde x_2$, tangential to the boundary. In this way we obtain a system for an unknown function $\boldsymbol q=\boldsymbol Q(\boldsymbol n)\boldsymbol w$, where

$$\boldsymbol Q(\boldsymbol n)=\begin{pmatrix}1, & 0, & 0, & 0\\ 0, & n_1, & n_2, & 0\\ 0, & -n_2, & n_1, & 0\\ 0, & 0, & 0, & 1\end{pmatrix} \tag{32}$$

is the rotational matrix. Now we neglect the derivative with respect to $\tilde{x}_2$ and linearize the system around the state $\boldsymbol{q}^{(L)} = \boldsymbol{Q}(\boldsymbol{n})\boldsymbol{w}^{(L)}$. In this way we obtain the linear system

$$\frac{\partial \boldsymbol{q}}{\partial t} + \mathbb{A}_1(\boldsymbol{q}^{(L)})\frac{\partial \boldsymbol{q}}{\partial \tilde{x}_1} = 0, \tag{33}$$

for the vector-valued function $\boldsymbol{q} = \boldsymbol{q}(\tilde{x}_1, t)$, considered in the set $(-\infty, 0) \times (0, \infty)$ and equipped with the initial and boundary conditions

$$\boldsymbol{q}(\tilde{x}_1, 0) = \boldsymbol{q}^{(L)},\ \tilde{x}_1 < 0, \quad \text{and} \quad \boldsymbol{q}(0, t) = \boldsymbol{q}^{(R)},\ t > 0. \tag{34}$$

The goal is to choose $\boldsymbol{q}^{(R)}$ in such a way that this initial-boundary value problem is well posed, i.e. has a unique solution. The method of characteristics leads to the following process:

Let us put $\boldsymbol{q}^* = \boldsymbol{Q}(\boldsymbol{n})\boldsymbol{w}^*$, where $\boldsymbol{w}^*$ is a given boundary state at the inlet. We calculate the eigenvectors $\boldsymbol{r}_s$ corresponding to the eigenvalues λ_s, $s = 1, \ldots, 4$, of the matrix $\mathbb{A}_1(\boldsymbol{q}^{(L)})$, arrange them as columns in the matrix $\mathbb{T}$ and calculate $\mathbb{T}^{-1}$. Now we set

$$\boldsymbol{\alpha} = \mathbb{T}^{-1}\boldsymbol{q}^{(L)}, \quad \boldsymbol{\beta} = \mathbb{T}^{-1}\boldsymbol{q}^* \tag{35}$$

and define the state $\boldsymbol{q}^{(R)}$ by the relations

$$\boldsymbol{q}^{(R)} := \sum_{s=1}^{4} \gamma_s \boldsymbol{r}_s, \quad \gamma_s = \begin{cases} \alpha_s, & \lambda_s \geq 0, \\ \beta_s, & \lambda_s < 0. \end{cases} \tag{36}$$

Finally, the sought boundary state $\boldsymbol{w}^{(R)}$ is defined as

$$\boldsymbol{w}^{(R)} = \boldsymbol{Q}^{-1}(\boldsymbol{n})\boldsymbol{q}^{(R)}. \tag{37}$$

On the impermeable moving wall we prescribe the normal component of the velocity

$$\boldsymbol{v} \cdot \boldsymbol{n} = \boldsymbol{z}_D \cdot \boldsymbol{n}, \tag{38}$$

where $\boldsymbol{n}$ is the unit outer normal to Γ_{W_t} and $\boldsymbol{z}_D$ is the wall velocity. This implies that two eigenvalues of $\mathbb{P}_g(\boldsymbol{w}, \boldsymbol{n})$ are equal to zero, one eigenvalue is positive and one eigenvalue is negative. Then, in analogy to [10], Section 3.3.6, we should prescribe one quantity, namely $\boldsymbol{v} \cdot \boldsymbol{n}$, and extrapolate three quantities - tangential velocity, density and pressure.

However, here we define the numerical flux on Γ_{W_t} as the physical flux through the boundary with the assumption (38) taken into account. Thus, on Γ_{W_t} we write

$$\begin{aligned} \sum_{s=1}^{2} \boldsymbol{g}_s(\boldsymbol{w}) n_s &= (\boldsymbol{v} \cdot \boldsymbol{n} - \boldsymbol{z}_D \cdot \boldsymbol{n})\boldsymbol{w} + p\,(0, n_1, n_2, \boldsymbol{v} \cdot \boldsymbol{n})^T \\ &= p\,(0, n_1, n_2, \boldsymbol{z}_D \cdot \boldsymbol{n})^T =: \boldsymbol{H}_g. \end{aligned} \tag{39}$$

On the outlet (which does not depend on time) the pressure is prescribed and other variables are extrapolated. However, numerical experiments show that this treatment of the outlet boundary conditions can lead to the reflection of a strong intensity vortex on an artificial outlet in the numerical simulation. This problem, which does not appear in the examples presented in Section 4, will require a special analysis in the future.

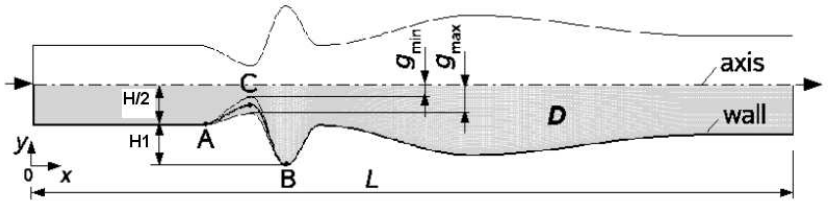


Figure 1: Computational domain for flow in human vocal folds.

Remark 1. *In practical computations, integrals appearing in the definitions of the forms* $\hat{a}_h$, $\hat{b}_h$, d_h, J_h, $\hat{J}_h$ *and* $\hat{\beta}_h$ *are evaluated with the aid of quadrature formulas.*

The developed numerical scheme can also be used for the numerical solution of inviscid flow, if we set $\mu = \lambda = k = 0$. *See [11].*

The linear algebraic system equivalent to (31) is solved either by a direct solver UMFPACK ([4]) or by the GMRES method with block diagonal preconditioning.

If we set $r = 0$, *we get a finite volume scheme.*

4 Numerical experiments

In order to demonstrate the applicability of the developed method, we present here results of two numerical experiments. In both cases piecewise quadratic finite elements ($r = 2$) in the space discretization are used. For the time discretization the second-order BDF formula from Section 3.2 is used.

4.1 Flow through a channel with moving walls

In the first example we present results of numerical experiments carried out for viscous compressible flow in a channel with geometry from [13] inspired by the shape of the human glottis and a part of supraglottal spaces as shown in Figure 1. The walls are moving in order to mimic the vibrations of vocal folds during voice production. The lower channel wall between the points A and B and the upper wall symmetric with respect to the axis of the channel are vibrating up and down periodically with frequency 100 Hz. This movement is interpolated into the domain resulting in the ALE mapping $\mathcal{A}_t$.

The width of the channel at the inlet (left part of the boundary) is $H = 0.016$ m and its length is $L = 0.16$ m. The width of the narrowest part of the channel (at the point C) oscillates between 0.0004 m and 0.0028 m with frequency 100 Hz. We consider the following input parameters and boundary conditions: magnitude of the inlet velocity $v_{in} = 4$ m/s, the viscosity $\mu = 15 \cdot 10^{-6}\ \mathrm{kg\,m^{-1}\,s^{-1}}$, the inlet density $\rho_{in} = 1.225\ \mathrm{kg\,m^{-3}}$, the outlet pressure $p_{out} = 97611$ Pa, the Reynolds number $Re = \rho_{in} v_{in} H/\mu = 5227$, heat conduction coefficient $k = 2.428 \cdot 10^{-2}\ \mathrm{kg\,m\,s^{-2}\,K^{-1}}$, the specific heat $c_v = 721.428\ \mathrm{m^2\,s^{-2}\,K^{-1}}$, the Poisson adiabatic constant $\gamma = 1.4$. The inlet Mach number is $M_{in} = 0.012$.

In [13], the described channel flow was solved by the first-order finite volume method under the assumption that the flow is symmetric with respect to the axis of the channel. This means that the results presented in [13] do not reflect the behaviour of real flow.

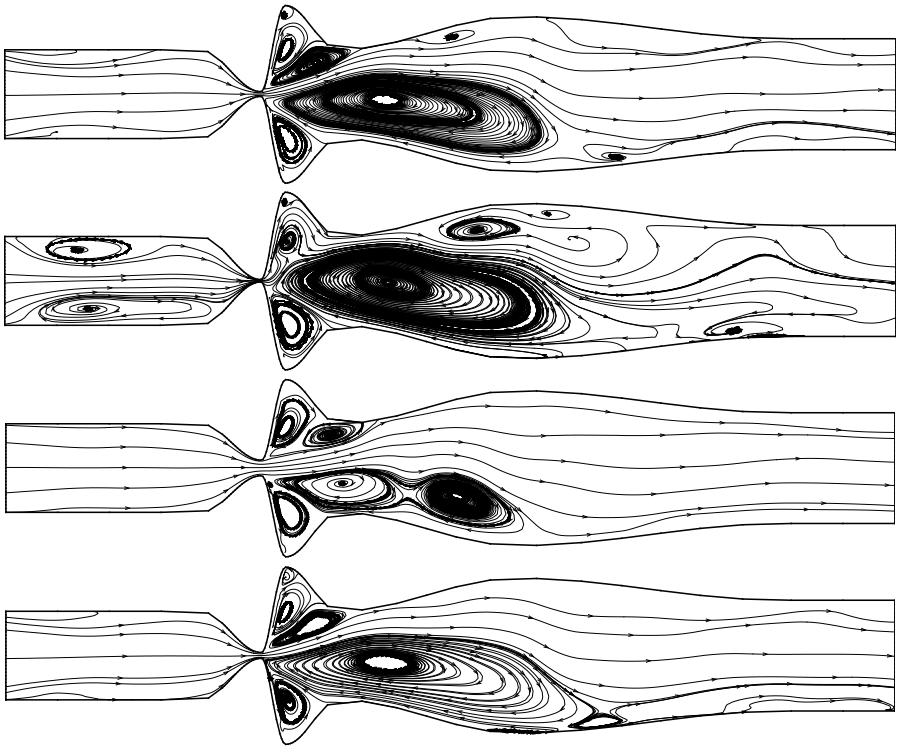


Figure 2: Streamlines at time instants $t = 2.016$, 2.124, 2.448, 2.664 s.

Here, we use piecewise quadratic finite elements and we do not require the symmetry of the flow field.

Figure 2 shows computed streamlines of the solution at different time instants 2.016, 2.124, 2.448, 2.664 s during the fourth period of the motion. In the solution we can observe large vortex formation convected through the domain. The flow field is neither periodic, nor axisymmetric, in spite of the fact that the computational domain is axisymmetric and the motion of the channel walls is periodic and symmetric as well. We can observe the so-called Coandă effect, when the main flow is attached to one of the walls. This effect is not present in results of the paper [13]

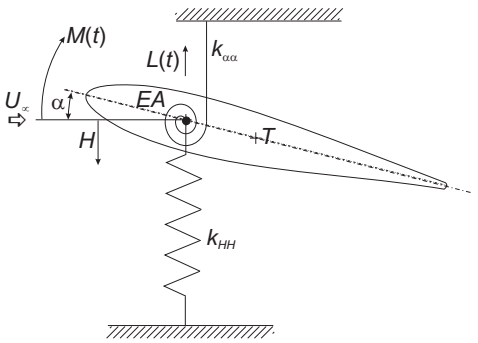


Figure 3: Computational domain for flow around a vibrating airfoil.

4.2 Flow induced airfoil vibrations

The second example is concerned with the simulation of vibrations of elastically supported airfoil NACA 0012 induced by compressible viscous flow. The airfoil has two degrees of freedom: the vertical displacement H (positively oriented downwards) and the angle of rotation around an elastic axis α (positively oriented clockwise), cf. Figure 3. The motion of the airfoil is described by the system of nonlinear ordinary differential equations for unknowns H, α:

$$\begin{aligned} m\ddot{H} + k_{HH}H + S_\alpha\,\ddot{\alpha}\,\cos\alpha - S_\alpha\dot{\alpha}^2\sin\alpha + d_{HH}\dot{H} &= -\mathcal{L}(t), \\ S_\alpha\ddot{H}\cos\alpha + I_\alpha\ddot{\alpha} + k_{\alpha\alpha}\alpha + d_{\alpha\alpha}\dot{\alpha} &= \mathcal{M}(t). \end{aligned} \tag{40}$$

The dot and two dots denote the first-order and second-order time derivative, respectively. We use the following notation: $\mathcal{L}(t)$ – aerodynamic lift force (upwards positive), $\mathcal{M}(t)$ – aerodynamic torsional moment (clockwise positive), m - mass of the airfoil, S_α – static moment around the elastic axis EO, I_α – inertia moment around the elastic axis EO, k_{HH} – bending stiffness, $k_{\alpha\alpha}$ – torsional stiffness, d_{HH} – structural damping in bending, $d_{\alpha\alpha}$ – structural damping in torsion, c - length of the chord of the airfoil, l – airfoil depth.

System (40) is equipped with the initial conditions prescribing the values $H(0), \alpha(0)$, $\dot{H}(0), \dot{\alpha}(0)$. It is transformed to a first-order ODE system and solved numerically by the fourth-order Runge-Kutta method. For the derivation of equations (40), see [14]. The aerodynamic lift force $\mathcal{L}$ acting in the vertical direction and the torsional moment $\mathcal{M}$ are defined by

$$\mathcal{L} = -\,l\int_{\Gamma_{Wt}}\sum_{j=1}^{2}\tau_{2j}n_j\,dS, \quad \mathcal{M} = l\int_{\Gamma_{Wt}}\sum_{i,j=1}^{2}\tau_{ij}n_j r_i^{\mathrm{ort}}\,dS, \tag{41}$$

where

$$\begin{aligned} \tau_{ij} &= (-p + \lambda\,\mathrm{div}\boldsymbol{v})\delta_{ij} + \mu\Big(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\Big), \\ r_1^{\mathrm{ort}} &= -(x_2 - x_{EO2}),\ r_2^{\mathrm{ort}} = x_1 - x_{EO1}. \end{aligned} \tag{42}$$

By τ_{ij} we denote the components of the stress tensor, δ_{ij} denotes the Kronecker symbol, $\boldsymbol{n} = (n_1, n_2)$ is the unit outer normal to $\partial\Omega_t$ on Γ_{Wt} (pointing into the airfoil) and $x_{EO} = (x_{EO1}, x_{EO2})$ is the position of the elastic axis (lying in the interior of the airfoil). Relations (41) and (42) define the coupling of the fluid dynamical model with the structural model.

4.2.1 Algorithm of the flow induced airfoil vibrations simulation

In the solution of the complete coupled fluid-structure interaction problem we apply the following algorithm:

1) Assume that the approximate solution of the discrete flow problem (31) at time levels t_{k-1} and t_k is known and the force $\mathcal{L}$ and torsional moment $\mathcal{M}$ are computed from (41).

2) Extrapolate $\mathcal{L}$ and $\mathcal{M}$ on the time interval $[t_k, t_{k+1}]$.

3) Compute the displacement H and angle α at time t_{k+1} as the solution of system (40).

4) Determine the position of the airfoil at time t_{k+1}, the domain $\Omega_{t_{k+1}}$, the ALE mapping and the domain velocity at time t_{k+1}.

5) Solve the discrete system at time t_{k+1}.

6) Compute $\mathcal{L}$ and $\mathcal{M}$ at time t_{k+1} and interpolate $\mathcal{L}$ and $\mathcal{M}$ on $[t_k, t_{k+1}]$.

7) Is higher accuracy needed? YES: go to 3); NO: $k := k + 1$, go to 2).

If in step 7) one goes to 2), the so-called loose (weak) coupling is applied. In our numerical experiments the stronger coupling was applied with 4 – 5 loops for obtaining the difference between two approximations of H and α less than 10^{-5}. The ALE mapping and the domain velocity are computed in the same way as in [8].

4.2.2 Results of numerical experiments

I) The simulation of flow induced airfoil vibrations was carried out for the following data: $m = 0.086622$ kg, $S_\alpha = -0.000779673$ kg m, $I_\alpha = 0.000487291$ kg m^2, $k_{HH} = 105.109$ N/m, $k_{\alpha\alpha} = 3.696682$ Nm/rad, $l = 0.05$ m, $c = 0.3$ m, $\mu = 1.8375 \cdot 10^{-5}$ kg m^{-1} s^{-1}, far-field density $\rho = 1.225$ kg m^{-3}, $H(0) = 0.02$ m, $\alpha(0) = 6$ degrees, $\dot{H}(0) = 0, \dot{\alpha} = 0$. We neglect the structural damping. The elastic axis is placed on the airfoil chord at the 40% distance from the leading edge.

The computational process starts at time $t = -\delta < 0$ by the solution of the flow, keeping the airfoil in a fixed position given by the prescribed initial translation H and the angle of attack α. Then, at time $t = 0$ the airfoil is released and we continue by the solution of a complete fluid-structure interaction problem.

Figure 4 shows the displacement H and the rotation angle α in dependence on time for the far-field velocity 10, 20, 30 and 40 m/s. The corresponding Reynolds number was in the range $2 \cdot 10^5 - 8 \cdot 10^5$. We see that for the velocities 10, 20 and 30 m/s the vibrations are damped, but for the velocity 40 m/s we get the flutter instability when the vibration amplitudes are increasing in time. The monotonous increase and decrease of the average values of H and α, respectively, shows that the flutter is combined with a divergence instability in the presented example.

II) In the above examples the flow was subsonic. The described method was also applied to transonic flow with far-field velocity 290 m/s, far-field Mach number 0.85, Reynolds number 5000 and initial data $H(0) = 0, \alpha(0) = 4$ degrees, $\dot{H}(0) = \dot{\alpha}(0) = 0$. In this case it was necessary to consider harder bending and torsional stiffnesses. We set $k_{HH} = 105109$ N/m and $k_{\alpha\alpha} = 36.956$ N m/rad. Figure 5 shows the time dependence of H and α. In Figure 6, Mach number isolines at time instants $t = 0.00261, 0.00661, 0.00831, 0.00961$ s are shown. We see an interesting system of shock waves, separated boundary layer, wake moving in time and vortices leaving the airfoil.

5 Conclusion

We have presented an efficient numerical scheme for the solution of the compressible Navier-Stokes equations in time dependent domains and the simulation of flow induced airfoil vibrations. It is based on several important ingredients:

- the ALE method applied to the compressible Navier-Stokes equations,
- the application of the discontinuous Galerkin method for the space discretization,
- semi-implicit time discretization,
- suitable treatment of boundary conditions,

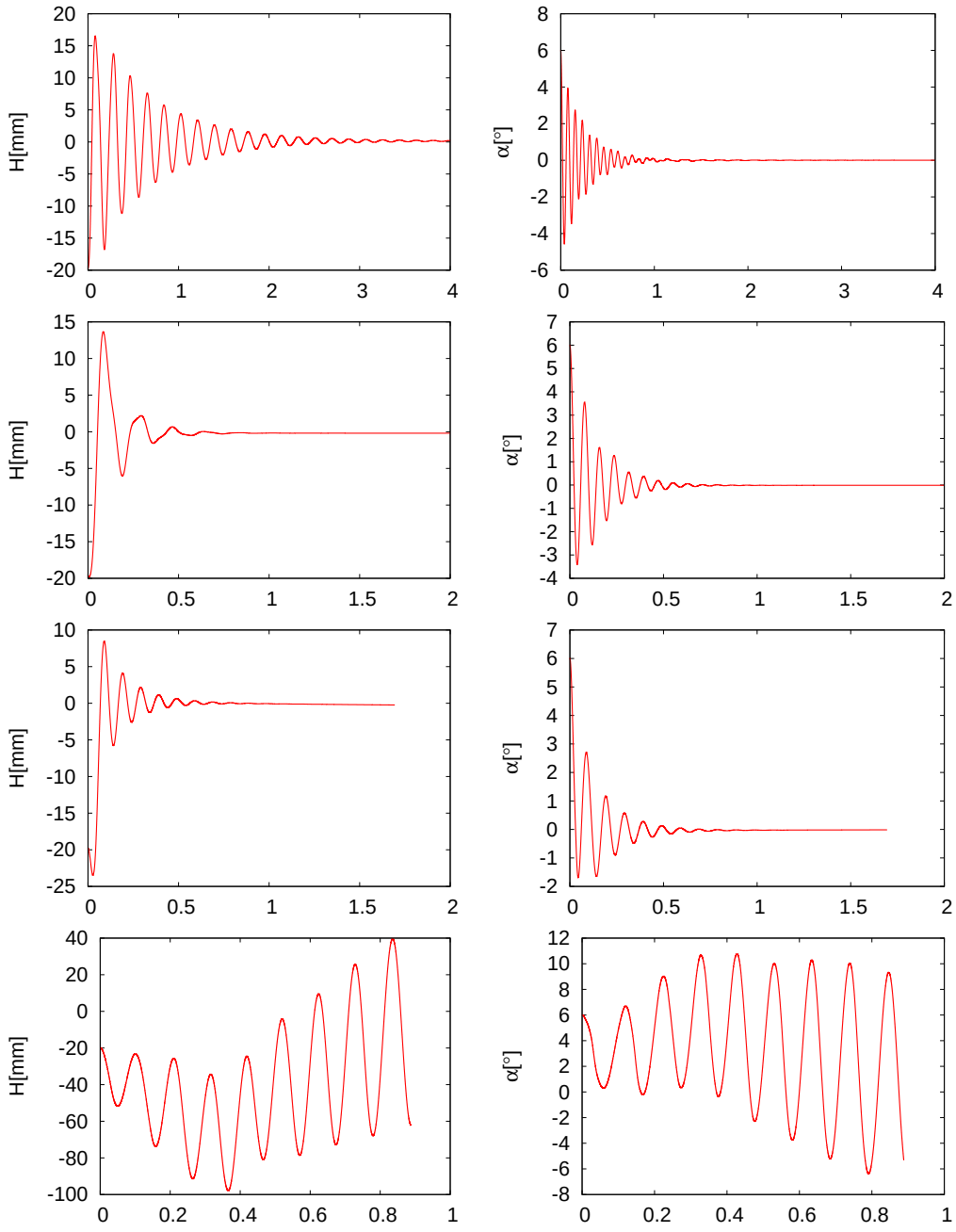


Figure 4: Displacement H (left) and rotation angle α (right) of the airfoil in dependence on time for far-field velocity 10, 20, 30 and 40 m/s.

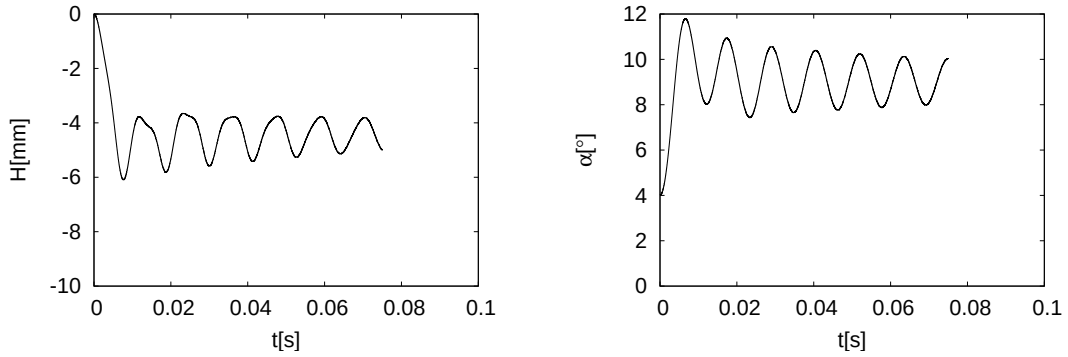


Figure 5: Displacement H (left) and rotation angle α (right) of the airfoil in dependence on time for far-field velocity 290 m/s and far-field Mach number 0.85.

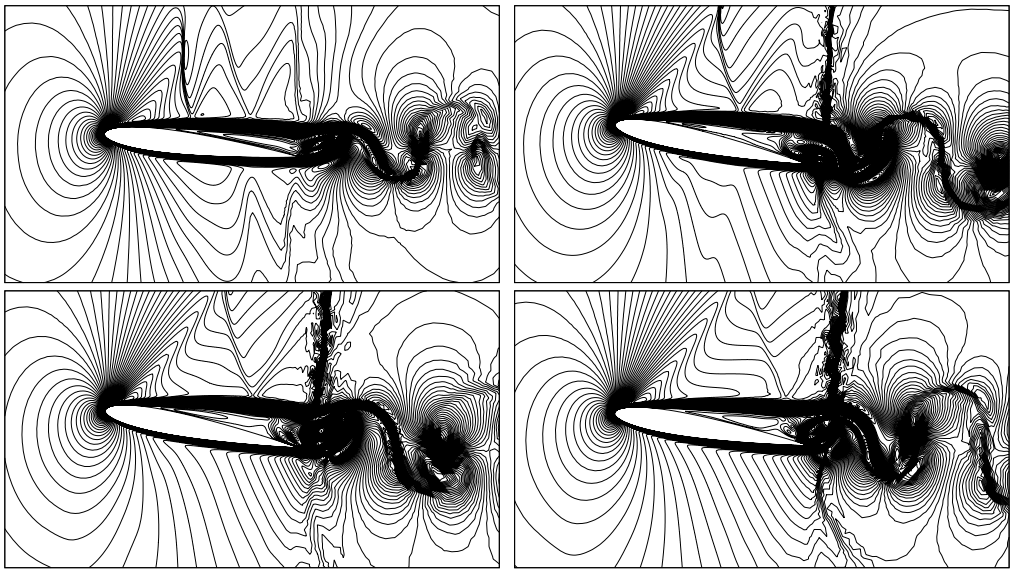


Figure 6: Flow past an airfoil: Mach number isolines for far-field velocity 290 m/s and far-field Mach number 0.85 at time instants $t = 0.00261, 0.00661, 0.00831, 0.00961$ s, ordered from left to right in rows.

- artificial viscosity applied in the vicinity of discontinuities.

The developed method behaves as unconditionally stable and appears to be robust with respect to the magnitude of the Mach number. The presented examples demonstrate that the method can be applied to the numerical solution of compressible flow with very low Mach numbers as well as high-speed flow with shock waves and contact discontinuities.

Future work will be concentrated on the following topics:

- further analysis of the robustness and accuracy of the method with respect to the Mach number and Reynolds number,
- investigation of various types of boundary conditions,
- the realization of a remeshing in case of closing the channel during the oscillation period of the channel walls,
- the coupling of the developed method with the solution of elasticity equations describing the deformation of vocal folds,
- the use of a suitable turbulence model.

Acknowledgements This work was supported by the research project MSM 0021620839 (M. Feistauer, V. Kučera) and by the Nečas Center for Mathematical Modelling, project LC06052 (J. Česenek), both financed by the Ministry of Education of the Czech Republic. It was also partly supported by the grants No. 201/08/0012 (M. Feistauer, V. Kučera) and P101/11/0207 (J. Horáček in the year 2011) of the Czech Science Foundation, and by the grant SVV-2010-261316 financed by the Charles University in Prague (J. Prokopová).

References

[1] S. Badia, R. Codina: On some fluid-structure iterative algorithms using pressure segregation methods. Application to aeroelasticity. *Int. J. Numer. Meth. Engng*, **72**, 46–71 (2007).

[2] F. Bassi, S. Rebay: High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, **138**, 251–285 (1997).

[3] C. E. Baumann, J. T. Oden: A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations. *Int. J. Numer. Methods Fluids*, **31**, 79–95 (1999).

[4] T. A. Davis, I. S. Duff: A combined unifrontal/multifrontal method for unsymmetric sparse matrices. *ACM Transactions on Mathematical Software*, **25**, 1–19 (1999).

[5] V. Dolejší: Semi-implicit interior penalty discontinuous Galerkin methods for viscous compressible flows. *Commun. Comput. Phys.*, **4**, 231–274 (2008).

[6] V. Dolejší, M. Feistauer: A semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow. *J. Comput. Phys.*, **198**, 727–746 (2004).

[7] V. Dolejší, M. Feistauer, C. Schwab: On some aspects of the discontinuous Galerkin finite element method for conservation laws. *Math. Comput. Simul.*, **61**, 333–346 (2003).

[8] L. Dubcová, M. Feistauer, J. Horáček, P. Sváček: Numerical simulation of interaction between turbulent flow and a vibrating airfoil. *Computing and Visualization in Science*, **12**, 207–225 (2009).

[9] M. Feistauer, V. Dolejší, V. Kučera: On the discontinuous Galerkin method for the simulation of compressible flow with wide range of Mach numbers. *Computing and Visualization in Science*, **10**, 17–27 (2007).

[10] M. Feistauer, J. Felcman, I. Straškraba: Mathematical and Computational Methods for Compressible Flow. *Clarendon Press*, Oxford (2003).

[11] M. Feistauer, V. Kučera: On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.*, **224**, 208–221 (2007).

[12] T. Nomura, T.J.R. Hughes: An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. *Comput. Methods Appl. Mech. Engrg.*, **95**, 115-138 (1992).

[13] P. Punčochářová, J. Fürst, K. Kozel, J. Horáček: Numerical solution of compressible flow with low Mach number through oscillating glottis. *Proceedings of the 9th International Conference On Flow-Induced Vibration (FIV 2008)*, Institute of Thermomechanics AS CR, Prague, 135-140 (2008).

[14] P. Sváček, M. Feistauer, J. Horáček: Numerical simulation of flow induced airfoil vibrations with large amplitudes. *J. of Fluids and Structures*, **23**, 391-411 (2007).

[15] J. J. W. van der Vegt, H. van der Ven: Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flow. *J. Comput. Phys.*, **182**, 546–585 (2002).

[16] G. Vijayasundaram: Transonic flow simulation using upstream centered scheme of Godunov type in finite elements. *J. Comput. Phys.*, **63**, 416–433 (1986).

DGFEM for dynamical systems describing interaction of compressible fluid and structures

Miloslav Feistauer[1,2], Jaroslava Hasnedlová-Prokopová[1], Jaromír Horáček[2], Adam Kosík[1,2], and Václav Kučera[1]

[1] *Charles University Prague, Faculty of Mathematics and Physics,, Sokolovská 83, 186 75 Praha 8, Czech Republic*
[2] *Institute of Thermomechanics, Academy of Sciences of the Czech Republic,, Dolejškova 5, 182 00 Praha 8, Czech Republic*

Published in December 2013,
Journal of Computational and Applied Mathematis

Abstract

The paper is concerned with the numerical solution of flow-induced vibrations of elastic structures. The dependence on time of the domain occupied by the fluid is taken into account with the aid of the ALE (Arbitrary Lagrangian-Eulerian) formulation of the compressible Navier-Stokes equations. The deformation of the elastic body, caused by aeroelastic forces, is described by the linear dynamical elasticity equations. These two systems are coupled by transmission conditions. The flow problem is discretized by the discontinuous Galerkin finite element method (DGFEM) in space and by the backward difference formula (BDF) in time. The structural problem is discretized by conforming finite elements and the Newmark method. The fluid-structure interaction is realized via weak or strong coupling algorithms. The developed technique is tested by numerical experiments and applied to the simulation of vibrations of vocal folds during phonation onset.

Keywords: compressible Navier-Stokes equations; time dependent domain; ALE method, discontinuous Galerkin method; semi-implicit time discretization; dynamic elasticity equations; conforming finite elements, Newmark method; weak and strong coupling; flow in glottis; flow-induced vibrations of vocal folds.

1 Introduction

The studies on flow-induced vibrations play an important role in a number of fields in science and technology (e.g., vibrations of airplane wings or turbine blades, interaction of wind with bridges, TV towers or cooling towers of power stations) but also in biomechanics, e.g., simulation of the vocal folds vibrations and voice production. In all of these examples the moving medium is gas, i.e. compressible fluid. For low Mach number flows incompressible models are used (as e.g. in [3], [15]), but in some cases compressibility plays an important role.

The goal of our research is the numerical finite element (FE) simulation of interaction of compressible 2D viscous flow in the glottal region with a compliant tissue of the human vocal folds modeled by a 2D elastic layered structure. A current challenging question is

a mathematical and physical description of the mechanism for transforming the airflow energy in the glottis into the acoustic energy representing the voice source in humans. The primary voice source is given by the airflow coming from the lungs that causes self-oscillations of the vocal folds. The voice source signal travels from the glottis to the mouth, exciting the acoustic supraglottal spaces, and becomes modified by acoustic resonance properties of the vocal tract ([16]).

An overview [2] presents the current state of mathematical models for the human phonation process. In current publications various simplified glottal flow models are used. They are based on the Bernoulli equation ([16]), 1D models for an incompressible inviscid fluid ([9]), 2D incompressible Navier-Stokes equations solved by the finite volume method ([1]) or finite element method ([18]). Acoustic wave propagation in the vocal tract is usually modelled separately using linear acoustic perturbation theory ([17]). Work [14] is concerned with the finite volume solution of the Navier-Stokes equations for a compressible fluid with prescribed periodic changes of the channel cross-section of the glottal channel. The phonation onset was studied by using the interaction of incompressible potential flow model with three-mass lumped model for the vibrating vocal folds in [8] and for a 2D elastic model of the vocal folds in [19].

Only in the paper [14], the model of compressible flow is used. It is solved by the finite volume method, but the vibrations of the moving walls are prescribed. Otherwise, all the above mentioned papers use the model of incompressible flow. In many cases the incompressible flow approximates the airflow well, but often the compressibility plays important role, even in the case of low Mach number flows. It is particularly the case, when acoustic effects as propagating pressure waves appear. In the domain representing the vocal tract, this can happen, when the vocal folds are very close to each other and when the voice source is generated in the glottis. Therefore, it is suitable to analyze the compressible flow in vocal flows as well. However, it is well-known that the numerical solution of low Mach number flow at incompressible limit is a very difficult task and standard finite volume and finite element schemes applied to this type of flow fail. Our goal is to develop a method, which overcomes this obstacle.

The present paper is devoted to the numerical simulation of vocal folds vibrations induced by compressible viscous flow. The air flow is described by the compressible Navier-Stokes equations written in the arbitrary Lagrangian-Eulerian (ALE) form in order to take into account the time dependence of the domain occupied by the air. The vocal folds are considered as isotropic elastic bodies. Their vibrations are described by the linear elasticity equations. The coupled fluid-structure interaction problem represents a strongly nonlinear dynamical system, which is analyzed numerically.

The flow problem is discretized in space by the discontinuous Galerkin finite element method (DGFEM), using piecewise polynomial approximations, in general discontinuous on interfaces between neighbouring elements. The time discretization is carried out by the backward difference formula (BDF) in time. The structural problem is approximated by conforming finite elements and the Newmark method. The fluid-structure interaction is realized via weak or strong coupling algorithms.

The main purpose of the paper is to present a numerical technique allowing the simulation of vocal fold vibrations induced by compressible flow. The developed method is tested on a model problem in order to show the applicability of the method to the compressible flow in time-dependent domains and to the interaction of gas flow with elastic bodies. The results of numerical experiments are qualitatively comparable with results of other works (using the model of incompressible flow) and with wind tunnel experiments.

The contents of the paper is the following. In Section 2, the continuous fluid-structure interaction (FSI) problem is formulated. Section 3 is concerned with the derivation of the discrete problem. Section 4 is devoted to the realization of the coupled FSI problem. It consists of the construction of the ALE mapping and the formulation of the coupling algorithms. In Section 5, we present results of numerical tests showing the applications to the simulation of flow-induced vibrations of vocal folds. In Conclusion, subjects for future work are formulated.

2 Continuous problem

In this section we shall formulate the problem of the interaction of a compressible flow with an elastic structure.

2.1 Formulation of the flow problem

We consider a compressible flow in a bounded domain $\Omega_t \subset I\!R^2$ depending on time $t \in [0,T]$. We assume that the boundary of Ω_t is formed by three disjoint parts: $\partial\Omega_t = \Gamma_I \cup \Gamma_O \cup \Gamma_{W_t}$, where Γ_I is the inlet, Γ_O is the outlet and Γ_{W_t} denotes impermeable walls that may move in dependence on time.

The dependence of the domain Ω_t on time is taken into account with the use of the *arbitrary Lagrangian-Eulerian* (ALE) method, see e.g. [13]. It is based on a regular one-to-one ALE mapping of the reference configuration Ω_0 onto the current configuration Ω_t:

$$\mathcal{A}_t : \overline{\Omega}_0 \longrightarrow \overline{\Omega}_t, \ i.e. \ \boldsymbol{X} \in \overline{\Omega}_0 \longmapsto \boldsymbol{x} = \boldsymbol{x}(\boldsymbol{X},t) = \mathcal{A}_t(\boldsymbol{X}) \in \overline{\Omega}_t.$$

We define the domain velocity:

$$\begin{aligned} \tilde{\boldsymbol{z}}(\boldsymbol{X},t) &= \frac{\partial}{\partial t}\mathcal{A}_t(\boldsymbol{X}), \quad t \in [0,T], \ \boldsymbol{X} \in \Omega_0, \\ \boldsymbol{z}(\boldsymbol{x},t) &= \tilde{\boldsymbol{z}}(\mathcal{A}^{-1}(\boldsymbol{x}),t), \quad t \in [0,T], \ \boldsymbol{x} \in \Omega_t \end{aligned} \tag{1}$$

and the ALE derivative of the vector function $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{x},t)$ defined for $\boldsymbol{x} \in \Omega_t$ and $t \in [0,T]$:

$$\frac{D^A}{Dt}\boldsymbol{w}(\boldsymbol{x},t) = \frac{\partial \tilde{\boldsymbol{w}}}{\partial t}(\boldsymbol{X},t), \tag{2}$$

where

$$\tilde{\boldsymbol{w}}(\boldsymbol{X},t) = \boldsymbol{w}(\mathcal{A}_t(\boldsymbol{X}),t), \ \boldsymbol{X} \in \Omega_0, \ \boldsymbol{x} = \mathcal{A}_t(\boldsymbol{X}).$$

Then, using the relations

$$\frac{D^A w_i}{Dt} = \frac{\partial w_i}{\partial t} + \operatorname{div}(\boldsymbol{z} w_i) - w_i \operatorname{div} \boldsymbol{z}, \quad i = 1,\dots,4,$$

we can write the governing system consisting of the continuity equation, the Navier-Stokes equations and the energy equation in the ALE form

$$\frac{D^A \boldsymbol{w}}{Dt} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{g}_s(\boldsymbol{w})}{\partial x_s} + \boldsymbol{w} \operatorname{div} \boldsymbol{z} = \sum_{s=1}^{2} \frac{\partial \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_s}. \tag{3}$$

See, for example [6]. Here

$$\begin{aligned}
\boldsymbol{w} &= (w_1,\dots,w_4)^T = (\rho,\rho v_1,\rho v_2,E)^T \in \mathbb{R}^4, \\
\boldsymbol{w} &= \boldsymbol{w}(x,t),\ x\in\Omega_t,\ t\in(0,T), \\
\boldsymbol{g}_s(\boldsymbol{w}) &= \boldsymbol{f}_s(\boldsymbol{w}) - z_s\boldsymbol{w}, \quad s=1,2, \\
\boldsymbol{f}_i(\boldsymbol{w}) &= (f_{i1},\cdots,f_{i4})^T = (\rho v_i,\rho v_1 v_i+\delta_{1i}\,p,\rho v_2 v_i+\delta_{2i}\,p,(E+p)v_i)^T, \\
\boldsymbol{R}_i(\boldsymbol{w},\nabla\boldsymbol{w}) &= (R_{i1},\dots,R_{i4})^T = \left(0,\tau_{i1}^V,\tau_{i2}^V,\tau_{i1}^V\,v_1+\tau_{i2}^V\,v_2+k\partial\theta/\partial x_i\right)^T, \\
\tau_{ij}^V &= \lambda\,\mathrm{div}\boldsymbol{v}\,\delta_{ij}+2\mu\,d_{ij}(\boldsymbol{v}),\ d_{ij}(\boldsymbol{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right).
\end{aligned} \tag{4}$$

We use the following notation: ρ – density, p – pressure, E – total energy, $\boldsymbol{v}=(v_1,v_2)$ – velocity, θ – absolute temperature, $\gamma>1$ – Poisson adiabatic constant, $c_v>0$ – specific heat at constant volume, $\mu>0, \lambda=-2\mu/3$ – viscosity coefficients, k – heat conduction, τ_{ij}^V – components of the viscous part of the stress tensor. The vector-valued function $\boldsymbol{w}$ is called state vector, the functions $\boldsymbol{f}_i$ are the so-called inviscid fluxes and $\boldsymbol{R}_i$ represent viscous terms. The above system is completed by the thermodynamical relations

$$p=(\gamma-1)(E-\frac{1}{2}\rho|\boldsymbol{v}|^2),\quad \theta=\frac{1}{c_v}\Big(\frac{E}{\rho}-\frac{1}{2}|\boldsymbol{v}|^2\Big). \tag{5}$$

The resulting system is equipped with the initial condition

$$\boldsymbol{w}(x,0)=\boldsymbol{w}^0(x),\quad x\in\Omega_0, \tag{6}$$

and the following boundary conditions:

$$\begin{aligned}
&\text{a)}\ \ \rho|_{\Gamma_I}=\rho_D,\quad \text{b)}\ \ \boldsymbol{v}|_{\Gamma_I}=\boldsymbol{v}_D=(v_{D1},v_{D2})^{\mathrm{T}}, \\
&\text{c)}\ \ \sum_{i,j=1}^{2}\tau_{ij}^V n_i v_j + k\frac{\partial\theta}{\partial n}=0 \quad \text{on } \Gamma_I, \\
&\text{d)}\ \ \boldsymbol{v}|_{\Gamma_{W_t}}=\boldsymbol{z}_D=\text{velocity of a moving wall},\ \text{e)}\ \ \frac{\partial\theta}{\partial n}|_{\Gamma_{W_t}}=0 \text{ on } \Gamma_{W_t}, \\
&\text{f)}\ \ \sum_{i=1}^{2}\tau_{ij}^V n_i=0,\quad j=1,\,2,\quad \text{g)}\ \ \frac{\partial\theta}{\partial n}=0 \text{ on } \Gamma_O,
\end{aligned} \tag{7}$$

with prescribed data $\rho_D, \boldsymbol{v}_D$ and $\boldsymbol{z}_D$.

2.2 Elasticity problem and fluid-structure interaction coupling

For the description of the deformation of an elastic structure we shall use the model of dynamical linear elasticity formulated in a bounded open set $\Omega^b\subset\mathbb{R}^2$ representing the elastic body, which has a common boundary with the reference domain Ω_0 occupied by the fluid at the initial time. We denote by $\boldsymbol{u}(\boldsymbol{X},t)=(u_1(\boldsymbol{X},t),u_2(\boldsymbol{X},t))$, $\boldsymbol{X}=(X_1,X_2)\in\Omega^b$, $t\in(0,T)$, the displacement of the body. The equations describing the deformation of the elastic body Ω^b have the form

$$\varrho^b\frac{\partial^2 u_i}{\partial t^2}+C\varrho^b\frac{\partial u_i}{\partial t}-\sum_{j=1}^{2}\frac{\partial\tau_{ij}^b}{\partial X_j}=0 \quad \text{in } \Omega^b\times(0,T),\quad i=1,2. \tag{8}$$

Here τ_{ij}^b are the components of the stress tensor defined by the generalized Hooke's law for isotropic bodies

$$\tau_{ij}^b = \lambda^b \mathrm{div}\, \boldsymbol{u}\, \delta_{ij} + 2\mu^b e_{ij}^b(\boldsymbol{u}), \quad i,j = 1,2. \tag{9}$$

By $\boldsymbol{e}^b = \{e_{ij}^b\}_{i,j=1}^2$ we denote the strain tensor defined by

$$e_{ij}^b(\boldsymbol{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right), \quad i,j = 1,2. \tag{10}$$

The Lamé coefficients λ^b and μ^b are related to the Young modulus E^b and the Poisson ratio σ^b as

$$\lambda^b = \frac{E^b\sigma^b}{(1+\sigma^b)(1-2\sigma^b)}, \quad \mu^b = \frac{E^b}{2(1+\sigma^b)}, \tag{11}$$

The expression $C\varrho^b \frac{\partial u_i}{\partial t}$, where $C \geq 0$, is the dissipative structural damping of the system and ϱ^b denotes the material density.

We complete the elasticity problem by initial and boundary conditions. The initial conditions read

$$\boldsymbol{u}(\cdot,0) = \boldsymbol{0}, \quad \frac{\partial \boldsymbol{u}}{\partial t}(\cdot,0) = \boldsymbol{0}, \quad \text{in } \Omega^b. \tag{12}$$

Further, we assume that $\partial\Omega^b = \Gamma_W^b \cup \Gamma_D^b$, where Γ_W^b and Γ_D^b are two disjoints parts of $\partial\Omega^b$. We assume that Γ_W^b is a common part between the fluid and structure at time $t = 0$. This means that $\Gamma_W^b \subset \Gamma_{W_0}$. On Γ_W^b we prescribe the normal component of the stress tensor and assume that the part Γ_D^b is fixed. This means that the following boundary conditions are used:

$$\sum_{j=1}^{2} \tau_{ij}^b n_j = T_i^{\boldsymbol{n}} \quad \text{on } \Gamma_W^b \times (0,T), \quad i = 1,2, \tag{13}$$

$$\boldsymbol{u} = 0 \quad \text{on } \Gamma_D^b \times (0,T). \tag{14}$$

By $\boldsymbol{T^n} = (T_1^n, T_2^n)$ we denote the prescribed normal component of the stress tensor.

The structural problem consists in finding the displacement $\boldsymbol{u}$ satisfying equations (8) and the initial and boundary conditions (12) – (14).

Now we shall deal with the formulation of the coupled FSI problem. We denote the common boundary between the fluid and the structure at time t by $\tilde{\Gamma}_{W_t}$. It is given by

$$\tilde{\Gamma}_{W_t} = \left\{\boldsymbol{x} \in \mathbb{R}^2;\ \boldsymbol{x} = \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{X},t),\ \boldsymbol{X} \in \Gamma_W^b\right\}. \tag{15}$$

Thus, the domain Ω_t is determined by the displacement $\boldsymbol{u}$ of the part Γ_W^b at time t. The ALE mapping $\mathcal{A}_t$ is constructed with the aid of a special stationary linear elasticity problem - see Section 4.1.

If the domain Ω_t occupied by the fluid at time t is known, we can solve the problem describing the flow and compute the surface force acting onto the body on the interface $\tilde{\Gamma}_{W_t}$, which can be transformed to the reference configuration, i.e. to the interface Γ_W^b. In case of the linear elasticity model, when only small deformations are considered, we get the transmission condition

$$\sum_{j=1}^{2} \tau_{ij}^b(\boldsymbol{X}) n_j(\boldsymbol{X}) = -\sum_{j=1}^{2} \tau_{ij}^f(\boldsymbol{x}) n_j(\boldsymbol{X}), \quad i = 1,2, \tag{16}$$

where τ_{ij}^f are the components of the stress tensor of the fluid:

$$\tau_{ij}^f = -p\delta_{ij} + \tau_{ij}^V, \quad i,j = 1,2, \tag{17}$$

the points $\boldsymbol{x}$ and $\boldsymbol{X}$ satisfy the relation

$$\boldsymbol{x} = \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{X}, t). \tag{18}$$

and $\boldsymbol{n}(\boldsymbol{X}) = (n_1(\boldsymbol{X}), n_2(\boldsymbol{X}))$ denotes the unit outer normal to the body Ω^b on Γ_W^b at the point $\boldsymbol{X}$. Further, the fluid velocity is defined on the moving part of the boundary $\tilde{\Gamma}_{W_t}$ by the second transmission condition

$$\boldsymbol{v}(\boldsymbol{x}, t) = \boldsymbol{z}_D(\boldsymbol{x}, t) = \frac{\partial \boldsymbol{u}(\boldsymbol{X}, t)}{\partial t}. \tag{19}$$

Now we formulate the *continuous FSI problem*: We want to determine the domain $\Omega_t,\ t \in (0,T]$ and functions $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{x}, t),\ \boldsymbol{x} \in \overline{\Omega}_t,\ t \in [0,T]$ and $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{X}, t),\ \boldsymbol{X} \in \overline{\Omega}^b,\ t \in [0,T]$ satisfying equations (3), (8), the initial conditions (6), (12), the boundary conditions (7), (13), (14) and the transmission conditions (16), (19).

This FSI problem represents a strongly nonlinear dynamical system. Theoretical analysis of qualitative properties of this problem, as the existence, uniqueness and regularity of its solution, is open. Therefore, in the sequel we shall be concerned with its numerical solution.

3 Discrete problem

First we describe numerical methods for the solution of separately considered flow and structural problems.

3.1 Discretization of the flow problem

3.1.1 Space discretization

For the space semidiscretization we use the discontinuous Galerkin finite element method (DGFEM). We construct a polygonal approximation Ω_{ht} of the domain Ω_t. By $\mathcal{T}_{ht}$ we denote a partition of the closure $\overline{\Omega}_{ht}$ of the domain Ω_{ht} into a finite number of closed triangles K with mutually disjoint interiors such that $\overline{\Omega}_{ht} = \bigcup_{K \in \mathcal{T}_{ht}} K$.

By $\mathcal{F}_{ht}$ we denote the system of all faces of all elements $K \in \mathcal{T}_{ht}$. Further, we introduce the set of all interior faces $\mathcal{F}_{ht}^I = \{\Gamma \in \mathcal{F}_{ht};\ \Gamma \subset \Omega_t\}$, the set of all boundary faces $\mathcal{F}_{ht}^B = \{\Gamma \in \mathcal{F}_{ht};\ \Gamma \subset \partial\Omega_{ht}\}$ and the set of all "Dirichlet" boundary faces $\mathcal{F}_{ht}^D = \left\{\Gamma \in \mathcal{F}_{ht}^B;\ \text{a Dirichlet condition is prescribed on } \Gamma\right\}$. Each $\Gamma \in \mathcal{F}_{ht}$ is associated with a unit normal vector $\boldsymbol{n}_\Gamma$ to Γ. For $\Gamma \in \mathcal{F}_{ht}^B$ the normal $\boldsymbol{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega_{ht}$. We set $d(\Gamma) =$ length of $\Gamma \in \mathcal{F}_{ht}$.

For each $\Gamma \in \mathcal{F}_{ht}^I$ there exist two neighbouring elements $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_{ht}$ such that $\Gamma \subset \partial K_\Gamma^{(R)} \cap \partial K_\Gamma^{(L)}$. We use the convention that $K_\Gamma^{(R)}$ lies in the direction of $\boldsymbol{n}_\Gamma$ and $K_\Gamma^{(L)}$ lies in the opposite direction to $\boldsymbol{n}_\Gamma$. If $\Gamma \in \mathcal{F}_{ht}^B$, then the element adjacent to Γ will be denoted by $K_\Gamma^{(L)}$.

The approximate solution will be sought in the space of piecewise polynomial functions

$$\boldsymbol{S}_{ht} = [S_{ht}]^4, \quad \text{with} \quad S_{ht} = \{v; v|_K \in P_r(K)\ \forall\, K \in \mathcal{T}_{ht}\}, \tag{20}$$

where $r \geq 1$ is an integer and $P_r(K)$ denotes the space of all polynomials on K of degree $\leq r$. A function $\boldsymbol{\varphi} \in \boldsymbol{S}_{ht}$ is, in general, discontinuous on interfaces $\Gamma \in \mathcal{F}_{ht}^I$. By $\boldsymbol{\varphi}_\Gamma^{(L)}$ and $\boldsymbol{\varphi}_\Gamma^{(R)}$ we denote the values of $\boldsymbol{\varphi}$ on Γ considered from the interior and the exterior of $K_\Gamma^{(L)}$, respectively, and set

$$\langle \boldsymbol{\varphi} \rangle_\Gamma = (\boldsymbol{\varphi}_\Gamma^{(L)} + \boldsymbol{\varphi}_\Gamma^{(R)})/2, \quad [\boldsymbol{\varphi}]_\Gamma = \boldsymbol{\varphi}_\Gamma^{(L)} - \boldsymbol{\varphi}_\Gamma^{(R)}. \tag{21}$$

The discrete problem is derived in the following way: We multiply system (3) by a test function $\boldsymbol{\varphi}_h \in \boldsymbol{S}_{ht}$, integrate over $K \in \mathcal{T}_{ht}$, apply Green's theorem, sum over all elements $K \in \mathcal{T}_{ht}$, use the concept of the numerical flux and introduce suitable terms mutually vanishing for a regular exact solution. Moreover, we carry out a linearization of nonlinear terms. In a similar way as in [6] we define the following forms.

Convection form: We set $\mathbb{A}_s(\boldsymbol{w}) = D\boldsymbol{f}_s(\boldsymbol{w})/D\boldsymbol{w}$, which is the Jacobi matrix of the mapping $\boldsymbol{f}_s$. Then $\frac{D\boldsymbol{g}_s(\boldsymbol{w})}{D\boldsymbol{w}} = \mathbb{A}_s(\boldsymbol{w}) - z_s\mathbb{I}$, and we write $\mathbb{P}_g(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^2 \frac{D\boldsymbol{g}_s(\boldsymbol{w})}{D\boldsymbol{w}} n_s = \sum_{s=1}^2 (\mathbb{A}_s(\boldsymbol{w})n_s - z_s n_s \mathbb{I})$. By [5], this matrix is diagonalizable. It means that there exists a nonsingular matrix $\mathbb{T} = \mathbb{T}(\boldsymbol{w}, \boldsymbol{n})$ such that $\mathbb{P}_g = \mathbb{T}\mathbb{\Lambda}\mathbb{T}^{-1}, \quad \mathbb{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_4)$ where $\lambda_i = \lambda_i(\boldsymbol{w}, \boldsymbol{n})$ are eigenvalues of the matrix $\mathbb{P}_g$. Further, we define the "positive" and "negative" parts of the matrix $\mathbb{P}_g$ by $\mathbb{P}_g^\pm = \mathbb{T}\mathbb{\Lambda}^\pm\mathbb{T}^{-1}, \quad \mathbb{\Lambda}^\pm = \mathrm{diag}(\lambda_1^\pm, \ldots, \lambda_4^\pm)$, where $\lambda^+ = \max(\lambda, 0)$, $\lambda^- = \min(\lambda, 0)$. Now, in the same way as in [6], for $\overline{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{ht}$ we define the linearized convection form

$$\begin{aligned}
&\hat{b}_h(\overline{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) \\
&= -\sum_{K \in \mathcal{T}_{ht_{k+1}}} \int_K \sum_{s=1}^2 (\mathbb{A}_s(\overline{\boldsymbol{w}}_h) - z_s(x))\mathbb{I})\boldsymbol{w}_h) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \,\mathrm{d}\boldsymbol{x} \\
&+ \sum_{\Gamma \in \mathcal{F}_{ht}^I} \int_\Gamma \Big(\mathbb{P}_g^+\big(\langle \overline{\boldsymbol{w}}_h \rangle, \boldsymbol{n}_\Gamma\big)\boldsymbol{w}_h^{(L)} + \mathbb{P}_g^-\big(\langle \overline{\boldsymbol{w}}_h \rangle, \boldsymbol{n}_\Gamma\big)\boldsymbol{w}_h^{(R)} \Big) \cdot [\boldsymbol{\varphi}_h] \,\mathrm{d}S \\
&+ \sum_{\Gamma \in \mathcal{F}_{ht}^B} \int_\Gamma \Big(\mathbb{P}_g^+\big(\langle \overline{\boldsymbol{w}}_h \rangle, \boldsymbol{n}_\Gamma\big)\boldsymbol{w}_h^{(L)} + \mathbb{P}_g^-\big(\langle \overline{\boldsymbol{w}}_h \rangle, \boldsymbol{n}_\Gamma\big)\overline{\boldsymbol{w}}_h^{(R)} \Big) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S.
\end{aligned} \tag{22}$$

If $\Gamma \in \mathcal{F}_{ht}^B$, it is necessary to specify the boundary state $\overline{\boldsymbol{w}}_{h\Gamma}^{(R)}$ appearing in the numerical flux $\boldsymbol{H}_g$ in the definition of the inviscid form $\hat{b}_h$. Here we use the approach applied in the case of inviscid flow simulation, treated in [5], using a linearized initial-boundary value 1D Riemann problem.

Viscous form: The linearization of the viscous terms is based on the fact that $\boldsymbol{R}_s(\boldsymbol{w}_h, \nabla \boldsymbol{w}_h)$ is linear in $\nabla \boldsymbol{w}$ and nonlinear in $\boldsymbol{w}$. We get the linearized viscous form

$$\begin{aligned}
\hat{a}_h(\overline{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) &= \sum_{K \in \mathcal{T}_{ht}} \int_K \sum_{s=1}^2 \boldsymbol{R}_s(\overline{\boldsymbol{w}}_h, \nabla \boldsymbol{w}_h) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \,\mathrm{d}\boldsymbol{x} \\
&- \sum_{\Gamma \in \mathcal{F}_{ht}^I} \int_\Gamma \sum_{s=1}^2 \big\langle \boldsymbol{R}_s(\overline{\boldsymbol{w}}_h, \nabla \boldsymbol{w}_h) \big\rangle (\boldsymbol{n}_\Gamma)_s \cdot [\boldsymbol{\varphi}_h] \,\mathrm{d}S \\
&- \sum_{\Gamma \in \mathcal{F}_{ht}^D} \int_\Gamma \sum_{s=1}^2 \boldsymbol{R}_s(\overline{\boldsymbol{w}}_h, \nabla \boldsymbol{w}_h)(\boldsymbol{n}_\Gamma)_s \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S.
\end{aligned} \tag{23}$$

(We use the so-called incomplete version of the approximation of the viscous terms.)

Interior and boundary penalty and right-hand side forms: Further, we set

$$J_h(\boldsymbol{w},\boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_{ht}^I}\int_\Gamma \sigma[\boldsymbol{w}]\cdot[\boldsymbol{\varphi}_h]\,dS + \sum_{\Gamma\in\mathcal{F}_{ht}^D}\int_\Gamma \sigma\boldsymbol{w}\cdot\boldsymbol{\varphi}_h\,dS, \tag{24}$$

$$\ell_h(\boldsymbol{w},\boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_{ht}^D}\int_\Gamma \sum_{s=1}^{2}\sigma\boldsymbol{w}_B\cdot\boldsymbol{\varphi}_h\,dS. \tag{25}$$

Here $\sigma|_\Gamma = C_W\mu/d(\Gamma)$ and $C_W > 0$ is a sufficiently large constant. The boundary state $\boldsymbol{w}_B$ is defined on the basis of the Dirichlet boundary conditions (7), a), b), d) and extrapolation:

$$\boldsymbol{w}_B = (\rho_D, \rho_D v_{D1}, \rho_D v_{D2}, c_v\rho_D\theta_\Gamma^{(L)} + \frac{1}{2}\rho_D|\boldsymbol{v}_D|^2) \quad \text{on } \Gamma_I, \tag{26}$$

$$\boldsymbol{w}_B = \boldsymbol{w}_\Gamma^{(L)} \quad \text{on } \Gamma_O, \tag{27}$$

$$\boldsymbol{w}_B = (\rho_\Gamma^{(L)}, \rho_\Gamma^{(L)} z_{D1}, \rho_\Gamma^{(L)} z_{D2}, c_v\rho_\Gamma^{(L)}\theta_\Gamma^{(L)} + \frac{1}{2}\rho_\Gamma^{(L)}|\boldsymbol{z}_D|^2) \quad \text{on } \Gamma_{W_t}. \tag{28}$$

Reaction form reads

$$d_h(\boldsymbol{w},\boldsymbol{\varphi}_h) = \sum_{K\in\mathcal{T}_{ht}}\int_K (\boldsymbol{w}\cdot\boldsymbol{\varphi}_h)\,\mathrm{div}\boldsymbol{z}\,\mathrm{d}\boldsymbol{x}. \tag{29}$$

3.1.2 Time discretization

Let us construct a partition $0 = t_0 < t_1 < t_2 \ldots$ of the time interval $[0,T]$ and define the time step $\tau_k = t_{k+1} - t_k$. We use the approximations $\boldsymbol{w}_h(t_n) \approx \boldsymbol{w}_h^n \in \boldsymbol{S}_{ht_n}$, $\boldsymbol{z}(t_n) \approx \boldsymbol{z}^n$, $n = 0,1,\ldots$, and introduce the function $\hat{\boldsymbol{w}}_h^k = \boldsymbol{w}_h^k \circ \mathcal{A}_{t_k} \circ \mathcal{A}_{t_{k+1}}^{-1}$, which is defined in the domain $\Omega_{ht_{k+1}}$. The ALE derivative at time t_{k+1} is approximated by the first- or second-order backward finite difference

$$\frac{D^A\boldsymbol{w}_h}{Dt}(x,t_{k+1}) \approx \frac{\boldsymbol{w}_h^{k+1}(x) - \hat{\boldsymbol{w}}_h^k(x)}{\tau_k}, \tag{30}$$

or

$$\frac{D^A\boldsymbol{w}_h}{Dt}(t_{k+1}) \approx \frac{2\tau_k+\tau_{k-1}}{\tau_k(\tau_k+\tau_{k-1})}\boldsymbol{w}_h^{k+1} - \frac{\tau_k+\tau_{k-1}}{\tau_k\tau_{k-1}}\hat{\boldsymbol{w}}_h^k + \frac{\tau_k}{\tau_{k-1}(\tau_k+\tau_{k-1})}\hat{\boldsymbol{w}}_h^{k-1}. \tag{31}$$

By the symbol $(\cdot,\cdot)$ we shall denote the scalar product in $L^2(\Omega_{ht_{k+1}})$, i.e.

$$(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \int_{\Omega_{ht_{k+1}}} \boldsymbol{w}_h\cdot\boldsymbol{\varphi}_h\,\mathrm{d}\boldsymbol{x}, \tag{32}$$

respectively.

In order to avoid spurious oscillations in the approximate solution in the vicinity of discontinuities or steep gradients, we apply artificial viscosity forms introduced in [7]. They are based on the discontinuity indicator

$$g^k(K) = \int_{\partial K}[\hat{\rho}_h^k]^2\,\mathrm{d}S\big/(h_K|K|^{3/4}), \quad K\in\mathcal{T}_{ht_{k+1}}. \tag{33}$$

By $[\hat{\rho}_h^k]$ we denote the jump of the function $\hat{\rho}_h^k$ on the boundary ∂K and $|K|$ denotes the area of the element K. Then for each $K\in\mathcal{T}_{ht_{k+1}}$ we define the discrete discontinuity

indicator $G^k(K) = 0$ if $g^k(K) < 1$, $G^k(K) = 1$ if $g^k(K) \geq 1$ and the artificial viscosity forms

$$\hat{\beta}_h(\hat{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) = \nu_1 \sum_{K \in \mathcal{T}_{ht_{k+1}}} h_K G^k(K) \int_K \nabla \boldsymbol{w}_h^{k+1} \cdot \nabla \boldsymbol{\varphi}_h \, \mathrm{d}\boldsymbol{x}, \tag{34}$$

$$\hat{J}_h(\hat{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) = \nu_2 \sum_{\Gamma \in \mathcal{F}^I_{ht_{k+1}}} \frac{1}{2}\big(G^k(K_\Gamma^{(L)}) + G^k(K_\Gamma^{(R)})\big) \int_\Gamma [\boldsymbol{w}_h^{k+1}] \cdot [\boldsymbol{\varphi}_h] \, \mathrm{d}\mathcal{S},$$

with parameters ν_1, $\nu_2 = O(1)$.

Finally, by $\overline{\boldsymbol{w}}_h^{k+1}$ we denote the state obtained by the extrapolation:

$$\overline{\boldsymbol{w}}_h^{k+1} = \hat{\boldsymbol{w}}_h^k \quad \text{and} \quad \overline{\boldsymbol{w}}_h^{k+1} = \frac{\tau_k + \tau_{k-1}}{\tau_{k-1}} \hat{\boldsymbol{w}}_h^k - \frac{\tau_k}{\tau_{k-1}} \hat{\boldsymbol{w}}_h^{k-1} \tag{35}$$

in the case of the first-order time discretization and second-order time discretization, respectively.

The resulting scheme has the following form: For each $k = 0, 1, \ldots$ we seek $\boldsymbol{w}_h^{k+1} \in \boldsymbol{S}_{ht_{k+1}}$ such that

$$\begin{aligned} &\Big(\frac{\boldsymbol{w}_h^{k+1} - \hat{\boldsymbol{w}}_h^k}{\tau_k}, \boldsymbol{\varphi}_h\Big) + \hat{b}_h(\overline{\boldsymbol{w}}_h^{k+1}, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \hat{a}_h(\overline{\boldsymbol{w}}_h^{k+1}, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) \\ &+ J_h(\boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) + d_h(\boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \hat{\beta}_h(\hat{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) \\ &+ \hat{J}_h(\hat{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) = \ell(\overline{\boldsymbol{w}}_B^{k+1}, \boldsymbol{\varphi}_h), \quad \forall \boldsymbol{\varphi}_h \in \boldsymbol{S}_{ht_{k+1}}, \end{aligned} \tag{36}$$

in the case of the first-order time discretization. In the case of the second-order time discretization the expression $(\boldsymbol{w}_h^{k+1} - \hat{\boldsymbol{w}}_h^k, \boldsymbol{\varphi}_h)/\tau_k$ is replaced by the approximation (31).

3.2 Discretization of the structural problem

3.2.1 Space semidiscretization

The space semidiscretization of the structural problem will be carried out by the conforming finite element method. By Ω_h^b we denote a polygonal approximation of the domain Ω^b. We construct a triangulation $\mathcal{T}_h^b$ of the domain Ω_h^b formed by a finite number of closed triangles with the following properties:

a) $\overline{\Omega}_h^b = \bigcup_{K \in \mathcal{T}_h^b} K$.

b) The intersection of two different elements $K, K' \in \mathcal{T}_h^b$ is either empty or a common edge of these elements or their common vertex.

c) The vertices lying on $\partial\Omega_h^b$ are elements of $\partial\Omega^b$.

d) The set $\overline{\Gamma}_W^b \cap \overline{\Gamma}_D^b$ is formed by vertices of some elements $K \in \mathcal{T}_h^b$.

Further, by Γ_{Wh}^b and Γ_{Dh}^b we denote the parts of $\partial\Omega_h^b$ approximating Γ_W^b and Γ_D^b.

The approximate solution of the structural problem will be sought in the finite-dimensional space $\boldsymbol{X_h} = X_h \times X_h$, where

$$X_h = \Big\{ v_h \in C(\overline{\Omega}_h^b);\ v_h|_K \in P_s(K),\ \forall K \in \mathcal{T}_h^b \Big\} \tag{37}$$

and $s \geq 1$ is an integer. In $\boldsymbol{X}_h$ we define the subspace $\boldsymbol{V_h} = V_h \times V_h$, where

$$V_h = \Big\{ y_h \in X_h; y_h|_{\overline{\Gamma}_{Dh}^b} = 0 \Big\}. \tag{38}$$

The derivation of the space semidiscretization can be obtained in a standard way. Multiplying system (8) by any test function $y_{hi} \in V_h,\ i = 1, 2$, applying Green's theorem and using the boundary condition (13), we obtain an identity containing the forms defined for $\boldsymbol{u}_h = (u_{h1}, u_{h2}),\ \boldsymbol{y}_h = (y_{h1}, y_{h2}) \in \boldsymbol{X}_h$:

$$a_h(\boldsymbol{u}_h, \boldsymbol{y}_h) = \int_{\Omega_h^b} \lambda^b \mathrm{div}\, \boldsymbol{u}_h \,\mathrm{div}\, \boldsymbol{y}_h \,\mathrm{d}\boldsymbol{X} + 2\int_{\Omega_h^b} \mu^b \sum_{i,j=1}^{2} e_{ij}^b(\boldsymbol{u}_h)\, e_{ij}^b(\boldsymbol{y}_h)\,\mathrm{d}\boldsymbol{X}, \tag{39}$$

and

$$(\boldsymbol{\varphi}, \boldsymbol{\psi})_{\Omega_h^b} = \int_{\Omega_h^b} \boldsymbol{\varphi}\cdot\boldsymbol{\psi}\,\mathrm{d}\boldsymbol{X}, \quad (\boldsymbol{\varphi}, \boldsymbol{\psi})_{\Gamma_{Wh}} = \int_{\Gamma_{Wh}} \boldsymbol{\varphi}\cdot\boldsymbol{\psi}\,\mathrm{d}S. \tag{40}$$

We shall use the approximation $\boldsymbol{T}_h^{\boldsymbol{n}} \approx \boldsymbol{T}^{\boldsymbol{n}}$ and the notation $\boldsymbol{u}_h'(t) = \frac{\partial \boldsymbol{u}_h(t)}{\partial t}$ and $\boldsymbol{u}_h''(t) = \frac{\partial^2 \boldsymbol{u}_h(t)}{\partial t^2}$. Then we define the *approximate solution of the structural problem* as a function $t \in [0, T] \to \boldsymbol{u}_h(t) \in \boldsymbol{V}_h$ such that there exist the derivatives $\boldsymbol{u}_h'(t)$, $\boldsymbol{u}_h''(t)$ and the identity

$$(\varrho^b \boldsymbol{u}_h''(t), \boldsymbol{y}_h)_{\Omega_h^b} + (C\varrho^b \boldsymbol{u}_h'(t), \boldsymbol{y}_h)_{\Omega_h^b} + a_h(\boldsymbol{u}_h(t), \boldsymbol{y}_h) = (\boldsymbol{T}_h^{\boldsymbol{n}}(t), \boldsymbol{y}_h)_{\Gamma_{Wh}},$$
$$\forall \boldsymbol{y}_h \in \boldsymbol{V}_h, \quad \forall t \in (0, T), \tag{41}$$

and the initial conditions

$$\boldsymbol{u}_h(\boldsymbol{X}, 0) = 0, \quad \boldsymbol{u}_h'(\boldsymbol{X}, 0) = 0, \quad \boldsymbol{X} \in \Omega_h^b. \tag{42}$$

are satisfied.

The discrete problem (41), (42) is equivalent to the solution of a system of ordinary differential equations. Let functions $\varphi_1, \ldots, \varphi_m$ form a basis of the space V_h. Then the system of $n = 2m$ of the vector functions $(\varphi_1, 0), \ldots, (\varphi_m, 0), (0, \varphi_1), \ldots, (0, \varphi_m)$ form a basis of the space $\boldsymbol{V}_h$. Let us denote them by $\boldsymbol{\varphi}_1, \ldots \boldsymbol{\varphi}_n$. Then the approximate solution $\boldsymbol{u}_h$ can be expressed in the form

$$\boldsymbol{u}_h(t) = \sum_{j=1}^{n} p_j(t)\boldsymbol{\varphi}_j, \quad t \in [0, T]. \tag{43}$$

Let us set $\boldsymbol{p}(t) = (p_1(t), \ldots, p_n(t))$. Using $\boldsymbol{\varphi}_j,\ j = 1, \ldots, n$, as test functions in (41), we get the following system of ordinary differential equations

$$\mathbb{M}\boldsymbol{p}'' = \boldsymbol{G} - \mathbb{K}\boldsymbol{p} - C\mathbb{M}\boldsymbol{p}', \tag{44}$$

where $\mathbb{M} = (m_{ij})_{i,j=1}^n$ is the mass matrix and $\mathbb{K} = (k_{ij})_{i,j=1}^n$ is the stiffness matrix with the elements $m_{ij} = (\rho^b \boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j)$ and $k_{ij} = a_h(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j)$, respectively. The aerodynamic force vector $\boldsymbol{G} = \boldsymbol{G}(t) = (G_1(t), \ldots, G_n(t))^T$ has the components $G_i(t) = (\boldsymbol{T}_h^n(t), \boldsymbol{\varphi}_i)_{\Gamma_{Wh}},\ i = 1, \ldots, n$. System (44) is equipped with the initial conditions

$$p_j(0) = 0, \quad p_j'(0) = 0, \quad j = 1, \ldots, n. \tag{45}$$

3.2.2 Time discretization of the structural problem

The discrete initial value problem (44), (45) is solved by the Newmark method ([4]). We consider the partition of the time interval $[0, T]$ formed by the time instants $0 = t_0 < t_1 < \ldots$ introduced in Section 3.1.2. Let us set $\boldsymbol{p}_0 = 0, \boldsymbol{z}_0 = 0, \boldsymbol{G}_k = \boldsymbol{G}(t_k)$,

and introduce the approximations $\boldsymbol{p}_k \approx \boldsymbol{p}(t_k)$ and $\boldsymbol{q}_k \approx \boldsymbol{p}'(t_k)$ for $k = 1, 2, \dots$. The Newmark scheme can be written in the form

$$\begin{aligned} \boldsymbol{p}_{k+1} &= \boldsymbol{p}_k + \tau_k \boldsymbol{q}_k + \tau_k^2 \Big(\beta \left(\mathbb{M}^{-1} \boldsymbol{G}_{k+1} - \mathbb{M}^{-1} \mathbb{K} \boldsymbol{p}_{k+1} - C \boldsymbol{q}_{k+1} \right) \\ &\quad + \left(\frac{1}{2} - \beta \right) \left(\mathbb{M}^{-1} \boldsymbol{G}_k - \mathbb{M}^{-1} \mathbb{K} \boldsymbol{p}_k - C \boldsymbol{q}_k \right) \Big), \end{aligned} \tag{46}$$

$$\begin{aligned} \boldsymbol{q}_{k+1} &= \boldsymbol{q}_k + \tau_k \Big(\gamma \left(\mathbb{M}^{-1} \boldsymbol{G}_{k+1} - \mathbb{M}^{-1} \mathbb{K} \boldsymbol{p}_{k+1} - C \boldsymbol{q}_{k+1} \right) \\ &\quad + (1 - \gamma) \left(\mathbb{M}^{-1} \boldsymbol{G}_k - \mathbb{M}^{-1} \mathbb{K} \boldsymbol{p}_k - C \boldsymbol{q}_k \right) \Big), \end{aligned} \tag{47}$$

where $\beta, \gamma \in \mathbb{R}$ are parameters. From equation (47) we get

$$\begin{aligned} \boldsymbol{q}_{k+1} &= \frac{1}{1 + C\gamma\tau_k} \Big(\boldsymbol{q}_k + \tau_k \Big(\gamma \left(\mathbb{M}^{-1} \boldsymbol{G}_{k+1} - \mathbb{M}^{-1} \mathbb{K} \boldsymbol{p}_{k+1} \right) \\ &\quad + (1 - \gamma) \left(\mathbb{M}^{-1} \boldsymbol{G}_k - \mathbb{M}^{-1} \mathbb{K} \boldsymbol{p}_k - C \boldsymbol{q}_k \right) \Big) \Big). \end{aligned} \tag{48}$$

The substitution of (48) in (46) yields the relation which can be written in the form

$$\begin{aligned} \left(\mathbb{I} + \xi_k \mathbb{M}^{-1} \mathbb{K} \right) \boldsymbol{p}_{k+1} &= \boldsymbol{p}_k + (\tau_k - C\xi_k) \, \boldsymbol{q}_k + \xi_k \mathbb{M}^{-1} \boldsymbol{G}_{k+1} + \\ &+ \left(C \left(\gamma - 1 \right) \xi_k \tau_k + \left(\frac{1}{2} - \beta \right) \tau_k^2 \right) \left(\mathbb{M}^{-1} \boldsymbol{G}_k - \mathbb{M}^{-1} \mathbb{K} \boldsymbol{p}_k - C \boldsymbol{q}_k \right). \end{aligned} \tag{49}$$

where we set for the sake of simplicity

$$\xi_k = \beta \tau_k^2 \left(1 - \frac{C\gamma\tau_k}{1 + C\gamma\tau_k} \right) = \frac{\beta \tau_k^2}{1 + C\gamma\tau_k}. \tag{50}$$

If $\boldsymbol{p}_k$ and $\boldsymbol{q}_k$ are known, then $\boldsymbol{p}_{k+1}$ is obtained from system (49) and afterwards $\boldsymbol{q}_{k+1}$ is computed from (48).

In numerical examples presented in Section 5, the parameters $\beta = 1/4$ and $\gamma = 1/2$ were used. This choice yields the Newmark method of the second order.

4 Realization of the coupled FSI problem

In this section we shall describe the algorithm of the numerical realization of the complete fluid-structure interaction problem.

4.1 Construction of the ALE mapping for fluid

The ALE mapping is constructed with the aid of an artificial stationary elasticity problem. We seek $\boldsymbol{d} = (d_1, d_2)$ defined in Ω_0 as a solution of the elastostatic system

$$\sum_{j=1}^{2} \frac{\partial \tau_{ij}^a}{\partial x_j} = 0 \quad \text{in } \Omega_0, \quad i = 1, 2, \tag{51}$$

where τ_{ij}^a are the components of the artificial stress tensor

$$\tau_{ij}^a = \lambda^a \mathrm{div}\, \boldsymbol{d} \, \delta_{ij} + 2\mu^a e_{ij}^a, \quad e_{ij}^a(\boldsymbol{d}) = \frac{1}{2} \left(\frac{\partial d_i}{\partial x_j} + \frac{\partial d_j}{\partial x_i} \right), \quad i, j = 1, 2. \tag{52}$$

The Lamé coefficients λ^a and μ^a are related to the artificial Young modulus E^a and to the artificial Poisson number σ_a as in (11). The boundary conditions for $\boldsymbol{d}$ are prescribed by

$$\boldsymbol{d}|_{\Gamma_I \cup \Gamma_O} = 0,\ \boldsymbol{d}|_{\Gamma_{W0h} \setminus \Gamma_{Wh}} = 0,\ \boldsymbol{d}(\boldsymbol{x}, t) = \boldsymbol{u}(\boldsymbol{x}, t),\ \boldsymbol{x} \in \Gamma_{Wh}. \tag{53}$$

The solution of (51) gives us the ALE mapping of $\overline{\Omega}_0$ onto $\overline{\Omega}_t$ in the form

$$\mathcal{A}_t(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{d}(\boldsymbol{x}, t), \quad \boldsymbol{x} \in \overline{\Omega}_0, \tag{54}$$

for each time t.

System (51) is discretized by the conforming piecewise linear finite elements on the mesh $\mathcal{T}_{h0}$ used for computing the flow field in the beginning of the computational process in the polygonal approximation Ω_{h0} of the domain Ω_0. The use of linear finite elements is sufficient, because we need only to know the movement of the points of the mesh.

In our computations we choose the Lamé coefficients λ^a and μ^a as constants corresponding to the Young modulus and Poisson ratio $E^a = 10000$ and $\sigma^a = 0.45$.

If the displacement $\boldsymbol{d}_h$ is computed at time t_{k+1}, then in view of (54), the approximation of the ALE mapping is obtained in the form

$$\mathcal{A}_{t_{k+1}h}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{d}_h(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega_{0h}. \tag{55}$$

The knowledge of the ALE mapping at the time instants t_{k-1}, t_k, t_{k+1} allows us to approximate the domain velocity with the aid of the second-order backward difference formula

$$\boldsymbol{z}_h^{k+1}(\boldsymbol{x}) = \frac{3\boldsymbol{x} - 4\mathcal{A}_{t_k h}(\mathcal{A}^{-1}_{t_{k+1}h}(\boldsymbol{x})) + \mathcal{A}_{t_{k-1}h}(\mathcal{A}^{-1}_{t_{k+1}h}(\boldsymbol{x}))}{2\tau}, \quad \boldsymbol{x} \in \Omega_{t_{k+1}h}. \tag{56}$$

4.2 Coupling procedure

In the solution of the complete coupled fluid-structure interaction problem it is necessary to apply a suitable coupling procedure. See, e.g. [3] for a general framework. Here we apply the following algorithm.

1. Assume that the approximate solution of the flow problem on the time level t_k is known as well as the deformation of the structure $\boldsymbol{u}_{h,k}$.
2. Set $\boldsymbol{u}^0_{h,k+1} := \boldsymbol{u}_{h,k},\ l := 1$ and apply the iterative process:
   (a) Compute the stress tensor τ^f_{ij} and the aerodynamical force acting on the structure and transform it to the interface Γ^b_{Wh}.
   (b) Solve the elasticity problem, compute the deformation $\boldsymbol{u}^l_{h,k+1}$ at time t_{k+1} and approximate the domain $\Omega^l_{ht_{k+1}}$.
   (c) Determine the ALE mapping $\mathcal{A}^l_{t_{k+1}h}$ and approximate the domain velocity $\boldsymbol{z}^l_{h,k+1}$.
   (d) Solve the flow problem on the approximation of $\Omega^l_{ht_{k+1}}$.
   (e) If the variation of the displacement $\boldsymbol{u}^l_{h,k+1}$ and $\boldsymbol{u}^{l-1}_{h,k+1}$ is larger than the prescribed tolerance, go to a) and $l := l + 1$. Else $k := k + 1$ and goto 2).

This represents the so-called strong coupling. If in the step e) we set $k := k + 1$ and go to 2) already in the case when $l = 1$, then we get the weak (loose) coupling.

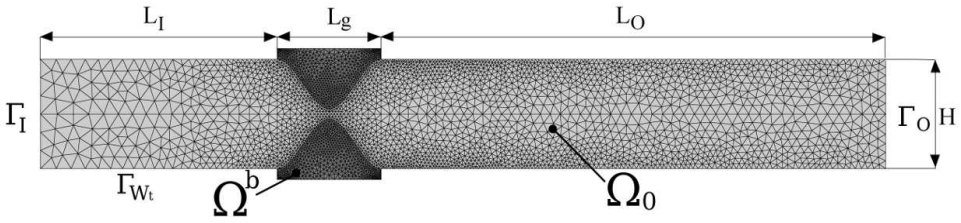


Figure 1: Computational domain at time $t = 0$ with a finite element mesh and the description of its size: $L_I = 50$ mm, $L_g = 15.4$ mm, $L_O = 94.6$ mm, $H = 16$ mm. The width of the channel in the narrowest part is 1.6 mm.

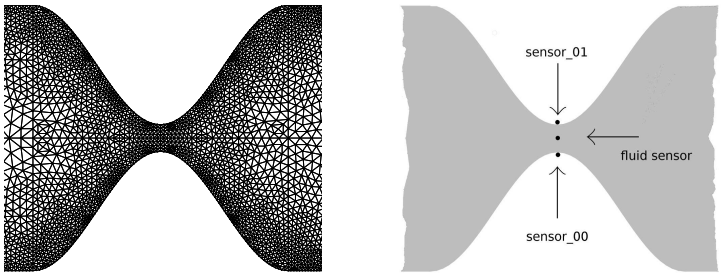


Figure 2: Detail of the flow computational mesh at time $t = 0$ near the narrowest part of the channel and the position of some sensors used in analysis.

5 Numerical examples

In order to demonstrate the applicability of the developed method, we present here results of some numerical experiments.

We consider a model of flow through a channel with two bumps which represent time dependent boundaries between the flow and a simplified model of vocal folds (see Figures 1 and 2). The numerical experiments were carried out for the following data: magnitude of the inlet velocity $v_{in} = 4$ m/s, the fluid viscosity $\mu = 15 \cdot 10^{-6}\ \mathrm{kg\,m^{-1}\,s^{-1}}$, the inlet density $\rho_{in} = 1.225\ \mathrm{kg\,m^{-3}}$, the outlet pressure $p_{out} = 97611$ Pa, the Reynolds number $Re = \rho_{in} v_{in} H/\mu = 5227$, heat conduction coefficient $k = 2.428 \cdot 10^{-2}\,\mathrm{kg\,m\,s^{-2}\,K^{-1}}$, the specific heat $c_v = 721.428\,\mathrm{m^2\,s^{-2}\,K^{-1}}$, the Poisson adiabatic constant $\gamma = 1.4$. The inlet Mach number is $M_{in} = 0.012$. The Young modulus and the Poisson ratio have values $E^b = 25000$ Pa and $\sigma^b = 0.4$, respectively, the structural damping coefficient is equal to the constant $C = 100\ \mathrm{s^{-1}}$ and the material density $\rho^b = 1040\,\mathrm{kg\,m^{-3}}$. The quadratic ($r = 2$) and linear ($s = 1$) elements were used for the approximation of flow and structural problem, respectively.

Figure 1 shows the situation at the initial time $t = 0$ the flow computational mesh consisting of 5398 elements and the structure computational mesh with 1998 elements. In Figure 2 we see a detail of the flow mesh near the narrowest part of the channel at the initial time and the positions of sensor points used in the analysis.

First we tested the influence of the density of the computational meshes on the oscillations of the pressure averaged over the outlet Γ_O and the corresponding Fourier analysis. We consider three successively refined meshes. Figure 3 shows the behaviour

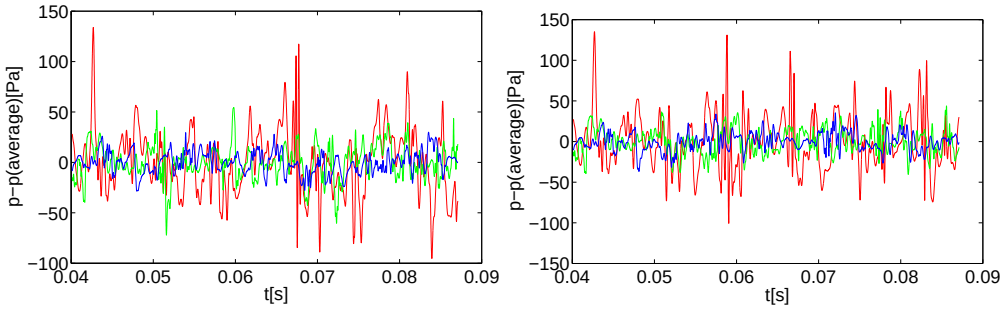


Figure 3: Dependence of the quantity p_{av} computed on three meshes: strong coupling (left), weak coupling (right).

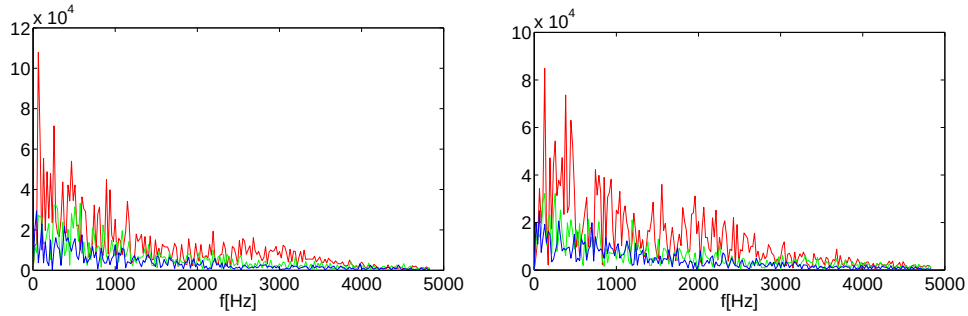


Figure 4: Fourier analysis of the quantity p_{av} computed on three meshes: strong coupling (left), weak coupling (right).

of the quantity

$$p_{av}(t) = \int_{\Gamma_O} \left(p(x,t) - \frac{1}{T} \int_0^T p(x,t)\,\mathrm{d}t \right) / \int_{\Gamma_O} \mathrm{d}S. \tag{57}$$

in dependence on time, computed on the flow/structure meshes with 5398/1998 elements (red), 10130/2806 elements (green) and 20484/4076 elements (blue) with the aid of the strong coupling (left) and the weak coupling (right). Figure 4 shows the corresponding Fourier analysis. During the successive mesh refinement one can observe the convergence tendency manifested by the decrease of the magnitude of the quantity p_{av} fluctuations and the decrease of the magnitude of the Fourier spectra. No peaks related to any basic acoustic modes of vibration in the channel were identified in the spectra. The difference between the results obtained by the strong and weak coupling is not too large. The main difference is in a higher stability of the strong coupling during solving the problem on a long time interval. On the other hand, the strong coupling requires naturally longer CPU time.

Flow-induced deformations of the vocal folds model with the computational mesh and the velocity field near the vocal folds are shown in Figure 5 at several time instants. Similarly as in experimental study presented in [10] and [11], we can see the Coanda effect represented by the attachment of the main stream (jet) successively to the upper and lower wall of the glottis and formation of large scale vortices behind the glottis. The same effects can be observed in numerical results from [12]. The character of the vocal folds vibration can be indicated in Figure 7, which shows the displacements of the sensor points on the vocal folds surface (marked in Figure 2) and the fluid pressure

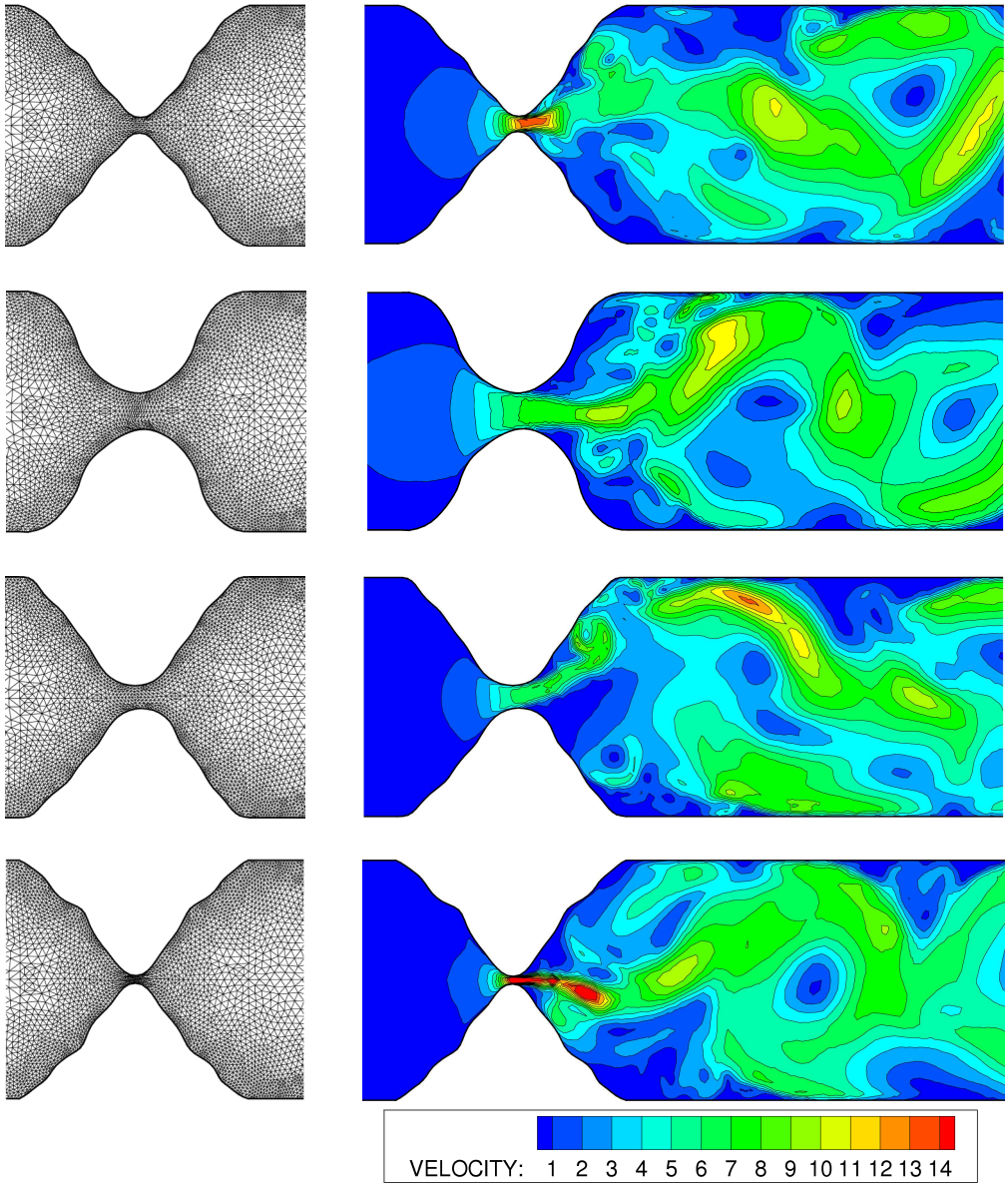


Figure 5: Detail of the mesh and the velocity distribution in the vicinity of the narrowest part of the channel at time instants $t = 0.2056$, 0.2072, 0.2088, 0.2104 s.

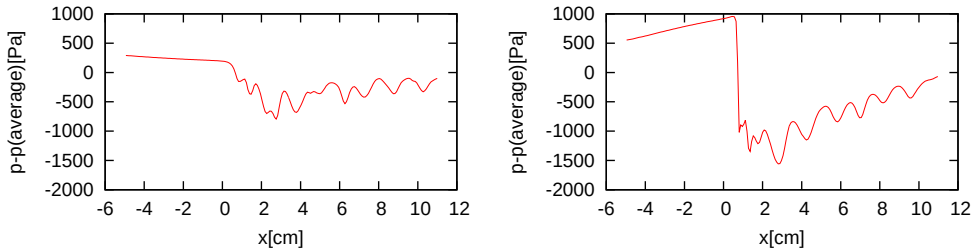


Figure 6: The difference between the pressure on the centreline of the channel and the averaged outlet pressure corresponding to time instants $t = 0.2072$ (left) and $t = 0.2104$ (right).

fluctuations in the middle of the gap as well as the Fourier analysis of the signals. The vocal folds vibrations are not symmetric due to the Coanda effect and are composed of the fundamental horizontal mode of vibration with the corresponding eigen-frequency 113 Hz and by the higher eigenmode with the eigenfrequency 439 Hz. The increase of horizontal vibrations due to the aeroelastic instability of the system results in a fast decrease of the glottal gap. Similarly as in [12], the displacement d_x in the vertical direction is larger than the vertical displacement d_y.

At about $t = 0.2$ s, when the gap is nearly closed, the fluid mesh deformation in this region is too high and the numerical simulation stopped. The dominant peak at 439 Hz in the spectrum of the pressure signal corresponds well to the vertical oscillations of the glottal gap, while the importance of the lower frequency 113 Hz associated with the horizontal vocal folds motion is in the pressure fluctuations negligible. The modeled flow-induced instability of the vocal folds is called phonation onset followed in reality by a complete closing of the glottis and consequently by the vocal folds collisions producing the voice source acoustic signal.

Figure 6 shows the distribution of the difference between the pressure on the centreline of the channel and the averaged outlet pressure , corresponding to time instants $t = 0.2072$ (maximal opening of the glottis) and $t = 0.2104$ (minimal opening of the glottis). The inlet pressure varies approximately between 300 and 500 Pa, which corresponds to subglottal pressure for humans during phonation. The pressure drop corresponds to the narrowest part of the channel similarly as in the paper [12]. Pressure oscillations behind the vocal folds are caused by propagating vortices.

6 Conclusion

We have presented a robust higher-order method for the numerical simulation of the interaction of compressible flow with elastic structures with applications to the computation of flow-induced vibrations of vocal folds during phonation. It is based on several important ingredients:

- the ALE method applied to the compressible Navier-Stokes equations,
- the application of the discontinuous Galerkin method for the space discretization and semi-implicit linearized time discretization,

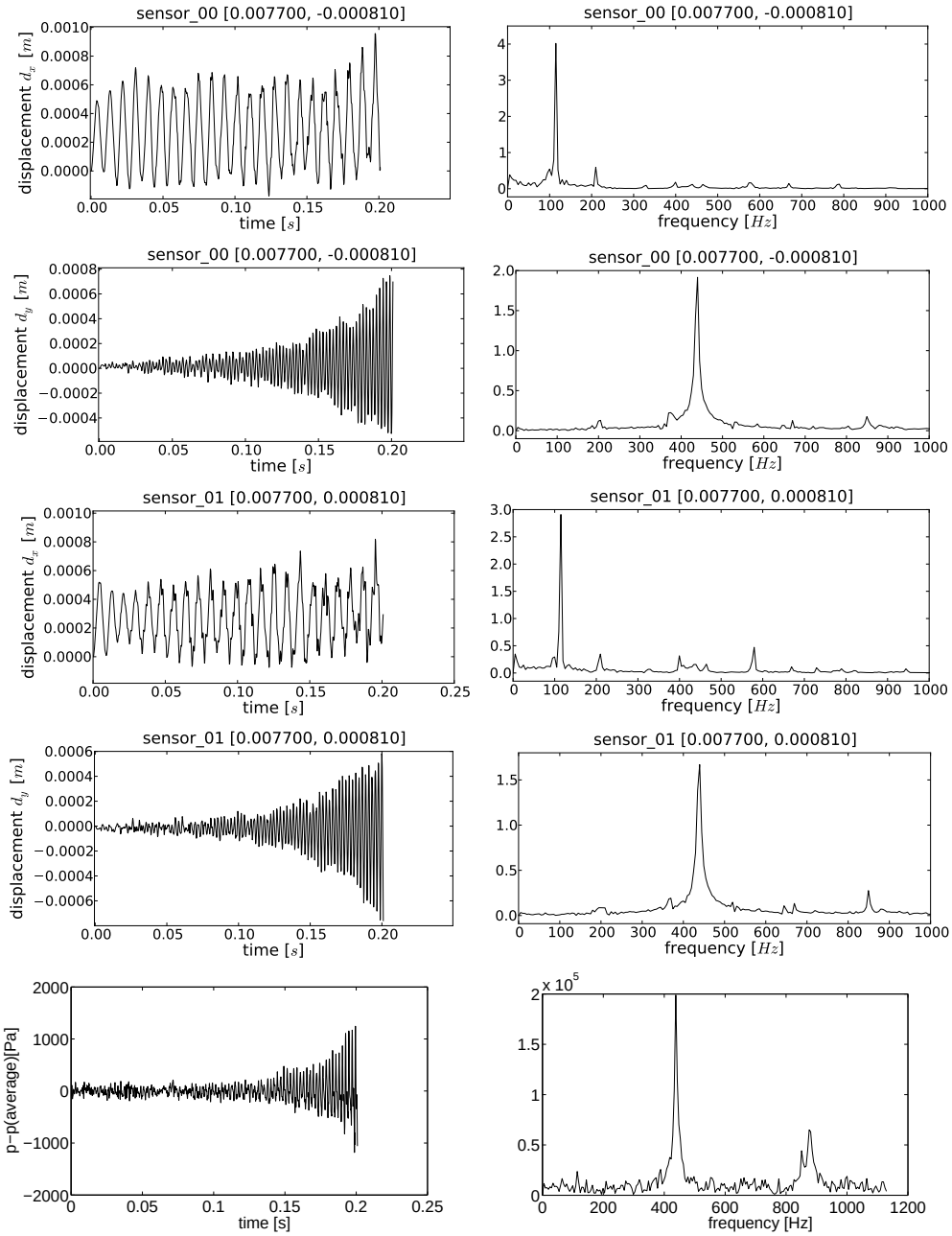


Figure 7: Vibrations of sensor points lying on the boundary of the vocal folds and the pressure oscillations in the middle of the gap (left), and the corresponding Fourier analyses (right).

- the use of conforming finite elements for the space discretization and of the Newmark method for the time discretization of the elasticity problem,
- technique for the construction of the ALE mapping,
- the application of coupling algorithms for the realization of the coupled FSI problem.

The numerical tests and experiments show that the developed method can be applied to the numerical solution of the interaction of compressible flow and elastic structures with applications to the simulation of air flow through vocal folds.

The computational results are qualitatively similar to other computations (cf., e.g., [12]) and wind-tunnel experiments ([10], [11]).

Future work will be concentrated on the following topics:

- further analysis of the robustness and accuracy of the method with respect to the Mach number and Reynolds number with the use of various types of the vocal folds geometry,
- quantitative examination of the worked out technique on suitable test problems and the comparison with results of other methods (if they are available),
- investigation of various types of boundary conditions,
- the realization of a remeshing in the case of closing the glottal channel during the oscillation period of the channel walls,
- the use of nonlinear elasticity models including vocal folds collision,
- the use of a suitable turbulence model,
- the identification of the acoustic signal.

Acknowledgements This work was supported by the grants No. 13-00522S (M. Feistauer, V. Kučera) and P101/11/0207 (J. Horáček) of the Czech Science Foundation, and by the grants SVV-2010-261316 and GAChU 549912 financed by the Charles University in Prague (J. Hasnedlová-Prokopová and A. Kosík).

References

[1] F. Alipour, I. R. Titze: Combined simulation of two-dimensional airflow and vocal fold vibration. *Vocal fold physiology, controlling complexity and chaos*, San Diego (1996).

[2] F. Alipour, Ch. Brücker, D.D. Cook, A. Gömmel, M. Kaltenbacher, W. Mattheus, L. Mongeau, E. Nauman, R. Schwarze, I. Tokuda and S. Zörner: Mathematical models and numerical schemes for the simulation of human phonation. *Current Bioinformatics* **6**, 323–343 (2011).

[3] S. Badia, R. Codina: On some fluid-structure iterative algorithms using pressure segregation methods. Application to aeroelasticity. *Int. J. Numer. Meth. Engng* **72**, 46–71 (2007).

[4] A. Curnier: *Computational Methods in Solid Mechanics*. Kluwer Academic Publishing Group, Dodrecht (1994).

[5] M. Feistauer, J. Felcman and I. Straškraba: *Mathematical and Computational Methods for Compressible Flow*. Clarendon Press, Oxford (2003).

[6] M. Feistauer, J. Horáček, V. Kučera, J. Prokopová: On numerical solution of compressible flow in time-dependent domains. *Mathematica Bohemica* **137**, 1–16 (2011).

[7] M. Feistauer and V. Kučera: On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.*, **224**, 208–221 (2007).

[8] J. Horáček and J. G. Švec: Aeroelastic model of vocal-fold-shaped vibrating element for studying the phonation threshold. *J. Fluids Struct.*, **16**, 931–955 (2002).

[9] J. Horáček, P. Šidlof and J. G. Švec: Numerical simulation of self-oscillations of human vocal folds with Hertz model of impact forces. *J. Fluids Struct.*, **20**, 853–69 (2005).

[10] J. Horáček, P. Šidlof, V. Uruba, J. Veselý, V. Radolf and V. Bula: Coherent structures in the flow inside a model of human vocal tract with self-oscillating vocal folds. *Acta Technica*, **55**, 327–343 (2010).

[11] J. Horáček, V. Uruba, V. Radolf, J. Veselý and V. Bula: Airflow visualization in a model of human glottis near the self-oscillating vocal folds model. *Appl. Comput. Mech.*, **5**, 21–28 (2011).

[12] H. Luo, R. Mittal, X. Zheng, S.A. Bielamowicz, R.J. Walsh, J.K. Hahn: An immersed-boundary method for flow-structure interaction in biological systems with application to phonation. *J. Comput. Phys.*, **227**, 9303–9332 (2008).

[13] T. Nomura and T.J.R. Hughes: An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. *Comput. Methods Appl. Mech. Engrg.*, **95**, 115–138 (1992).

[14] P. Punčochářová-Pořízková, K. Kozel and J. Horáček: Simulation of unsteady compressible flow in a channel with vibrating walls - influence of the frequency. *Computers and Fluids*, **46**, 404–410 (2011).

[15] P. Sváček, M. Feistauer and J. Horáček: Numerical simulation of flow induced airfoil vibrations with large amplitudes. *J. Fluids Struct.*, **23**, 391–411 (2007).

[16] I.R. Titze: *Principles of Voice Production*. National Centre for Voice and Speech, Iowa City (2000).

[17] I.R. Titze, *The Myoelastic Aerodynamic Theory of Phonation*. National Centre for Voice and Speech, Denver and Iowa City (2006).

[18] M. P. De Vries, H. K. Schutte, A. E. P. Veldman and G. J. Verkerke: Glottal flow through a two-mass model: comparison of Navier-Stokes solutions with simplified models. *J. Acoust. Soc. Am.*, **111** (4), 1847–53 (2002).

[19] Z. Zhang, J. Neubauer and D. A. Berry: Physical mechanisms of phonation onset: A linear stability analysis of an aeroelastic continuum model of phonation. *J. Acoust. Soc. Am.*, **122**, 2279–2295 (2007).