Numerical Analysis of Higher Order Discontinuous Galerkin Finite Element Methods

Ralf Hartmann
Institute of Aerodynamics and Flow Technology
DLR (German Aerospace Center)
Lilienthalplatz 7, 38108 Braunschweig, Germany

Contents

1 Introduction

In recent years higher order discretization methods are of increasing importance in computational fluid dynamics. In particular, for compressible flows as considered in aerodynamic flow simulations the development of high order accurate, stable and efficient discretization methods is a hot topic. The European project ADIGMA [1] concentrates and focuses the European research effort on the development of these methods to aerodynamic applications in industry.

1.1 Higher order discretization methods

The discretization error of higher order discretization methods decreases with a higher order in the mesh size h than that of low order schemes. A discretization method is of order n if the discretization error behaves like $\mathcal{O}(h^n)$. When halving the mesh size h by performing one global mesh refinement step the discretization error decreases by e.g. a factor of 16 for a forth order scheme in comparison to a factor of only 4 for a second order scheme. As a consequence a required accuracy in the solution can be obtained on coarser meshes and in general with less degrees of freedom and computing resources required than for second order schemes.

The advantages of higher order methods over second order methods are particularly important in aerodynamic flow simulations:

- Higher order methods allow a significantly improved resolution of flow features like vortices in comparison to second order methods. This is particularly important for the simulation of *vortex creation* and *blade-vortex interaction* at helicopter rotor blades as well as for the simulation of *wake-vortices* behind transport aircrafts. Current *second order* based flow solvers are too dissipative leading to strong damping of flow features and a premature dissipation of vortices in numerical simulation although being still present in reality. In contrast to that, the vortices can be well resolved and accurately tracked for a significantly longer time/distance by higher order methods, see Figure 20. This is particular important for improving the shape and control of helicopter rotor blades which is required for reducing helicopter noise. It is even more important for optimizing aircraft shapes in order to reduce wake-vortices and to cause wake-vortices to interact and vanish earlier, which is required for reducing the minimum distance of aircrafts at take-off or landing at airports, eventually increasing the transport capacity of airports.

- Most computing resources of current second order flow solvers are required for resolving viscous and turbulent boundary layers, represented by the fact that typically about 50% of all mesh points are concentrated near the boundary layers. As higher order methods are particularly suited for resolving boundary layers, the enormous number of mesh points required for resolving them can be dramatically reduced. In fact, for laminar flows it has been shown, see [31] or Figure 14 and Table 1, that a 4th order discretization requires 3 elements in the boundary layer to give the same accuracy as a 2nd order discretization with 36 elements in the boundary layer. This promises a significant reduction of mesh sizes potentially allowing for larg*er*-scale application with the same computing resources than with current flow solver technologies.

The maximum order one encounters when applying a higher order discretization method to a particular problem depends on the smoothness of the solution. Whereas for (arbitrary) smooth solutions a method of order n shows in fact a discretization error of order $\mathcal{O}(h^n)$, the order of convergence is reduced for non-smooth solutions.

In general, solutions are not smooth in the whole computational domain; in fact, they might exhibit some irregularities like shocks or singularities in some parts of the domain but are perfectly smooth in other parts. In order to fully exploit the regularity of the solution the order of the

discretization should be adapted to the smoothness of the solution. Here, the general idea is to employ discretization methods of higher order in smooth parts of the solution and of low order in irregular parts of the solution (p-refinement). Together with local mesh refinement (h-refinement) this leads to the so-called hp-refinement.

We note, that this lecture is concerned with higher order discretization methods, it does not cover the wide field of h-, p- and hp-adaptation methods. Having in mind that in practice solutions are of different regularity in different parts of the domain which would require the use of hp-adaptive methods, in this lecture we always assume that solutions are of a specific regularity globally, i.e. in the *whole* computational domain.

1.2 Discontinuous Galerkin discretizations

Over the last about ten years the development of the discontinuous Galerkin (DG) methods has attracted more and more research groups all over the world, significantly increasing the pace of the development of these methods, to work on DG [5, 6, 9, 10, 16, 17, 18, 23, 30, 31, 32, 38, 41, 42]. In fact, it can be observed that to an increasing extent discontinuous Galerkin methods are now applied to problems which traditionally where solved using finite volume methods. The reason for this trend can be identified in several advantages of the discontinuous Galerkin methods over finite volume methods. Second order finite volume methods are achieved by employing a second order accurate reconstruction. The extension of a second order finite volume scheme to a (theoretically) third order scheme requires a third order accurate reconstruction which on unstructured meshes is very cumbersome and which in practice shows deterioration of order. On unstructured meshes finite volume methods of even higher order are virtually impossible. These difficulties bound the order of numerical computations in industrial applications to second order. In contrast to this, the order of discontinuous Galerkin methods, applied to problems with regular solutions, depends on the degree of the approximating polynomials only which can easily be increased, dramatically simplifying the use of higher order methods on unstructured meshes. Furthermore, the stencil of most discontinuous Galerkin schemes is minimal in the sense that each element communicates only with its direct neighbors. In contrast to the increasing number of elements or mesh points communicating for increasing accuracy of finite volume methods, the inter-element communication of discontinuous Galerkin methods is the same for any order. The compactness of the discontinuous Galerkin method has clear advantages in parallelization, which does not require additional element layers at partition boundaries. Also due to simple communication at element interfaces, elements with so-called 'hanging nodes' can be treated just as easily as elements without hanging nodes, a fact that simplifies local mesh refinement (h-refinement). In addition to this, the communication at element interfaces is identical for any order of the method which simplifies the use of methods of differing orders in adjacent elements. This allows for the variation of the order of the numerical scheme over the computational domain, which in combination with h-refinement leads to the so-called hp-refinement algorithms, where p-refinement denotes the variation of the polynomial degree p.

1.3 Numerical analysis of finite element methods

Discontinuous Galerkin methods are a special type of *finite element methods*. Thus, there are many powerful tools of finite element analysis available which – with some DG specific modifications – can be applied to the numerical analysis of discontinuous Galerkin discretizations.

Consider, for simplicity, a linear partial differential equation of the form

$$Lu = f \quad \text{in } \Omega, \qquad\qquad Bu = g \quad \text{on } \Gamma, \tag{1}$$

with $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$, where L denotes a linear differential operator on Ω, and B denotes a linear differential (boundary) operator on the boundary Γ of domain Ω.

Furthermore, consider the following finite element discretization: find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_h. \tag{2}$$

Here, V_h is a discrete function space and $B_h : V \times V \to \mathbb{R}$ is a bilinear form, where V is an appropriately chosen function space such that $V_h \subset V$ and $u \in V$, where u is the exact, i.e. analytical, solution to (1). Then, some of the most important topics in the numerical analysis of this discretization are the following:

- **Consistency:** Does relation (2) still hold when we replace u_h by the exact solution u to the differential equation (1)? I.e. do we have

  $$B_h(u, v) = F_h(v) \quad \forall v \in V. \tag{3}$$

  This answers the question: Do we solve the right equations?

  If the discretization is consistent, we can subtract (2) from (3) for $v_h \in V_h \subset V$ which immediately gives us the so-called *Galerkin orthogonality*:

  $$B_h(u - u_h, v_h) = 0 \quad \forall v_h \in V_h, \tag{4}$$

  which means that the discretization error $e = u - u_h$ is orthogonal (with respect to the bilinear form B_h) to the discrete test space V_h. This is a basic property of all Galerkin finite element methods, among them e.g. the standard Galerkin (or continuous) finite element method as well as the discontinuous Galerkin finite element method.

- **Coercivity & Stability**: Is there a constant $\gamma > 0$, such that

  $$B_h(v_h, v_h) \geq \gamma |\!|\!| v_h |\!|\!|^2 \quad \forall v_h \in V_h, \tag{5}$$

  where $|\!|\!| v |\!|\!|$ is a norm (or seminorm) on V. Furthermore, we assume that F_h in (2) is continuous, i.e. there is a $C_F > 0$ such that

  $$F_h(v_h) \leq C_F |\!|\!| v_h |\!|\!| \quad \forall v_h \in V_h. \tag{6}$$

  Then, for the solution $u_h \in V_h$ to the discrete problem (2) we obtain

  $$\gamma |\!|\!| u_h |\!|\!|^2 \leq B_h(u_h, u_h) = F_h(u_h) \leq C_F |\!|\!| u_h |\!|\!|, \tag{7}$$

  and thus $|\!|\!| u_h |\!|\!| \leq \frac{C_F}{\gamma}$, i.e. we have control over all terms occurring in $|\!|\!| u_h |\!|\!|$. If $|\!|\!| \cdot |\!|\!|$ is a norm (and not only a semi-norm) on the space in which weak solutions to (1) are to be searched then the discretization (2) is stable.

- **Convergence (Order of convergence):** Does the discrete solution u_h converge to the exact solution u? What is the order of convergence, i.e., given a solution u with $\|u\|_{**} < \infty$, what is (the maximum) r such that

  $$\|u - u_h\|_* \leq c h^r \|u\|_{**}, \tag{8}$$

  where $\|\cdot\|_*$ is an appropriate (global) norm to measure the error in, e.g. $\|\cdot\|_* = \|\cdot\|_{L^2}$, and $\|\cdot\|_{**}$ is a norm on (possibly a subset of) V.

- **Convergence in specific target quantities $J(\cdot)$:** Instead of measuring the error in terms of (global) norms, one might be interested in the error measured in terms of some physically relevant quantity. Let $J : V \to \mathbb{R}$ be a functional, like e.g. a (weighted) mean value of the solution on Ω or on parts of the boundary $\partial\Omega$. Then we are interested in the order of

convergence with respect to $J(\cdot)$, i.e. given a u with $\|u\|_{**} < \infty$, what is (the maximum) s such that

$$|J(u) - J(u_h)| \leq ch^s \|u\|_{**}. \tag{9}$$

We note, that in aerodynamics the functional $J(\cdot)$ might represent important quantities like aerodynamical force coefficients (drag, lift or moment coefficients).

Some error estimates like the L^2-estimate in the case of Poisson's equation and the error estimates with respect to target functionals $J(\cdot)$ require the use of duality arguments including the solutions to appropriately defined dual or adjoint problems. Therefore, we continue the above list as follows:

- **Adjoint consistency:** Given the primal problem (1) and a target functional

$$J(u) = \int_\Omega j_\Omega \, u \, \mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma \, u \, \mathrm{d}s, \tag{10}$$

with $j_\Omega \in L^2(\Omega)$ and $j_\Gamma \in L^2(\Gamma)$, we define the adjoint problem

$$L^* z = j_\Omega \quad \text{in } \Omega, \qquad\qquad B^* z = j_\Gamma \quad \text{on } \Gamma. \tag{11}$$

where L^* and B^* denote the adjoint operators to L and B, respectively. Then we say that discretization the (2) together with $J(\cdot)$ in (10) is *adjoint consistent* if the exact solution z to the adjoint problem (11) satisfies:

$$B_h(w, z) = J(w) \quad \forall w \in V. \tag{12}$$

Depending on the discretization being adjoint consistent or not the corresponding discretization errors measured in $J(\cdot)$ (or in L^2) are optimal or not. In fact, there are discontinuous Galerkin discretizations which are adjoint inconsistent, e.g. the non-symmetric interior penalty (NIPG) method for the discretization of Poisson's equation, and which show a reduced order of convergence as compared to adjoint consistent discretizations like the symmetric interior penalty (SIPG) method or the method of Bassi and Rebay (BR2). Whereas consistency can be considered as basic requirement of a discretization to be reasonable at all (without consistency the discrete solutions might even not converge to the exact solution) the adjoint consistency property represents an additional, and very desirable, quality of the discretization.

There are further topics of high interest in the numerical analysis of finite element methods, among them *a posteriori* error estimates and indicators for local h-refinement and/or p-refinement which will not be covered in this lecture (but in the VKI lecture on adaptivity planned for 2009):

- **A priori and a posteriori error estimates:** We distinguish between *a priori* error estimates and *a posteriori* error estimates.
  - *A priori* error estimates involve norms of the exact solution u. As u is unknown (otherwise we would not need to solve the problem numerically) an *a priori* error estimate gives no quantitative size of the error of the numerical solution. It gives, however, the order the error converges under mesh refinement, $h \to 0$; see e.g. estimates (8) and (9).
  - *A posteriori* error estimates do not include the exact solution u but only computable values which depend on e.g. the numerical solution u_h like in

$$J(u) - J(u_h) \approx E(u_h). \tag{13}$$

- **Indicators for local refinement:** In most cases global refinement of the computational mesh or global enrichment of the polynomial degree is a very inefficient way of improving the accuracy of a numerical solution. In practice, usually only local mesh refinement can be afforded. For deciding which elements to refine local error indicators η_κ are needed. Here, a variety of different indicators exist, many of which are purely heuristic, some are designed to reduce the error in specific global norms and some to reduce the error in specific target quantities $J(\cdot)$. The derivation of reliable refinement indicators is a non-trivial task.

1.4 Outline

In discontinuous Galerkin methods the discrete functions might be discontinuous between neighboring elements. There, continuity of the discrete functions are not enforced strongly like in continuous finite elements but only weakly by introducing flux, jump and/or penalization terms on the faces between neighboring elements. Due to these interior face terms the estimates like (5), (8) or (9) are more complicated to prove in the DG world than for continuous finite element (FE) methods. In order to understand the derivation of estimates for the DG discretizations it is in some cases illustrative to first recall the respective estimates and their derivations for the continuous Galerkin methods. Therefore, this lecture starts off with recalling well-known results from the numerical analysis of the continuous finite element methods. In particular, we recall *a priori* error estimates in the energy norm and the L^2-norm including their proofs for higher order standard finite element methods of Poisson's equation in Section 2 and for the standard and the streamline diffusion finite element method of the linear advection equation in Section 3.

We then introduce the discontinuous Galerkin discretization of the linear advection equation in Section 4. Following [13] we consider two numerical flux functions, the mean-value flux and the upwind flux, and derive the corresponding *a priori* error estimates. Whereas the standard Galerkin discretization of the linear advection equation is unstable and requires e.g. streamline diffusion for stabilization, we will see in Section 4 that the discontinuous Galerkin discretization of the linear advection based on upwind is stable without addition of streamline diffusion.

Then in Section 5, we follow [2] and derive and analyze a variety of discontinuous Galerkin discretizations of Poisson's equations. In particular, we derive the symmetric and non-symmetric interior penalty Galerkin method (SIPG and NIPG), the method of Baumann-Oden (BO) and the first and second method of Bassi and Rebay (BR1 and BR2). The analysis of the methods includes the consistency and adjoint consistency of the schemes, continuity and coercivity of the respective bilinear forms and *a priori* error estimates for the interior penalty methods. In particular, we will see that the adjoint consistent SIPG scheme is of optimal order in the L^2-norm whereas the adjoint inconsistent NIPG scheme is not.

Motivated by the connection of adjoint consistency of DG discretizations to the availability of optimal order error estimates in the L^2-norm we concentrate on the adjoint consistency property in Section 6. In particular, here we follow [27] and give a general framework for analyzing the consistency and adjoint consistency of DG discretizations for linear problems with inhomogeneous boundary conditions. This includes the derivation of continuous adjoint problems associated to specific target quantities, the derivation of primal and adjoint residual forms of the discretizations and the discussion whether the discretizations in combination with specific target quantities $J(\cdot)$ are adjoint consistent or not. This analysis is performed in Sections 6.3 and 6.4 for the interior penalty DG discretization of the Dirichlet-Neumann boundary value problem of Poisson's equations and for the upwind DG discretization of the linear advection equation, respectively.

Then in Section 7 the previously shown properties and estimates for the interior penalty and the upwind DG discretization are used to derive *a priori* estimates for the error measured in terms of target quantities $J(\cdot)$. Here again, we will see that a discretization must be consistent and adjoint consistent in order to provide optimal error estimates in $J(\cdot)$.

This lecture is finalized with the Sections 8 and 9 which introduce the DG discretizations of the compressible Euler and Navier-Stokes equations. Additionally, the consistency and adjoint consistency analysis which has been introduced in Section 6 for linear problems is now generalized to nonlinear problems in Section 8.5. This analysis is performed for the compressible Euler and Navier-Stokes equations in Sections 8.6 and 9.3, respectively. This includes the derivation of an adjoint consistent discretization of boundary conditions and of target functionals. Here particular emphasis is placed on the aerodynamic force coefficients like the drag, lift and moment coefficients.

Various examples in Sections 5.6, 7.3, 8.7 and 9.4 illustrate the numerical methods described.

2 Higher order continuous FE methods for Poisson's equation

In this section we consider the continuous finite element discretization of Poisson's equation. In particular, we recall some standard results including the H^1 and L^2 *a priori* error estimates for 2nd and higher order discretizations.

2.1 Poisson's equation

Let $\Omega \in \mathbb{R}^d$, $d \geq 1$, a bounded open domain. We consider the elliptic model problem,

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D, \qquad \mathbf{n} \cdot \nabla u = g_N \quad \text{on } \Gamma_N, \tag{14}$$

where $f \in L^2(\Omega)$, $g_D \in L^2(\Gamma_D)$ and $g_N \in L^2(\Gamma_N)$ are given functions. We assume that Γ_D and Γ_N are disjoint subsets with union $\Gamma = \partial\Omega$, that is $\Gamma_D \cup \Gamma_N = \Gamma$ and $\Gamma_D \cap \Gamma_N = \emptyset$.

2.1.1 The homogeneous Dirichlet problem

We first consider the case $\Gamma_D = \Gamma$ and $g_D = 0$, i.e. the homogeneous Dirichlet problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma_D. \tag{15}$$

We multiply (15) by a test function $v \in H_0^1(\Omega)$, integrate over Ω, and integrate by parts, to obtain

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} - \int_{\partial\Omega} \mathbf{n} \cdot \nabla u \, v \, \mathrm{d}s = \int_\Omega f v \, \mathrm{d}\mathbf{x}.$$

Due to $v \in H_0^1(\Omega)$ the boundary integral vanishes. Thereby, the weak form of (15) is given as follows: find $u \in V := H_0^1(\Omega)$ such that

$$B(u, v) = F(v) \qquad \forall v \in V, \tag{16}$$

where

$$B(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x}, \qquad F(v) = \int_\Omega f v \, \mathrm{d}\mathbf{x}. \tag{17}$$

Theorem 2.1 (Lax-Milgram theorem (Existence and Uniqueness)) *Let V be a Hilbert space, i.e. a complete space with scalar product $(\cdot, \cdot)$. Let the linear form $F : V \to \mathbb{R}$ be* continuous*, i.e. there is a $C_F > 0$ such that*

$$F(v) \leq C_F \|v\|_V \quad \forall v \in V. \tag{18}$$

Let the bilinear form $B : V \times V \to \mathbb{R}$ be continuous*, i.e. there is a constant $C_B > 0$ such that*

$$B(u, v) \leq C_B \|u\|_V \|v\|_V, \quad \forall u, v \in V. \tag{19}$$

Furthermore, let B be V-coercive (or V-elliptic), i.e. there is a constant $\gamma > 0$, such that

$$B(v, v) \geq \gamma \|v\|_V^2, \quad \forall v \in V. \tag{20}$$

Then, there is a unique solution $u \in V$ such that

$$B(u, v) = F(v) \quad \forall v \in V. \tag{21}$$

We say, Problem (21) *is well-posed.*

In the following we want to employ the Lax-Milgram Theorem 2.1 to show that there exists a unique solution to problem (16). Using the Cauchy-Schwarz inequality, we find that

$$|F(v)| = |\int_\Omega fv \,\mathrm{d}\mathbf{x}| \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)},$$
$$|B(u,v)| = |\int_\Omega \nabla u \nabla v \,\mathrm{d}\mathbf{x}| \leq \|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}.$$

Hence, the linear and bilinear functionals F and B are continuous. Furthermore, we see that B is coercive with respect to the H^1-seminorm:

$$B(v,v) = \int_\Omega |\nabla v|^2 \,\mathrm{d}\mathbf{x} = |v|^2_{H^1(\Omega)}. \tag{22}$$

Using the Poincaré inequality 24 we obtain

$$B(v,v) = \int_\Omega |\nabla v|^2 \,\mathrm{d}\mathbf{x} = |v|^2_{H^1(\Omega)} \geq \gamma \left(\|u\|^2_{L^2(\Omega)} + |v|^2_{H^1(\Omega)}\right) = \gamma \|v\|^2_{H^1(\Omega)}, \tag{23}$$

with $\gamma = 1/(C_p + 1)$, i.e. $B(\cdot,\cdot)$ is H^1-coercive. Existence and uniqueness of a solution to (16) now follows by application of the Lax-Milgram theorem 2.1.

Theorem 2.2 (Poincaré inequality) *Let $\Omega \in \mathbb{R}^d$ be contained in a d-dimensional (hyper)cube with edge length s. Then, there is a constant $C_p = s > 0$ such that*

$$\|v\|_{L^2(\Omega)} \leq C_p |v|_{H^1(\Omega)} \quad \forall v \in H^1_0(\Omega). \tag{24}$$

Proof: Any book on linear functional analysis or finite element methods. □

2.1.2 The inhomogeneous Dirichlet problem

Let us now consider the case of an *in*homogeneous Dirichlet problem,

$$-\Delta u = f \quad \text{in } \Omega, \qquad\qquad u = g_D \quad \text{on } \Gamma_D, \tag{25}$$

with $g_D \in L^2(\Gamma_D)$ on $\Gamma_D = \Gamma$, and $g_D \not\equiv 0$. Assume that there is a $u_D \in H^1(\Omega)$ with $u_D = g_D$ on Γ_D. Then, the weak formulation to (25) is given by: find $u = u_D + u_0$ with $u_0 \in H^1_0(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}\mathbf{x} = \int_\Omega fv \,\mathrm{d}\mathbf{x} \qquad \forall v \in H^1_0(\Omega). \tag{26}$$

We deduce the existence and uniqueness of a solution to this problem similar to the homogeneous case by rewriting the problem as follows: find $u_0 \in H^1_0(\Omega)$ such that

$$\int_\Omega \nabla u_0 \cdot \nabla v \,\mathrm{d}\mathbf{x} = \int_\Omega fv \,\mathrm{d}\mathbf{x} - \int_\Omega \nabla u_D \cdot \nabla v \,\mathrm{d}\mathbf{x} \qquad \forall v \in H^1_0(\Omega).$$

For $B(u_0, v)$ and $F(v)$ denoting the left and right hand side of this equation, we find that

$$F(v) \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + |u_D|_{H^1(\Omega)}|v|_{H^1(\Omega)} \leq C\|v\|_{H^1(\Omega)}.$$

Furthermore, it has been shown in the previous section that $B(\cdot,\cdot)$ is a continuous and H^1-coercive bilinear functional. Well-posedness, i.e. existence and uniqueness of a solution to (26) then follows through the application of the Lax-Milgram theorem 2.1.

2.1.3 The Neumann problem

Finally, we consider the Neumann problem,

$$-\Delta u = f \quad \text{in } \Omega, \qquad\qquad \mathbf{n}\cdot\nabla u = g_N \quad \text{on } \Gamma_N, \tag{27}$$

with $\Gamma_N = \Gamma$ and $g_N \in L^2(\Gamma_N)$ is a given function. As we do not have Dirichlet boundary conditions to impose we consider the space $H^1(\Omega)$ instead of $H^1_0(\Omega)$. We multiply (27) by a test function $v \in H^1(\Omega)$, integrate over Ω, and integrate by parts, to obtain

$$\int_\Omega \nabla u\cdot\nabla v\,\mathrm{d}\mathbf{x} - \int_{\partial\Omega} \mathbf{n}\cdot\nabla u\, v\,\mathrm{d}s = \int_\Omega f v\,\mathrm{d}\mathbf{x}.$$

In view of (27) we obtain following weak formulation: find $u \in V := H^1(\Omega)$ such that

$$B(u,v) = F(v), \quad \forall v \in V,$$

with

$$B(u,v) \equiv \int_\Omega \nabla u\cdot\nabla v\,\mathrm{d}\mathbf{x}, \qquad\qquad F(v) = \int_\Omega f v\,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N\, v\,\mathrm{d}s. \tag{28}$$

Remark 2.3 (Well-posedness of the Neumann problem) *We observe that*

$$B(1,v) = B(u,1) = 0 \tag{29}$$

for all $u, v \in H^1(\Omega)$, i.e. $\ker B = span\{1\}$. As $B(1,1) = 0$, the Lax-Milgram theorem can not be applied. However, considering $H^1(\Omega)/\ker B$ which is the set of all equivalent classes

$$\tilde{H}^1(\Omega) := H^1(\Omega)/\ker B \cong \{u \in H^1(\Omega) : \int_\Omega u\,d\mathbf{x} = 0\}, \tag{30}$$

we employ a Poincaré inequality for $v \in \tilde{H}^1(\Omega)$, [40], and arrive at

$$|B(v,v)| = \int_\Omega |\nabla v|^2\,d\mathbf{x} = |v|^2_{H^1(\Omega)} \geq \gamma\|v\|^2_{H^1(\Omega)} \quad v \in \tilde{H}^1(\Omega). \tag{31}$$

Using a trace inequality $\|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)}$ for all $v \in H^1(\Omega)$, see Theorem 2.5, we obtain

$$|F(v)| \leq |\int_\Omega f v\,d\mathbf{x}| + |\int_{\partial\Omega} g_N\, v\,ds| \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g_N\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)}.$$

Using a generalization to the Lax-Milgram theorem, we obtain existence and uniqueness in $\tilde{H}^1(\Omega)$ (i.e. uniqueness up to constants) of a solution, provided following compatibility condition is satisfied

$$F(1) = \int_\Omega f\,d\mathbf{x} + \int_{\partial\Omega} g_N\,ds = 0. \tag{32}$$

Remark 2.4 *In order to avoid the theoretical difficulties of the pure Neumann problem, we usually consider the mixed Dirichlet-Neumann problem (14), with $\Gamma_D \neq \emptyset$. In fact, imposing Dirichlet boundary conditions at a single point only, i.e. $\Gamma_D = \{p\} \subset \Gamma$, is sufficient to obtain a solution which is unique, and not only unique up to a constant.*

Theorem 2.5 (Trace theorem) *Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain with piecewise smooth boundary. Furthermore, Ω satisfies a cone condition. Then there is a unique continuous linear map*

$$\gamma : H^1(\Omega) \to L^2(\partial\Omega),$$

and a constant $C > 0$ such that

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)} \quad \forall u \in H^1(\Omega), \tag{33}$$

and

$$\gamma(u) = u|_{\partial\Omega} \quad \forall u \in H^1(\Omega) \cap C^0(\bar{\Omega}),$$

where $\gamma(u)$ denotes the trace *of u on $\partial\Omega$ and γ the* trace operator*. Note, that usually the trace operator is omitted and a notation $u|_{\partial\Omega}$ is used instead of $\gamma(u)$.*

2.2 The standard finite element method for Poisson's equation

In this section we introduce the standard (continuous) finite element method for Poisson's equation,

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D, \qquad \mathbf{n} \cdot \nabla u = g_N \quad \text{on } \Gamma_N, \tag{34}$$

where $f \in L^2(\Omega)$, $g_D \in L^2(\Gamma_D)$ and $g_N \in L^2(\Gamma_N)$ are given functions. We assume that Γ_D and Γ_N are disjoint subsets with union Γ, that is $\Gamma_D \cup \Gamma_N = \Gamma$ and $\Gamma_D \cap \Gamma_N = \emptyset$. Furthermore, we assume that $\Gamma_D \neq \emptyset$, see Remark 2.4. As described in the previous section, this problem is rewritten in a weak formulation: find $u \in V$ such that

$$B(u, v) = F(v) \quad \forall v \in V, \tag{35}$$

where V is an appropriately chosen function space with $H_0^1(\Omega) \subset V \subset H^1(\Omega)$.

The finite element method generates approximate solutions to (35). To this end, let $\mathcal{T}_h = \{\kappa\}$ be a geometric discretization of Ω consisting of elements κ, where h denotes the maximum diameter of all κ. Let $\phi_j(\mathbf{x}) \in V, 0 \leq j < N_h$, be N_h linearly independent functions in V and u_j, $0 \leq j < N_h$, real numbers. Then

$$u_h(\mathbf{x}) = \sum_{0 \leq j < N_h} u_j \phi_j(\mathbf{x}), \tag{36}$$

is a discrete function in the *discrete* function space V_h defined by

$$V_h = \text{span}\{\phi_j(\mathbf{x})\}_{j=1}^{N_h} \subset V.$$

We note, that $V_h \subset V$, where V is the *continuous* (and infinite-dimensional) function space the exact solution u to (35) is to be sought in.

Definition 2.6 *For $p \geq 1$ we define the space of continuous piecewise polynomials of degree p by*

$$\begin{aligned} V_{h,p}^c = \{v_h \in C^0(\Omega) : & v_h|_\kappa \circ \sigma_\kappa \in Q_p(\hat{\kappa}) \text{ if } \hat{\kappa} \text{ is the unit hypercube, and} \\ & v_h|_\kappa \circ \sigma_\kappa \in P_p(\hat{\kappa}) \text{ if } \hat{\kappa} \text{ is the unit simplex}, \kappa \in \mathcal{T}_h\}, \end{aligned} \tag{37}$$

where P_p and Q_p are the spaces of polynomials and tensor product polynomials of degree p. While dealing with continuous finite element discretizations, we use the short notation $V_h := V_{h,p}^c$.

Replacing u and v in (35) by discrete functions $u_h, v_h \in V_h$, the discrete problem is given by: find $u_h \in V_h$ such that

$$B(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h. \tag{38}$$

We note that the bilinear and linear forms in the discrete problem (38) are the same as in the weak formulation (35). Therefore, in this section we do not need to introduce notations $B_h(\cdot,\cdot)$ and $F_h(\cdot)$ as we did in the general case in Section 1.3.

Due to the trial (ansatz) and test functions u_h and v_h taken from the same discrete function space V_h this is a so-called *Galerkin* finite element discretization. There are also so-called *Petrov-Galerkin* finite element discretizations where the trial and test functions, $u_h \in U_h$ and $v_h \in V_h$, belong to different discrete function spaces $U_h \neq V_h$.

Definition 2.7 *Finite elements based on discrete functions $u_h \in V_h$ being contained in the continuous function space V, i.e. $V_h \subset V$, are called* conforming *finite elements.*

Finite elements methods where the discrete functions $u_h \in V_h$ do not belong to V, i.e. $V_h \not\subset V$, are called non-conforming *finite elements.*

Note, that continuous Galerkin finite elements, as discussed in this section, are conforming finite elements. In contrast to that, the discontinuous Galerkin finite elements being discussed in Section 4 onwards are non-conforming as the discontinuous discrete functions spaces are generally not subspaces of the classic continuous function spaces V.

2.2.1 Consistency

From the weak formulation (35) we see that the discretization (38) is consistent, i.e. the exact (weak) solution $u \in V$ to (34) satisfies

$$B(u,v) = F(v) \quad \forall v \in V. \tag{39}$$

From consistency we immediately deduce following important property of Galerkin finite element methods, the so-called *Galerkin-orthogonality*:

$$B(u-u_h, v_h) = 0 \quad \forall v_h \in V_h, \tag{40}$$

which we obtain by subtracting (38) from (39) for $v_h \in V_h \subset V$ and using linearity of B_h with respect to its first argument. This means that the error $e = u - u_h$ is orthogonal (with respect to the bilinear form B_h) to the discrete function space V_h.

2.2.2 Existence and uniqueness of discrete solutions

In Section 2.1 we have shown that the linear and bilinear functionals $F : V \to \mathbb{R}$ and $B : V \times V \to \mathbb{R}$ in the weak problem: find u such that

$$B(u,v) = F(v) \quad \forall v \in V \tag{41}$$

are continuous. Furthermore, we have shown that B is V-coercive. We then applied the Lax-Milgram theorem to show that a solution u to (41) exists and that this solution is unique. As discussed above, standard finite elements are conforming finite elements, i.e. $V_h \subset V$. Therefore, F and B are continuous also on V_h. Furthermore, B is coercive also on V_h. Again by using the Lax-Milgram theorem we deduce that the discrete problem: find $u_h \in V_h$ such that

$$B(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h, \tag{42}$$

has a unique solution, i.e. (42) is well-posed.

Remark 2.8 *We see that, for all weak problems* (41) *where we are able to show well-posedness by using the Lax-Milgram theorem we immediately obtain well-posedness of any discrete problem* (42) *provided we have $V_h \subset V$, i.e. provided we use conforming finite elements. In contrast to that, discontinuous Galerkin methods are non-conforming, $V_h \not\subset V$. Therefore, well-posedness of discrete problems originating from discontinuous Galerkin discretizations is not immediate, but must be shown for each discrete function space V_h under consideration.*

2.2.3 Best approximation property

Lemma 2.9 (Céa Lemma) *Let the bilinear form B be continuous and V-coercive, with $H_0^1(\Omega) \subset V \subset H^1(\Omega)$. Furthermore, let $u \in V$ and $u_h \in V_h \subset V$ be the solutions to*

$$B(u, v) = F(v), \quad \forall v \in V,$$

and

$$B(u_h, v_h) = F(v_h), \quad \forall v_h \in V_h,$$

respectively. Then,

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C_B}{\gamma} \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}. \tag{43}$$

Proof: Let $v_h \in V_h$. First we use coercivity of B, then we use the Galerkin orthogonality, $B(u - u_h, w_h) = 0$ for $w_h = v_h - u_h \in V_h$, and finally we use continuity of B to obtain

$$\begin{aligned}
\gamma \|u - u_h\|_{H^1(\Omega)}^2 &\leq B(u - u_h, u - u_h) \\
&= B(u - u_h, u - v_h) + B(u - u_h, v_h - u_h) \\
&= B(u - u_h, u - v_h) \\
&\leq C_B \|u - u_h\|_{H^1(\Omega)} \|u - v_h\|_{H^1(\Omega)}
\end{aligned}$$

Dividing by $\gamma \|u - u_h\|_{H^1(\Omega)}$ we obtain (43). □

In (43) we see that apart from a constant the approximation error $e = u - u_h$ is bounded by the difference $v - v_h$ for any discrete function $v_h \in V_h$. This is the so-called best approximation property. As we are free to choose any v_h we can for example take an interpolation $v_h = I_h u \in V_h$ of u to obtain

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C_B}{\gamma} \|u - I_h u\|_{H^1(\Omega)}. \tag{44}$$

Hence, the discretization error $e = u - u_h$ can be bounded by the interpolation error $u - I_h u$. In particular, the order of the discretization is the same as the order of an interpolation into V_h.

2.2.4 Interpolation estimates

In this section we recall some standard interpolation estimates.

Definition 2.10 (Interpolation operator $I_{h,p}^c$ onto $V_{h,p}^c$) *For $p \geq 1$ let ϕ_i, $0 \leq i < N_h$, be a nodal basis of $V_{h,p}^c$ with*

$$\phi_i(\mathbf{x}^{(j)}) = \delta_{ij} \quad \forall\ 0 \leq i, j < N_h,$$

where $\mathbf{x}^{(j)}$, $0 \leq j < N_h$, are the nodal points of $V_{h,p}^c$ and $N_h := \#(V_{h,p}^c)$ is the dimension of the discrete space $V_{h,p}^c$, i.e. the number of degrees of freedom in $V_{h,p}^c$. Given a function $u \in H^2(\Omega)$, we define $I_{h,p}^c u \in V_{h,p}^c$ to be the interpolation of u onto $V_{h,p}^c$ given by

$$I_{h,p}^c u(\mathbf{x}) = \sum_{0 \leq i < N_h} u(\mathbf{x}_i) \phi_i(\mathbf{x}). \tag{45}$$

Possible nodal basis functions are the Lagrange interpolation polynomials $\phi_i(\mathbf{x}) := L_i^{(p)}(\mathbf{x})$. In the following we use the short notation $I_h u$ instead of $I_{h,p}^c u$ when it is clear that an interpolation into $V_{h,p}^c$ is meant.

Theorem 2.11 (Interpolation estimate) *Let $\mathcal{T}_h$ be a shape regular mesh of $\Omega \subset \mathbb{R}^d$, $1 \leq d \leq 3$. Let $p \geq 1$ and I_h be an interpolation operator onto $V_{h,p}^c$. Furthermore, let $0 \leq m \leq p+1$. Then there is a constant C, independent of h, such that for all $u \in H^{p+1}(\Omega)$ we have*

$$\|u - I_h u\|_{H^m(\Omega)} \leq C h^{p+1-m} |u|_{H^{p+1}(\Omega)}. \tag{46}$$

Proof: See e.g. [40]. □

Example 2.12 *In particular, for $u \in H^{p+1}(\Omega)$ and $m = 0, 1$, the estimate (46) reduces to*

$$\|u - I_h u\|_{L^2(\Omega)} \leq C h^{p+1} |u|_{H^{p+1}(\Omega)}, \tag{47}$$

and to

$$\|u - I_h u\|_{H^1(\Omega)} \leq C h^{p} |u|_{H^{p+1}(\Omega)}. \tag{48}$$

I.e. the interpolation error is of $\mathcal{O}(h^{p+1})$ in the $L^2(\Omega)$-norm and of $\mathcal{O}(h^p)$ in the $H^1(\Omega)$-norm.

We note that Theorem 2.11 requires the solution u to be in $H^{p+1}(\Omega)$. However, for the case that $u \in H^{s+1}(\Omega)$ with $s < p$ we obtain only a reduced interpolation order.

Corollary 2.13 (Interpolation estimate) *Let $\mathcal{T}_h$ be a shape regular mesh of $\Omega \subset \mathbb{R}^d$, $1 \leq d \leq 3$. Let $p \geq 1$ and I_h be the interpolation operator onto $V_{h,p}^c$ as defined in Definition 2.10. Then there is a constant C, independent of h, such that for all $u \in H^{s+1}(\Omega)$ we have*

$$\|u - I_h u\|_{H^m(\Omega)} \leq C h^{t+1-m} |u|_{H^{t+1}(\Omega)}. \tag{49}$$

where $t = \min\{s, p\}$ and $0 \leq m \leq t+1$.

We note that for sufficiently smooth functions $u \in H^{p+1}(\Omega)$, i.e. $s \geq p$, this estimate reduces to (46), i.e. we have

$$\|u - I_h u\|_{H^m(\Omega)} \leq C h^{p+1-m} |u|_{H^{p+1}(\Omega)}, \tag{50}$$

while for functions with a lower smoothness, i.e. $u \in H^{s+1}(\Omega)$ with $s < p$, we have

$$\|u - I_h u\|_{H^m(\Omega)} \leq C h^{s+1-m} |u|_{H^{s+1}(\Omega)}, \tag{51}$$

for $m \leq s+1$. The interpolation estimates given above can be combined with a trace theorem for obtaining interpolation estimates in the $L^2(\partial\Omega)$-norm:

Theorem 2.14 (Interpolation estimate in the $L^2(\partial\Omega)$-norm) *Let $\mathcal{T}_h$ be a shape regular mesh of $\Omega \subset \mathbb{R}^d$, $1 \leq d \leq 3$. Let $p \geq 1$ and I_h be the interpolation operator onto $V_{h,p}^c$ as defined in Definition 2.10. Then there is a constant C, independent of h, such that for all $u \in H^{p+1}(\Omega)$ we have*

$$\|u - I_h u\|_{L^2(\partial\Omega)} \leq C h^{p+1/2} |u|_{H^{p+1}(\Omega)}. \tag{52}$$

2.2.5 *A priori* error estimates in the H^1- and L^2-norm

In this section we derive error estimates in the $H^1(\Omega)$- and the $L^2(\Omega)$-norm for the discretization error of the standard finite element discretization of Poisson's equation.

Corollary 2.15 (H^1-error estimate) *Let $u \in H^{p+1}(\Omega)$ and $u_h \in V_h = V_{h,p}^c$ be the solutions to (35) and (38), respectively. Then*

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{p} |u|_{H^{p+1}(\Omega)}. \tag{53}$$

Proof: Using (44) and (46) for $m = 1$ we obtain

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C_B}{\gamma}\|u - I_h u\|_{H^1(\Omega)} \leq Ch^p |u|_{H^{p+1}(\Omega)}.$$

□

Example 2.16 *For $u \in H^2(\Omega)$ and $u_h \in V_{h,1}^c$ we obtain the standard result:*

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch|u|_{H^2(\Omega)}. \tag{54}$$

In the following we derive an error estimate in the L^2-norm by using a duality argument.

Theorem 2.17 (L^2-error estimate (Aubin-Nitsche)) *Let $u \in H^{p+1}(\Omega)$ and $u_h \in V_h = V_{h,p}^c$ be the solutions to* (35) *and* (38)*, respectively. Then*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{p+1}|u|_{H^{p+1}(\Omega)}. \tag{55}$$

Proof: For simplicity, we assume homogeneous Dirichlet boundary conditions. For $v \in L^2(\Omega)$ let z be the solution to following dual (or adjoint) problem: find $z \in H_0^1(\Omega)$ such that

$$B(w, z) = \int_\Omega wv \,\mathrm{d}\mathbf{x} \quad \forall w \in H_0^1(\Omega).$$

We assume that $z \in H_0^1(\Omega) \cap H^2(\Omega)$ and $\|z\|_{H^2} \leq C\|v\|_{L^2}$ which is satisfied if Ω is a convex polygon, for example. Now choosing $v = w = e = u - u_h \in H_0^1(\Omega)$ yields

$$\|e\|_{L^2(\Omega)}^2 = \int_\Omega e^2 \,\mathrm{d}\mathbf{x} = B(e, z) = B(e, z - z_h) \leq C\|u - u_h\|_{H^1(\Omega)}\|z - z_h\|_{H^1(\Omega)}, \tag{56}$$

where we used Galerkin-orthogonality $B(e, z_h) = 0$ for any $z_h \in V_h$. Choosing $z_h = I_h z \in V_h$ and using the interpolation estimate (51) for $s = m = 1$ we obtain

$$\|e\|_{L^2(\Omega)}^2 \leq C\|u - u_h\|_{H^1(\Omega)} h\|z\|_{H^2(\Omega)} \leq Ch\|u - u_h\|_{H^1(\Omega)}\|e\|_{L^2(\Omega)}. \tag{57}$$

Division by $\|e\|_{L^2(\Omega)}$, and using the H^1-error estimate (53) we get (55). □

Example 2.18 *For $u \in H^2(\Omega)$ and $u_h \in V_{h,1}^c$ we obtain the standard result:*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2|u|_{H^2(\Omega)}. \tag{58}$$

In summary, we have seen that the standard (continuous) finite element discretization of Poisson's equation based on piecewise polynomials of degree p is of order $p + 1$ in the $L^2(\Omega)$-norm provided the exact solution u is sufficiently smooth.

3 Higher order continuous FE methods for the linear advection equation

In this section we consider standard (continuous) finite element discretizations for the linear advection equation. In particular, we recall some standard results including *a priori* error estimates in the L^2-norm for the standard Galerkin method and in the $H^{1,\mathbf{b}}$-norm for the streamline diffusion finite element method (SDFEM).

3.1 The linear advection equation

For $\Omega \in \mathbb{R}^d$, $d \geq 1$, we consider the linear advection equation

$$Lu := \nabla \cdot (\mathbf{b}u) + cu = f \quad \text{in } \Omega, \qquad\qquad u = g \quad \text{on } \Gamma_-, \tag{59}$$

where $f \in L^2(\Omega)$, $\mathbf{b} \in [C^1(\Omega)]^d$, $c \in L^\infty(\Omega)$ and $g \in L^2(\Gamma_-)$, where

$$\Gamma_- = \{\mathbf{x} \in \Gamma, \mathbf{b}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\} \tag{60}$$

denotes the inflow part of the boundary $\Gamma = \partial\Omega$. Furthermore, we adopt following hypothesis: there exists a $c_0 \in L^\infty(\Omega)$ and a number $\gamma_0 > 0$ such that

$$c(\mathbf{x}) + \frac{1}{2}\nabla \cdot \mathbf{b}(\mathbf{x}) = c_0^2(\mathbf{x}) \geq \gamma_0 > 0. \tag{61}$$

This condition is required for ensuring stability below.

Remark 3.1 *In order to demonstrate the similarities with the compressible Euler equations, see Section 8, we consider the linear advection equation* (59) *in conservative form. We note, that problem* (59) *is equivalent to the linear advection equation in non-conservative form*

$$\mathbf{b} \cdot \nabla u + \tilde{c}u = f \tag{62}$$

with the hypothesis $\tilde{c} - \frac{1}{2}\nabla \cdot \mathbf{b} = c_0^2$ *and* $\tilde{c} = c + \nabla \cdot \mathbf{b}$.

In the following we derive the variational formulation of the linear advection equation and define the proper function space the solution is to be sought in. To this end, we multiply (59) by a test function $v \in L^2(\Omega)$ and integrate over the domain Ω,

$$\int_\Omega (\nabla \cdot (\mathbf{b}u) + cu)\, v \,\mathrm{d}\mathbf{x} = \int_\Omega f v \,\mathrm{d}\mathbf{x} \quad \forall v \in L^2(\Omega), \tag{63}$$

where the function space the solution is to be searched in will be defined in the following.

First we note, that for the integral on the left hand side to exist we require $\nabla \cdot (\mathbf{b}u) + cu \in L^2(\Omega)$, i.e. we consider the function space

$$H^{1,\mathbf{b}}(\Omega) = \{u \in L^2(\Omega) : Lu = \nabla \cdot (\mathbf{b}u) + cu \in L^2(\Omega)\}. \tag{64}$$

Then, it remains to incorporate boundary conditions in (63). There are two ways of doing so, first by a so-called *strong* and second by a so-called *weak* imposition of boundary conditions, which both are discussed in the following two subsections.

3.1.1 Variational formulation with strong boundary conditions

We recall that for the Dirichlet problem of Poisson's equation we have used $H_0^1(\Omega)$ instead of $H^1(\Omega)$ for realizing Dirichlet boundary conditions on $\Gamma = \partial\Omega$. For the linear advection equation (59) imposition of boundary conditions is allowed only on the inflow boundary part Γ_- of the boundary. Consequently, the function space to search the solution in is

$$H_-^{1,\mathbf{b}}(\Omega) = \{u \in L^2(\Omega) : Lu \in L^2(\Omega), \mathbf{b}\cdot\mathbf{n}\, u = 0 \text{ on } \Gamma_-\} \subset H^{1,\mathbf{b}}(\Omega). \tag{65}$$

Then the variational formulation of the linear advection with homogeneous inflow boundary conditions, i.e. (59) with $g \equiv 0$ on Γ_-, is given by: find $u \in H_-^{1,\mathbf{b}}(\Omega)$ such that

$$\int_\Omega (\nabla\cdot(\mathbf{b}u) + cu)\, v \,\mathrm{d}\mathbf{x} = \int_\Omega fv \,\mathrm{d}\mathbf{x} \quad \forall v \in L^2(\Omega). \tag{66}$$

The realization of inhomogeneous inflow boundary conditions is similar to the imposition of inhomogeneous Dirichlet boundary conditions described for Poisson's equation in Section 2.1.2. Assume that there is a $u_g \in H^{1,\mathbf{b}}(\Omega)$ with $u_g = g$ on Γ_-. Then the variational formulation of (59) is given by: find $u = u_- + u_g$ with $u_- \in H_-^{1,\mathbf{b}}(\Omega)$ such that

$$\int_\Omega (\nabla\cdot(\mathbf{b}u) + cu)\, v \,\mathrm{d}\mathbf{x} = \int_\Omega fv \,\mathrm{d}\mathbf{x} \quad \forall v \in L^2(\Omega). \tag{67}$$

In summary, strong boundary conditions are realized by considering an appropriate function (sub)space $H_-^{1,\mathbf{b}}(\Omega) \subset H^{1,\mathbf{b}}(\Omega)$ which incorporates boundary conditions on Γ_-.

3.1.2 Variational formulation with weak boundary conditions

In the following we derive a variational formulation which imposes boundary conditions in a weak sense. To this end, we multiply (59) by a test function $v \in H^{1,\mathbf{b}}(\Omega)$, integrate over the domain Ω, integrate by parts and replace u by g on Γ_- which gives

$$-\int_\Omega (\mathbf{b}u)\cdot\nabla v \,\mathrm{d}\mathbf{x} + \int_\Omega cuv \,\mathrm{d}\mathbf{x} + \int_{\Gamma_+} \mathbf{b}\cdot\mathbf{n}\, uv \,\mathrm{d}s = \int_\Omega fv \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, gv \,\mathrm{d}s,$$

where $\Gamma_+ = \Gamma \setminus \Gamma_-$ is the outflow part of the boundary. Integrating back by parts we obtain following variational formulation: find $u \in H^{1,\mathbf{b}}(\Omega)$ such that

$$\int_\Omega (\nabla\cdot(\mathbf{b}u) + cu)\, v \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, uv \,\mathrm{d}s = \int_\Omega fv \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, gv \,\mathrm{d}s \quad \forall v \in H^{1,\mathbf{b}}(\Omega). \tag{68}$$

Note, that here the boundary condition $u = g$ on the inflow boundary Γ_- is weakly imposed. Furthermore, here u is sought in the (full) function space $H^{1,\mathbf{b}}(\Omega)$ which is in contrast to $u \in H_-^{1,\mathbf{b}}(\Omega)$ in the case of strongly imposed boundary conditions in Section 3.1.1. The transport equation (59) has a unique weak solution $u \in H^{1,\mathbf{b}}(\Omega)$ given by (68) and the boundary condition is satisfied as an equality in $L^2(\Gamma_-)$, see [34].

3.2 The standard Galerkin method with weak boundary conditions

Starting from the variational formulation (68) the standard Galerkin method with weak boundary conditions is given as follows: find $u_h \in V_h := V_{h,p}^c$ such that

$$B(u_h, v_h) = F(v_h) \quad \forall v \in V_h, \tag{69}$$

where

$$\begin{aligned} B(u,v) &= \int_\Omega \left(\nabla\cdot(\mathbf{b}u) + cu\right) v \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, uv \,\mathrm{d}s, \\ F(v) &= \int_\Omega fv \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, gv \,\mathrm{d}s. \end{aligned} \tag{70}$$

From (68) we see that the discretization is consistent, i.e.

$$B(u,v) = F(v), \quad \forall v \in V, \tag{71}$$

holds for the exact solution u. By subtracting (69) from (71) for $v_h \in V_h \subset V$ we obtain the Galerkin orthogonality

$$B(u-u_h, v_h) = 0, \quad \forall v \in V_h. \tag{72}$$

Lemma 3.2 *For any $v \in H^{1,\mathbf{b}}(\Omega)$ we have*

$$B(v,v) \geq \gamma_0 \|v\|^2_{L^2(\Omega)} + \frac{1}{2}\int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \, ds. \tag{73}$$

Proof: From (70) we have

$$B(v,v) = \int_\Omega \nabla\cdot(\mathbf{b}v)\, v + cv^2 \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, v^2 \,\mathrm{d}s.$$

Noting that

$$\begin{aligned} \nabla\cdot(\mathbf{b}v)\, v &= (\nabla\cdot\mathbf{b})\, v^2 + (\mathbf{b}\cdot\nabla v)\, v = (\nabla\cdot\mathbf{b})\, v^2 + \frac{1}{2}\mathbf{b}\cdot\nabla v^2 \\ &= (\nabla\cdot\mathbf{b})\, v^2 + \frac{1}{2}\nabla\cdot\left(\mathbf{b}v^2\right) - \frac{1}{2}(\nabla\cdot\mathbf{b})\, v^2 \\ &= \frac{1}{2}(\nabla\cdot\mathbf{b})\, v^2 + \frac{1}{2}\nabla\cdot\left(\mathbf{b}v^2\right), \end{aligned} \tag{74}$$

we obtain by using the divergence theorem and hypothesis (61)

$$\begin{aligned} B(v,v) &= \int_\Omega \left(\frac{1}{2}\nabla\cdot\mathbf{b} + c\right) v^2 \,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Omega \nabla\cdot\left(\mathbf{b}v^2\right) \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, v^2 \,\mathrm{d}s \\ &= \int_\Omega c_0^2 v^2 \,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Gamma \mathbf{b}\cdot\mathbf{n} v^2 \,\mathrm{d}s - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, v^2 \,\mathrm{d}s \\ &\geq \gamma_0 \int_\Omega v^2 \,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_{\Gamma_+} \mathbf{b}\cdot\mathbf{n} v^2 \,\mathrm{d}s - \frac{1}{2}\int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, v^2 \,\mathrm{d}s \\ &= \gamma_0 \int_\Omega v^2 \,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s, \end{aligned} \tag{75}$$

as $\mathbf{b}\cdot\mathbf{n} < 0$ on Γ_- and $\mathbf{b}\cdot\mathbf{n} \geq 0$ on Γ_+. □

Hence $B(\cdot,\cdot)$ is coercive with respect to

$$\|v\|^2_{L^2(\Omega)} + \int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s. \tag{76}$$

However, this is *not* a norm on $H^{1,\mathbf{b}}(\Omega)$. In fact, there are functions $v \in L^2(\Omega)$ with $\|v\|^2_{L^2(\Omega)} + \int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s < \infty$ but with unbounded directional derivatives, i.e. $\|\mathbf{b}\cdot\nabla v\| = \infty$, hence $v \notin H^{1,\mathbf{b}}(\Omega)$.

We note, that the lack of $H^{1,\mathbf{b}}$-coercivity of the standard Galerkin method for the linear advection equation has direct consequences on the stability of the method. In fact, it is well known that the standard Galerkin method for the linear advection is *unstable*. This instability may lead to highly oscillating numerical solutions.

Theorem 3.3 *For $p \geq 1$ let $u \in H^{p+1}(\Omega)$ be the solution to (59) and $u_h \in V_{h,p}^c$ the solution to (69). Then,*

$$\|u - u_h\|_{L^2(\Omega)} + \left(\int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, |u - u_h|^2 \, ds \right)^{1/2} \leq C h^p |u|_{H^{p+1}(\Omega)}. \tag{77}$$

Proof: Let $e = u - u_h = \eta - \xi$ with $\eta = u - I_h u$ and $\xi = u_h - I_h u$ where $I_h : H^2(\Omega) \to V_{h,p}^c$ is the interpolation operator as defined in Definition 2.10. Then by using triangle inequality

$$\|e\|^2 + \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, |e|^2 \, \mathrm{d}s \leq \|\eta\|^2 + \|\xi\|^2 + \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, |\eta|^2 \, \mathrm{d}s + \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, |\xi|^2 \, \mathrm{d}s, \tag{78}$$

where we write $\|\cdot\| = \|\cdot\|_{L^2(\Omega)}$ for short. For the first term and third term on the right hand side we can use interpolation estimates given in Section 2.2.4. Furthermore, by using (73) we can bound the second and fourth term in (78) as follows

$$\gamma_0 \|\xi\|^2 + \tfrac{1}{2} \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, \xi^2 \, \mathrm{d}s \leq B(\xi, \xi) = B(\eta - e, \xi) = B(\eta, \xi) - B(e, \xi) = B(\eta, \xi), \tag{79}$$

where we used the Galerkin orthogonality property $B(e, \xi) = 0$ for $\xi \in V_{h,p}^c$, see (72). Using the Definition (70) of $B(\cdot, \cdot)$, and the Young's inequality, $ab \leq \frac{\epsilon}{4} a^2 + \frac{1}{\epsilon} b^2$, we obtain

$$\begin{aligned}
\gamma_0 \|\xi\|^2 + \tfrac{1}{2} \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, \xi^2 \, \mathrm{d}s \leq B(\eta, \xi) &= \int_\Omega (\nabla \cdot (\mathbf{b}\eta) + c\eta) \, \xi \, \mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n} \, \eta \xi \, \mathrm{d}s \\
&\leq \tfrac{\gamma_0}{4} \|\xi\|^2 + \tfrac{1}{\gamma_0} \|\mathbf{b} \cdot \nabla \eta\|^2 + \tfrac{\gamma_0}{4} \|\xi\|^2 + \tfrac{\tilde{c}^2}{\gamma_0} \|\eta\|^2 + \tfrac{1}{4} \int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}| \, \xi^2 \, \mathrm{d}s + \int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}| \, \eta^2 \, \mathrm{d}s,
\end{aligned}$$

where $\tilde{c} = \|\nabla \cdot \mathbf{b} + c\|_{L^\infty(\Omega)}$. By subtracting all ξ terms of the right hand side we obtain

$$\tfrac{\gamma_0}{2} \|\xi\|^2 + \tfrac{1}{4} \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, \xi^2 \, \mathrm{d}s \leq C \left(\|\mathbf{b} \cdot \nabla \eta\|^2 + \|\eta\|^2 + \int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}| \, \eta^2 \, \mathrm{d}s \right).$$

Hence, together with (78) we have

$$\|e\|^2 + \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, |e|^2 \, \mathrm{d}s \leq C \left(\|\mathbf{b} \cdot \nabla \eta\|^2 + \|\eta\|^2 + \int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}| \, \eta^2 \, \mathrm{d}s \right). \tag{80}$$

Using the interpolation estimates (47), (48) and (52):

$$\begin{aligned}
\|\eta\|_{L^2(\Omega)} &\leq C h^{p+1} |u|_{H^{p+1}(\Omega)}, \\
\|\mathbf{b} \cdot \nabla \eta\|_{L^2(\Omega)} \leq C \|\eta\|_{H^1(\Omega)} &\leq C h^p |u|_{H^{p+1}(\Omega)}, \\
\left(\int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}| \, \eta^2 \, \mathrm{d}s \right)^{1/2} \leq C \|\eta\|_{L^2(\partial\Omega)} &\leq C h^{p+1/2} |u|_{H^{p+1}(\Omega)},
\end{aligned} \tag{81}$$

and using $(a + b)^2 \leq 2(a^2 + b^2)$ we finally obtain (77). □

The estimate (77) shows that if the exact solution u to problem (59) happens to be smooth so that $\|u\|_{H^{p+1}(\Omega)}$ is finite, then the standard Galerkin method (69) will converge at the rate $O(h^p)$. Although this rate is one power of h from being optimal, it shows that the standard Galerkin method will perform satisfactorily in this case. However, in general u will not be smooth and in this case the standard Galerkin method gives poor results. In fact, for $u \in H^{s+1}(\Omega)$ with $s < p$ estimate (77) reduces to

$$\|u - u_h\|_{L^2(\Omega)} + \left(\int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, |u - u_h|^2 \, \mathrm{d}s \right)^{1/2} \leq C h^s |u|_{H^{s+1}(\Omega)}. \tag{82}$$

In particular, for $u \in H^1(\Omega)$ the solution to the standard Galerkin method does not converge as $h \to 0$. Instead, the solution might become oscillatory as there is no control on $\mathbf{b} \cdot \nabla u_h$.

3.3 The streamline diffusion method with weak boundary conditions

We have seen in the previous section that the standard Galerkin discretization for the linear advection equation is unstable. For obtaining a stable discretization scheme we need to add some artificial diffusion to the scheme. However, we do not require a stabilizing effect in all directions. As we will see in the following, for obtaining a stable scheme it is sufficient to add diffusion in streamline direction, only. Starting from the variational formulation (68) we replace u by a discrete function $u_h \in V_h$, and the test function v by $v_h + \delta h \mathbf{b} \cdot \nabla v_h$ on Ω with $\delta > 0$, and by v_h on Γ. Then, the streamline diffusion method is given by

$$B_h(u_h, v_h) = F_h(v_h), \quad \forall v_h \in V_h, \tag{83}$$

where

$$\begin{aligned} B_h(u,v) &= \int_\Omega \left(\nabla \cdot (\mathbf{b}u) + cu\right)\left(v + \delta h \mathbf{b} \cdot \nabla v\right)\, \mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, uv \, \mathrm{d}s, \\ F_h(v) &= \int_\Omega f\left(v + \delta h \mathbf{b} \cdot \nabla v\right)\, \mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, gv \, \mathrm{d}s. \end{aligned} \tag{84}$$

We note, that here trial and test functions u_h, v_h are taken from different discrete function spaces, $u_h \in V_h$ and $v_h \in \tilde{V}_h := \{v_h = w_h + \delta h \mathbf{b} \cdot \nabla w_h, \text{ with } w_h \in V_h\}$, $\tilde{V}_h = V_h + \delta h \mathbf{b} \cdot \nabla V_h$. Thereby, this is a so-called *Petrov-Galerkin* discretization which is in contrast to the standard *Galerkin* discretizations discussed so far, where the ansatz and test functions are taken from the same discrete function space $u_h, v_h \in V_h$. First, we note that (83) is consistent, i.e. we have

$$B_h(u,v) = F_h(v) \quad \forall v \in V,$$

for the exact solution u. Furthermore, we see that

$$B_h(u,v) = B(u,v) + \int_\Omega \left(\nabla \cdot \mathbf{b}u + \mathbf{b} \cdot \nabla u + cu\right) \delta h \mathbf{b} \cdot \nabla v \, \mathrm{d}\mathbf{x}, \tag{85}$$

where $B(\cdot,\cdot)$ is as defined in (17) for the standard Galerkin method. In particular, there is the additional term

$$\delta h \int_\Omega (\mathbf{b} \cdot \nabla u)(\mathbf{b} \cdot \nabla v)\, \mathrm{d}\mathbf{x} \tag{86}$$

which represents artificial viscosity (diffusion) in streamline direction $\mathbf{b}$. In fact, integrating by parts we see, that this term corresponds to $\partial^2_{\mathbf{b}} u$ where $\partial_{\mathbf{b}} u = \mathbf{b} \cdot \nabla u$.

In the following we show that B_h is $H^{1,\mathbf{b}}$-coercive.

Lemma 3.4 *Let* $\mathbf{b}$ *and* c *be constant and* $\delta h c \le \frac{1}{2}$. *Then there is a* $C > 0$ *such that for all* $v \in H^{1,\mathbf{b}}(\Omega)$ *we have*

$$B_h(v,v) \ge \gamma \left(h \|\mathbf{b} \cdot \nabla v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)} + \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| v^2 \, ds \right). \tag{87}$$

Proof: Using $\nabla \cdot \mathbf{b} = 0$ and $c = const$ we have

$$\int_\Omega cv(\mathbf{b} \cdot \nabla v)\, \mathrm{d}\mathbf{x} = \frac{1}{2} c \int_\Omega \nabla \cdot \left(\mathbf{b} v^2\right)\, \mathrm{d}\mathbf{x} = \frac{1}{2} c \int_\Gamma \mathbf{b} \cdot \mathbf{n} v^2 \, \mathrm{d}s.$$

Then, using (85), (73) and $\delta hc \leq \frac{1}{2}$ we obtain

$$\begin{aligned} B_h(v,v) &= B(v,v) + \delta h \|\mathbf{b}\cdot\nabla v\|^2_{L^2(\Omega)} + \frac{1}{2}\delta hc \int_\Gamma \mathbf{b}\cdot\mathbf{n} v^2 \,\mathrm{d}s \\ &\geq \gamma_0 \|v\|^2_{L^2(\Omega)} + \frac{1}{2}\int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s + \delta h \|\mathbf{b}\cdot\nabla v\|^2_{L^2(\Omega)} + \frac{1}{2}\delta hc \int_\Gamma \mathbf{b}\cdot\mathbf{n} v^2 \,\mathrm{d}s \\ &\geq \delta h \|\mathbf{b}\cdot\nabla v\|^2_{L^2(\Omega)} + \gamma_0 \|v\|^2_{L^2(\Omega)} + \frac{1}{4}\int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s \\ &\geq \gamma \left(h\|\mathbf{b}\cdot\nabla v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)} + \int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s \right), \end{aligned} \tag{88}$$

where $\gamma = \min(\delta, \gamma_0, 1/4)$. □

Motivated by (87) we define following norm on $H^{1,\mathbf{b}}(\Omega)$:

$$\|v\|_{H^{1,\mathbf{b}}(\Omega)} = \left(h\|\mathbf{b}\cdot\nabla v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)} + \int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s \right)^{\frac{1}{2}}. \tag{89}$$

Hence, with (87) we have shown $H^{1,\mathbf{b}}$-coercivity of $B_h(\cdot,\cdot)$,

$$B_h(v,v) \geq \gamma \|v\|^2_{H^{1,\mathbf{b}}(\Omega)}.$$

Additionally, by using Cauchy-Schwarz's inequality we find that $F_h(\cdot)$ is continuous,

$$\begin{aligned} |F_h(v)| &\leq |\int_\Omega f v \,\mathrm{d}\mathbf{x}| + \delta h |\int_\Omega f \mathbf{b}\cdot\nabla v \,\mathrm{d}\mathbf{x}| + |\int_\Gamma \mathbf{b}\cdot\mathbf{n}\, g v \,\mathrm{d}s| \\ &\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \delta h \|f\|_{L^2(\Omega)}\|\mathbf{b}\cdot\nabla v\|_{L^2(\Omega)} + C\|g\|_{L^2(\Gamma)} \left(\int_\Gamma |\mathbf{b}\cdot\mathbf{n}| v^2 \,\mathrm{d}s \right)^{\frac{1}{2}} \\ &\leq C_F \|v\|_{H^{1,\mathbf{b}}(\Omega)}. \end{aligned}$$

We can then deduce stability of the method as follows

$$\gamma \|u_h\|^2_{H^{1,\mathbf{b}}(\Omega)} \leq B_h(u_h,u_h) = F_h(u_h) \leq C_F \|u_h\|_{H^{1,\mathbf{b}}(\Omega)}. \tag{90}$$

Hence, we have

$$h\|\mathbf{b}\cdot\nabla u_h\|^2_{L^2(\Omega)} + \|u_h\|^2_{L^2(\Omega)} + \int_\Gamma |\mathbf{b}\cdot\mathbf{n}| u_h^2 \,\mathrm{d}s = \|u_h\|^2_{H^{1,\mathbf{b}}(\Omega)} \leq \left(\tfrac{C_F}{\gamma}\right)^2, \tag{91}$$

i.e. we have control of $\|u_h\|$ and $\|\mathbf{b}\cdot\nabla u_h\|$. In fact, the streamline diffusion method is stable.

Remark 3.5 *We recall, see Section 3.2, that the standard Galerkin method for the linear advection equation offers no control of $\|\mathbf{b}\cdot\nabla u_h\|$ which might lead to highly oscillating numerical solutions, hence an unstable scheme. Only by adding diffusion in streamline direction we gain control of $\|\mathbf{b}\cdot\nabla u_h\|$ and a stable scheme. We will show later, that in contrast to that, the discontinuous Galerkin discretizations of the linear advection equation is stable without streamline diffusion.*

Theorem 3.6 *Let $u \in H^{p+1}(\Omega)$ be the solution to* (59) *with constant* $\mathbf{b}$ *and* c*. Furthermore, let $u_h \in V^c_{h,p}$ the solution to* (83)*. Then,*

$$\|u - u_h\|_{H^{1,\mathbf{b}}(\Omega)} \leq C h^{p+1/2} |u|_{H^{p+1}(\Omega)}. \tag{92}$$

Proof: Let $e = u - u_h = \eta - \xi$ with $\eta = u - I_h u$ and $\xi = u_h - I_h u$ where $I_h : H^2(\Omega) \to V_{h,p}^c$ is the interpolation operator as defined in Definition 2.10. Then

$$\gamma \|e\|^2_{H^{1,\mathbf{b}}(\Omega)} \leq B_h(e,e) = B_h(e, \eta - \xi) = B_h(e, \eta), \tag{93}$$

where we used Galerkin orthogonality property $B_h(e, \xi) = 0$ for $\xi \in V_{h,p}^c$. Using the definition of $B_h(\cdot,\cdot)$, $\nabla \cdot \mathbf{b} = 0$ and the inequality $ab \leq \frac{\epsilon}{4}a^2 + \frac{1}{\epsilon}b^2$ we obtain

$$\begin{aligned}
\gamma \|e\|^2_{H^{1,\mathbf{b}}(\Omega)} \leq B_h(e,\eta) &= \int_\Omega \left(\nabla \cdot (\mathbf{b}e) + ce\right)\left(\eta + \delta h \mathbf{b} \cdot \nabla \eta\right) \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, e\eta \,\mathrm{d}s \\
&= \int_\Omega \left(\mathbf{b} \cdot \nabla e\right)\eta + \delta h \left(\mathbf{b} \cdot \nabla e\right)\left(\mathbf{b} \cdot \nabla \eta\right) + ce\eta + c\delta h e \left(\mathbf{b} \cdot \nabla \eta\right) \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, e\eta \,\mathrm{d}s \\
&\leq \tfrac{\gamma h}{4}\|\mathbf{b} \cdot \nabla e\|^2 + \tfrac{1}{\gamma h}\|\eta\|^2 + \tfrac{\gamma h}{4}\|\mathbf{b} \cdot \nabla e\|^2 + \tfrac{\delta^2 h}{\gamma}\|\mathbf{b} \cdot \nabla \eta\|^2 + \tfrac{\gamma}{4}\|e\|^2 + \tfrac{c^2}{\gamma}\|\eta\|^2 \\
&\quad + \tfrac{\gamma}{4}\|e\|^2 + \tfrac{c^2 h^2}{\gamma}\|\mathbf{b} \cdot \nabla \eta\|^2 + \tfrac{\gamma}{4}\int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}|\, e^2 \,\mathrm{d}s + \tfrac{1}{\gamma}\int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}|\, \eta^2 \,\mathrm{d}s \\
&\leq \tfrac{\gamma}{2}\|e\|^2_{H^{1,\mathbf{b}}(\Omega)} + C\left(h^{-1}\|\eta\|^2 + h\|\mathbf{b} \cdot \nabla \eta\|^2 + \int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}|\, \eta^2 \,\mathrm{d}s\right).
\end{aligned}$$

Subtracting $\frac{\gamma}{2}\|e\|^2_{H^{1,\mathbf{b}}(\Omega)}$ on both sides and multiplying with $\frac{2}{\gamma}$ we obtain

$$\|e\|^2_{H^{1,\mathbf{b}}(\Omega)} \leq C\left(h^{-1}\|\eta\|^2 + h\|\mathbf{b} \cdot \nabla \eta\|^2 + \int_{\Gamma_-} |\mathbf{b} \cdot \mathbf{n}|\, \eta^2 \,\mathrm{d}s\right). \tag{94}$$

Using the interpolation estimates (81) we obtain (92). □

We see, that the streamline diffusion discretization is of order $\mathcal{O}(h^{p+1/2})$ which is half an order higher than the $\mathcal{O}(h^p)$ obtained for the standard Galerkin discretization, see (77). Due to the streamline diffusion term (86), the streamline diffusion discretization is coercive with respect to the $H^{1,\mathbf{b}}$-norm which includes the $h\|\mathbf{b} \cdot \nabla v\|^2_{L^2(\Omega)}$ term, see (87) and (89). Thereby, in the proof of Theorem 3.6 we could subtract/hide the term $\frac{\gamma h}{4}\|\mathbf{b} \cdot \nabla e\|^2$ from/in the left hand side term $\gamma\|e\|^2_{H^{1,\mathbf{b}}(\Omega)}$. The remaining term $h^{1/2}\|\mathbf{b} \cdot \nabla \eta\|$ is $\mathcal{O}(h^{p+1/2})$.

In contrast to that the standard Galerkin discretization is coercive with respect to a weaker norm which does not include $\|\mathbf{b} \cdot \nabla v\|^2$, see (73). Thereby, in the proof of Theorem 3.3 the $\|\mathbf{b} \cdot \nabla \eta\|^2$ term could not be "hidden" in the left hand side and is finally estimated as $\mathcal{O}(h^p)$.

Remark 3.7 *Note, that there are many possibilities known as how to stabilize the standard Galerkin method of the linear advection equation: e.g. by using residual-free bubbles [12, 14], or subgrid modeling [21], among many others. However, for the purpose of this lecture it is sufficient to have error estimation results available for the classic SDFEM, only.*

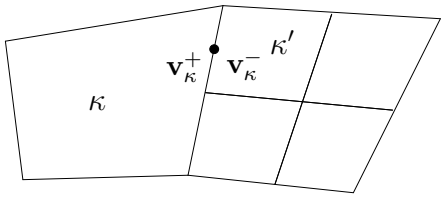


Figure 1: Definition of the interior and exterior traces $\mathbf{v}_\kappa^\pm$ wrt. element κ.

4 Higher order DG discretizations of the linear advection equation

4.1 Mesh related function spaces

We begin by introducing some notation. As before, we assume that the domain Ω can be subdivided into shape regular meshes $\mathcal{T}_h = \{\kappa\}$ consisting of elements κ. Here, h denotes the piecewise constant mesh function defined by $h|_\kappa \equiv h_\kappa = \mathrm{diam}(\kappa)$ for all $\kappa \in \mathcal{T}_h$.

In the following we define some broken (mesh related) function spaces on $\mathcal{T}_h$:

Definition 4.1 (Broken Sobolev space $H^m(\mathcal{T}_h)$) *By $H^m(\mathcal{T}_h)$ we denote the space of L^2 functions on Ω whose restriction to each element κ belongs to the Sobolov space $H^m(\kappa)$, i.e.*

$$H^m(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|_\kappa \in H^m(\kappa), \kappa \in \mathcal{T}_h\}. \tag{95}$$

Definition 4.2 (Interior faces: $\Gamma_\mathcal{I}$) *Let κ and κ' be two adjacent elements of $\mathcal{T}_h$ with common edge (interior face) $e = \partial\kappa \cap \partial\kappa'$. We define $\Gamma_\mathcal{I}$ to be the union of all interior faces of $\mathcal{T}_h$.*

Definition 4.3 (Traces v_κ^+, v_κ^- and the space $T(\mathcal{T}_h)$) *Suppose that $v \in H^1(\mathcal{T}_h)$, i.e. $v|_\kappa \in H^1(\kappa)$ for each $\kappa \in \mathcal{T}_h$. By $v_\kappa^\pm$ we denote the traces of v taken from within the interior of κ and κ', respectively, see Figure 1. We note, that for $v|_\kappa \in H^1(\kappa)$ the trace v_κ^+ belongs to $L^2(\partial\kappa)$, and traces of $v \in H^1(\mathcal{T}_h)$ belong to $T(\mathcal{T}_h) := \prod_{\kappa\in\mathcal{T}_h} L^2(\partial\kappa)$.*

Finally, we define mesh related (or broken) gradient, divergence and Laplace operators.

Definition 4.4 (∇_h, $\nabla_h\cdot$ and Δ_h) *We define broken operators by restriction to each element $\kappa \in \mathcal{T}_h$ as follows:*

- *The broken gradient operator $\nabla_h : H^1(\mathcal{T}_h) \to [L^2(\mathcal{T}_h)]^d$ is defined by*

$$(\nabla_h v)|_\kappa := \nabla(v|_\kappa), \quad \kappa \in \mathcal{T}_h, \tag{96}$$

  for $v \in H^1(\mathcal{T}_h)$, where $(\nabla v)_i = \partial_{x_i} v, i = 1, \ldots, d$.

- *The broken divergence operator $\nabla_h\cdot : [H^1(\mathcal{T}_h)]^d \to L^2(\mathcal{T}_h)$ is defined by*

$$(\nabla_h \cdot \tau)|_\kappa = \nabla \cdot (\tau|_\kappa), \quad \kappa \in \mathcal{T}_h, \tag{97}$$

  for $\tau \in [H^1(\mathcal{T}_h)]^d$, where $\nabla \cdot \tau = \sum_{1\le i\le d} \partial_{x_i}\tau_i$.

- *Finally, the broken Laplace operator $\Delta_h : H^2(\mathcal{T}_h) \to L^2(\mathcal{T}_h)$ is defined by*

$$(\Delta_h u)|_\kappa := \Delta(u|_\kappa), \quad \kappa \in \mathcal{T}_h, \tag{98}$$

  for $u \in H^2(\mathcal{T}_h)$, where $\Delta u = \nabla \cdot \nabla u = \sum_{1\le i\le d} \partial^2_{x_i} u$.

4.2 A variational formulation of the linear advection equation

Like in Section 3.1 here we consider the linear advection equation

$$Lu := \nabla \cdot (\mathbf{b}u) + cu = f \quad \text{in } \Omega, \qquad\qquad u = g \quad \text{on } \Gamma_-, \tag{99}$$

where $f \in L^2(\Omega)$, $\mathbf{b} \in [C^1(\Omega)]^d$, $c \in L^\infty(\Omega)$ and $g \in L^2(\Gamma_-)$, where

$$\Gamma_- = \{\mathbf{x} \in \Gamma, \mathbf{b}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\} \tag{100}$$

denotes the inflow part of the boundary $\Gamma = \partial\Omega$. Furthermore, we adopt following hypothesis: there exists a $c_0 \in L^\infty(\Omega)$ and a number $\gamma_0 > 0$ such that

$$c(\mathbf{x}) + \frac{1}{2}\nabla \cdot \mathbf{b}(\mathbf{x}) = c_0^2(\mathbf{x}) \geq \gamma_0 > 0. \tag{101}$$

In the following we derive a variational formulation for $u \in H^{1,\mathbf{b}}(\mathcal{T}_h)$ where

$$H^{1,\mathbf{b}}(\mathcal{T}_h) = \{u \in L^2(\Omega) : Lu = \nabla \cdot (\mathbf{b}u) + cu|_\kappa \in L^2(\kappa), \kappa \in \mathcal{T}_h\}, \tag{102}$$

is a broken space which is the mesh related (broken) counterpart of the function space $H^{1,\mathbf{b}}(\Omega)$ defined in (64). Note, that $H^{1,\mathbf{b}}(\Omega) \subset H^{1,\mathbf{b}}(\mathcal{T}_h)$. Given an element $\kappa \in \mathcal{T}_h$, we multiply (99) by a test function $v \in H^{1,\mathbf{b}}(\mathcal{T}_h)$, integrate over κ

$$\int_\kappa (\nabla \cdot (\mathbf{b}u) + cu)\, v \,\mathrm{d}\mathbf{x} = \int_\kappa fv \,\mathrm{d}\mathbf{x},$$

and integrate by parts

$$-\int_\kappa (\mathbf{b}u) \cdot \nabla v \,\mathrm{d}\mathbf{x} + \int_\kappa cuv \,\mathrm{d}\mathbf{x} + \int_{\partial\kappa} \mathbf{b} \cdot \mathbf{n}\, uv \,\mathrm{d}s = \int_\kappa fv \,\mathrm{d}\mathbf{x}. \tag{103}$$

We sum over all elements $\kappa \in \mathcal{T}_h$ and replace u on Γ_- by the boundary value function g,

$$-\int_\Omega (\mathbf{b}u) \cdot \nabla_h v \,\mathrm{d}\mathbf{x} + \int_\Omega cuv \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} \mathbf{b} \cdot \mathbf{n}\, uv \,\mathrm{d}s + \int_{\Gamma_+} \mathbf{b} \cdot \mathbf{n}\, uv \,\mathrm{d}s$$
$$= \int_\Omega fv \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, gv \,\mathrm{d}s \quad \forall v \in H^{1,\mathbf{b}}(\mathcal{T}_h).$$

As functions $u \in H^{1,\mathbf{b}}(\mathcal{T}_h)$ may be discontinuous across edges $e = \partial\kappa \cap \partial\kappa'$ between neighboring elements κ and κ' we replace $\mathbf{b}\cdot\mathbf{n}\, u$ on $\partial\kappa$ by a numerical flux function $\mathcal{H}(u^+, u^-, \mathbf{n})$, where $u^+ := u_\kappa^+$ and $u^- := u_\kappa^-$, respectively, are the interior and exterior traces of u on $\partial\kappa$. Thereby, the variational formulation of (99) is given by: find $u \in H^{1,\mathbf{b}}(\mathcal{T}_h)$ such that

$$B_h(u, v) = F(v) \quad \forall v \in H^{1,\mathbf{b}}(\mathcal{T}_h), \tag{104}$$

where

$$B_h(u, v) = -\int_\Omega (\mathbf{b}u) \cdot \nabla_h v \,\mathrm{d}\mathbf{x} + \int_\Omega cuv \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} \mathcal{H}(u^+, u^-, \mathbf{n})v \,\mathrm{d}s + \int_{\Gamma_+} \mathbf{b} \cdot \mathbf{n}\, uv \,\mathrm{d}s,$$
$$F(v) = \int_\Omega fv \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, gv \,\mathrm{d}s. \tag{105}$$

4.3 Consistency, conservation property, coercivity and stability

Definition 4.5 *A numerical flux function $\mathcal{H}(u^+, u^-, \mathbf{n})$ is said to be* consistent *if*

$$\mathcal{H}(u, u, \mathbf{n}) = \mathbf{b} \cdot \mathbf{n}\, u. \tag{106}$$

Furthermore, $\mathcal{H}(u^+, u^-, \mathbf{n})$ is said to be conservative *if*

$$\mathcal{H}(u^+, u^-, \mathbf{n}) = -\mathcal{H}(u^-, u^+, -\mathbf{n}). \tag{107}$$

Lemma 4.6 (Consistency) *Let problem (99) be discretized based on (104) and (105). Then, the discretization is consistent, i.e. the exact solution $u \in H^{1,\mathbf{b}}(\Omega) \subset H^{1,\mathbf{b}}(\mathcal{T}_h)$ to (99) satisfies*

$$B_h(u, v) = F(v) \quad \forall v \in H^{1,\mathbf{b}}(\mathcal{T}_h), \tag{108}$$

if and only if the numerical flux function $\mathcal{H}$ is consistent, i.e.

$$\mathcal{H}(u, u, \mathbf{n}) = \mathbf{b} \cdot \mathbf{n}\, u. \tag{109}$$

Proof: After integrating by parts (104), and rearranging terms we see that (104) is equivalent to

$$\int_\Omega (f - \nabla_h \cdot (\mathbf{b}u) - cu)\, v \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} \left(\mathbf{b} \cdot \mathbf{n}\, u^+ - \mathcal{H}(u^+, u^-, \mathbf{n})\right) v \,\mathrm{d}s - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n} \left(g - u^+\right) v \,\mathrm{d}s = 0 \quad \forall v \in H^{1,\mathbf{b}}(\mathcal{T}_h). \tag{110}$$

For the exact (and smooth) solution $u \in H^{1,\mathbf{b}}(\Omega)$ to problem (99), the first and third term in (110) vanishes. Thereby, we obtain

$$\sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} (\mathbf{b} \cdot \mathbf{n}\, u - \mathcal{H}(u, u, \mathbf{n}))\, v \,\mathrm{d}s = 0 \quad \forall v \in H^{1,\mathbf{b}}(\mathcal{T}_h), \tag{111}$$

hence we have consistency if and only if $\mathcal{H}(u, u, \mathbf{n}) = \mathbf{b} \cdot \mathbf{n}\, u$. □

Lemma 4.7 (Global conservation property) *Let problem (99) with $c \equiv 0$ be discretized based on (104) and (105). Then, the discretization is conservative, i.e.*

$$\int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, g \, ds + \int_{\Gamma_+} \mathbf{b} \cdot \mathbf{n}\, u \, ds = \int_\Omega f \, d\mathbf{x}, \tag{112}$$

if and only if the numerical flux function $\mathcal{H}$ is conservative, i.e.

$$\mathcal{H}(u^+, u^-, \mathbf{n}) = -\mathcal{H}(u^+, u^-, -\mathbf{n}).$$

Proof: By setting $v \equiv 1$ in (104) with $c \equiv 0$, we obtain

$$\sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} \mathcal{H}(u^+, u^-, \mathbf{n}) \,\mathrm{d}s + \int_{\Gamma_+} \mathbf{b} \cdot \mathbf{n}\, u \,\mathrm{d}s = \int_\Omega f \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, g \,\mathrm{d}s.$$

Then we note that in the sum $\sum_\kappa \int_{\partial\kappa}$ of all element faces each edge $e = \partial\kappa \cap \partial\kappa'$ between neighboring elements κ and κ' occurs twice with opposite normals and states u^+ and u^-. Thereby writing in terms of interior faces $e \in \Gamma_\mathcal{I}$ we obtain

$$\sum_{e \in \Gamma_\mathcal{I}} \int_e \mathcal{H}(u^+, u^-, \mathbf{n}) \;+\; \mathcal{H}(u^-, u^+, -\mathbf{n}) \,\mathrm{d}s \;+\; \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, g \,\mathrm{d}s \;+\; \int_{\Gamma_+} \mathbf{b} \cdot \mathbf{n}\, u \,\mathrm{d}s \;=\; \int_\Omega f \,\mathrm{d}\mathbf{x}.$$

Hence we see that (112) holds if and only if the numerical flux function is conservative. □

In the following we introduce two different numerical flux functions which are both consistent and conservative.

The mean value flux First we define the *mean value flux* (also named *central flux* in finite volume schemes) as follows

$$\mathcal{H}_{\mathrm{mv}}(u^+, u^-, \mathbf{n}) = \mathbf{b}\cdot\mathbf{n}\,\{u\}, \tag{113}$$

where

$$\{u\} = \frac{1}{2}\left(u^+ + u^-\right) \tag{114}$$

denotes the mean value of u^+ and u^-. This seems to be the most natural choice of a numerical flux function approximating $\mathbf{b}\cdot\mathbf{n}\,u$ based on u^+ and u^-. In fact, this flux is consistent and conservative. However, as we will show later, this flux leads to an *unstable* discontinuous Galerkin discretization.

The upwind flux We now define the *upwind flux* as follows:

$$\mathcal{H}_{\mathrm{uw}}(u^+, u^-, \mathbf{n}) = \begin{cases} \mathbf{b}\cdot\mathbf{n}\,u^-, & \text{for } (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) < 0, \text{ i.e. } \mathbf{x}\in\partial\kappa_-, \\ \mathbf{b}\cdot\mathbf{n}\,u^+, & \text{for } (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) \geq 0, \text{ i.e. } \mathbf{x}\in\partial\kappa_+, \end{cases}, \tag{115}$$

where $\partial\kappa_-$ and $\partial\kappa_+$ are the inflow and outflow boundaries of element κ defined as follows

$$\begin{aligned} \partial\kappa_- &= \{\mathbf{x}\in\partial\kappa, \mathbf{b}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) < 0\}, \\ \partial\kappa_+ &= \{\mathbf{x}\in\partial\kappa, \mathbf{b}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) \geq 0\} = \partial\kappa \setminus \partial\kappa_-. \end{aligned} \tag{116}$$

This flux always takes the value from upstream (upwind) direction. This numerical flux is consistent and conservative. Additionally, as we will show later, a discretization based on this flux is *stable*.

Generic flux The mean value flux and the upwind flux can be written as follows

$$\mathcal{H}_{b_0}(u^+, u^-, \mathbf{n}) = \mathbf{b}\cdot\mathbf{n}\,\{u\} + b_0\,[u], \tag{117}$$

where

$$[u] = u^+ - u^- \tag{118}$$

denotes the (simple) jump of u. By setting $b_0 = 0$ the generic flux (117) reduces to the mean value flux (113) and by setting $b_0 = \frac{1}{2}|\mathbf{b}\cdot\mathbf{n}|$ we obtain the upwind flux (115).

Theorem 4.8 (Coercivity) *Let $B_h(\cdot,\cdot)$ be given by*

$$\begin{aligned} B_h(u,v) = &-\int_\Omega (\mathbf{b}u)\cdot\nabla_h v\,d\mathbf{x} + \int_\Omega cuv\,d\mathbf{x} \\ &+ \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathcal{H}_{b_0}(u^+, u^-, \mathbf{n})v\,ds + \int_{\Gamma_+} \mathbf{b}\cdot\mathbf{n}\,uv\,ds, \end{aligned} \tag{119}$$

where $\mathcal{H}_{b_0}$ as defined in (117) *represents the mean value flux or the upwind flux depending on $b_0 = 0$ or $b_0 = \frac{1}{2}|\mathbf{b}\cdot\mathbf{n}|$ respectively. Then for all $v\in H^{1,\mathbf{b}}(\mathcal{T}_h)$ we have*

$$B_h(v,v) = \|c_0 v\|^2 + \sum_{e\in\Gamma_{\mathcal{I}}}\int_e b_0\,[v]^2\,ds + \frac{1}{2}\int_\Gamma |\mathbf{b}\cdot\mathbf{n}|\,v^2\,ds. \tag{120}$$

Proof: First we rewrite

$$\begin{aligned} -\int_\kappa (\mathbf{b}v)\cdot\nabla v\,\mathrm{d}\mathbf{x} &= -\frac{1}{2}\int_\kappa \mathbf{b}\cdot\nabla v^2\,\mathrm{d}\mathbf{x} \\ &= -\frac{1}{2}\int_\kappa \nabla\cdot\left(\mathbf{b}v^2\right)\,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\kappa \nabla\cdot\mathbf{b}v^2\,\mathrm{d}\mathbf{x} \\ &= -\frac{1}{2}\int_{\partial\kappa} \mathbf{b}\cdot\mathbf{n}(v^+)^2\,\mathrm{d}s + \frac{1}{2}\int_\kappa \nabla\cdot\mathbf{b}v^2\,\mathrm{d}\mathbf{x}. \end{aligned}$$

Furthermore, we have

$$-\frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa}\mathbf{b}\cdot\mathbf{n}(v^+)^2\,\mathrm{d}s+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathbf{b}\cdot\mathbf{n}\,\{v\}\,v\,\mathrm{d}s$$
$$=-\frac{1}{2}\sum_{e\in\Gamma_{\mathcal{I}}}\int_e\mathbf{b}\cdot\mathbf{n}^+\left((v^+)^2-(v^-)^2\right)\,\mathrm{d}s+\sum_{e\in\Gamma_{\mathcal{I}}}\mathbf{b}\cdot\mathbf{n}^+\frac{1}{2}(v^++v^-)(v^+-v^-)\,\mathrm{d}s=0.$$

Thereby,

$$\begin{aligned}B_h(v,v)&=-\int_\Omega(\mathbf{b}v)\cdot\nabla_h v\,\mathrm{d}\mathbf{x}+\int_\Omega cv^2\,\mathrm{d}\mathbf{x}+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\left(\mathbf{b}\cdot\mathbf{n}\,\{v\}+b_0\,[v]\right)v\,\mathrm{d}s+\int_{\Gamma_+}\mathbf{b}\cdot\mathbf{n}\,v^2\,\mathrm{d}s\\&=\int_\Omega\left(c+\frac{1}{2}\nabla\cdot\mathbf{b}\right)v^2\,\mathrm{d}\mathbf{x}+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}b_0\,[v]v\,\mathrm{d}s+\int_{\Gamma_+}\mathbf{b}\cdot\mathbf{n}\,v^2\,\mathrm{d}s-\frac{1}{2}\int_\Gamma\mathbf{b}\cdot\mathbf{n}\,v^2\,\mathrm{d}s\\&=\int_\Omega c_0^2v^2\,\mathrm{d}\mathbf{x}+\sum_{e\in\Gamma_{\mathcal{I}}}b_0\,[v][v]\,\mathrm{d}s+\frac{1}{2}\int_{\Gamma_+}\mathbf{b}\cdot\mathbf{n}\,v^2\,\mathrm{d}s-\frac{1}{2}\int_{\Gamma_-}\mathbf{b}\cdot\mathbf{n}v^2\,\mathrm{d}s\\&=\|c_0v\|^2+\sum_{e\in\Gamma_{\mathcal{I}}}\int_e b_0\,[v]^2\,\mathrm{d}s+\frac{1}{2}\int_\Gamma|\mathbf{b}\cdot\mathbf{n}|\,v^2\,\mathrm{d}s,\end{aligned}$$

where we used hypothesis (101). Hence, we have shown (120). □

Definition 4.9 *Motivated by the coercivity* (120) *we define the DG norm* $|||\cdot|||_{b_0}$ *by*

$$|||v|||_{b_0}^2=\|c_0v\|^2+\sum_{e\in\Gamma_{\mathcal{I}}}\int_e b_0\,[v]^2\,ds+\frac{1}{2}\int_\Gamma|\mathbf{b}\cdot\mathbf{n}|\,v^2\,ds\tag{121}$$

From the coercivity of B_h, (120), we immediately obtain the stability of the discontinuous Galerkin discretization in the DG norm $|||\cdot|||_{b_0}$ as follows:

$$\begin{aligned}|||v|||_{b_0}^2&=B_h(v,v)=F(v)=\int_\Omega fv\,\mathrm{d}\mathbf{x}+\int_{\Gamma_-}|\mathbf{b}\cdot\mathbf{n}|\,gv\,\mathrm{d}s\\&\le\|f\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)}+\left(\int_{\Gamma_-}|\mathbf{b}\cdot\mathbf{n}|\,g^2\,\mathrm{d}s\right)^{1/2}\left(\int_{\Gamma_-}|\mathbf{b}\cdot\mathbf{n}|\,v^2\,\mathrm{d}s\right)^{1/2}\le C|||v|||_{b_0},\end{aligned}$$

for $f\in L^2(\Omega)$ and $g\in L^2(\Gamma_-)$. After division by $|||v|||_{b_0}$ we obtain $|||v|||_{b_0}\le C$, and hence

$$|||v|||_{b_0}^2=\|c_0v\|^2+\sum_{e\in\Gamma_{\mathcal{I}}}\int_e b_0\,[v]^2\,\mathrm{d}s+\frac{1}{2}\int_\Gamma|\mathbf{b}\cdot\mathbf{n}|\,v^2\,\mathrm{d}s\le C^2.\tag{122}$$

Remark 4.10 *We note, that the discontinuous Galerkin discretization based on the upwind flux, i.e.* $b_0=\frac{1}{2}|\mathbf{b}\cdot\mathbf{n}|$*, has an improved stability as compared to the discretization based on the mean value flux where* $b_0=0$*. In fact, from* (122) *we can see that we have control of* $\sum_{e\in\Gamma_{\mathcal{I}}}\int_e[v]^2\,ds$ *for* $b_0\neq0$ *which we do not have for* $b_0=0$*. As we will see later, this translates to a reduced order of convergence for* $b_0=0$ *as compared to the case* $b_0=\frac{1}{2}|\mathbf{b}\cdot\mathbf{n}|$*. In fact, the DG discretization based on the upwind flux turns out to be stable whereas the DG discretization based on the mean value flux is unstable.*

This behavior corresponds to the difference in terms of stability and order of convergence of the standard (continuous) Galerkin method in comparison to the streamline diffusion method as discussed in Sections 3.2 and 3.3, respectively.

4.4 The discontinuous Galerkin discretization

Definition 4.11 *For $p \geq 0$ we define the space of discontinuous piecewise polynomials of degree p:*

$$V_{h,p}^d = \{v_h \in L^2(\Omega) : v_h|_\kappa \circ \sigma_\kappa \in Q_p(\hat{\kappa}) \text{ if } \hat{\kappa} \text{ is the unit hypercube, and } v_h|_\kappa \circ \sigma_\kappa \in P_p(\hat{\kappa}) \text{ if } \hat{\kappa} \text{ is the unit simplex}, \kappa \in \mathcal{T}_h\}, \tag{123}$$

where P_p and Q_p are the spaces of polynomials and tensor product polynomials of degree p.

Remark 4.12 *Note that $V_h := V_{h,p}^d \subset H^m(\mathcal{T}_h) \subset H^1(\mathcal{T}_h) \subset H^{1,\mathbf{b}}(\mathcal{T}_h)$, $m > 1$, and $u \in H^{1,\mathbf{b}}(\Omega) \subset H^{1,\mathbf{b}}(\mathcal{T}_h)$ but $V_h \not\subset H^{1,\mathbf{b}}(\Omega)$, i.e. DG methods are non-conforming finite element methods.*

For obtaining the DG discretization of the linear advection equation (99) we now replace the functions $u, v \in H^{1,\mathbf{b}}(\mathcal{T}_h)$ in (104) by discrete functions $u_h, v_h \in V_h$: find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h, \tag{124}$$

where

$$\begin{aligned} B_h(u, v) &= -\int_\Omega (\mathbf{b}u) \cdot \nabla_h v \,\mathrm{d}\mathbf{x} + \int_\Omega cuv \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} \mathcal{H}(u^+, u^-, \mathbf{n}) v \,\mathrm{d}s + \int_{\Gamma_+} \mathbf{b} \cdot \mathbf{n}\, uv \,\mathrm{d}s, \\ F(v) &= \int_\Omega fv \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, gv \,\mathrm{d}s. \end{aligned} \tag{125}$$

We recall that the numerical flux function $\mathcal{H}$ must be consistent and conservative. Examples are the mean value flux $\mathcal{H}_{\text{mv}}$ and the upwind flux $\mathcal{H}_{\text{uw}}$ defined in (113) and (115), respectively. They can be represented by the numerical flux function

$$\mathcal{H}_{b_0}(u^+, u^-, \mathbf{n}) = \mathbf{b} \cdot \mathbf{n}\, \{u\} + b_0\, [u], \tag{126}$$

which reduces to the mean value flux $\mathcal{H}_{\text{mv}}$ for $b_0 = 0$ and the upwind flux $\mathcal{H}_{\text{uw}}$ for $b_0 = \frac{1}{2}|\mathbf{b} \cdot \mathbf{n}|$.

Existence and uniqueness of a discrete solution Writing the discrete function $u_h(\mathbf{x}) = \sum_{0 \leq j < N_h} u_j \phi_j(\mathbf{x})$, in terms of the basis functions $\phi_j \in V_h$, $0 \leq j < N_h$, where $N_h = \# V_h$, and varying v_h over all basis functions, $v_h = \phi_i$, $0 \leq i < N_h$, problem (124) can be rewritten as a linear system

$$A\mathbf{u} = \mathbf{b}, \tag{127}$$

where $A \in \mathbb{R}^{N_h \times N_h}$ with $A_{ij} = B_h(\phi_j, \phi_i)$, $0 \leq i, j < N_h$, $\mathbf{b} \in \mathbb{R}^{N_h}$ with $b_i = F(\phi_i)$, $0 \leq i < N_h$, and $\mathbf{u} \in \mathbb{R}^{N_h}$. First we note, that from the coercivity (120) of B_h we have

$$\xi_i A_{ij} \xi_j = \xi_i B_h(\phi_j, \phi_i) \xi_j = B_h(\xi_j \phi_j, \xi_i \phi_i) = B_h(\boldsymbol{\xi}, \boldsymbol{\xi}) \geq \|c_0 \boldsymbol{\xi}\|^2 \geq \gamma_0 \|\boldsymbol{\xi}\|^2,$$

with a constant $\gamma_0 > 0$, i.e. A is positive definite. Given two solution vectors $\mathbf{u}^1$ and $\mathbf{u}^2$, with $A\mathbf{u}^1 = \mathbf{b}$ and $A\mathbf{u}^2 = \mathbf{b}$, respectively, we obtain $A\mathbf{u}^1 - A\mathbf{u}^2 = 0$ and $0 = (\mathbf{u}^1 - \mathbf{u}^2)_i A_{ij} (\mathbf{u}^1 - \mathbf{u}^2)_j$, thus $\mathbf{u}^1 = \mathbf{u}^2$. Hence the linear mapping $\mathbf{u} \to A\mathbf{u}$ is injective. We then use following

Lemma 4.13 *For a linear mapping $A : U \to W$ of finite dimensional spaces with $\dim U = \dim W$ following properties are equivalent: A is injective, A is surjective, and A is bijective.*

Proof: Any introductory book to linear algebra □

We conclude that $\mathbf{u} \to A\mathbf{u}$ is bijective. Hence there is a unique discrete solution u_h to (124).

Consistency and Galerkin orthogonality: According to Lemma 4.6 and due to the consistency of the numerical flux $\mathcal{H}$ we conclude that this discretization is consistent, i.e. the exact solution $u \in H^{1,\mathbf{b}}(\Omega)$ to (99) satisfies

$$B_h(u, v) = F(v) \quad \forall v \in H^{1,\mathbf{b}}(\mathcal{T}_h). \tag{128}$$

Again we obtain the Galerkin orthogonality

$$B_h(u - u_h, v_h) = 0 \quad \forall v \in V_h, \tag{129}$$

Before deriving *a priori* error estimates for the discontinuous Galerkin discretization (124), we give the definition of the L^2-projector onto $V_{h,p}^d$ together with some approximation estimates.

4.5 The local L^2-projection and approximation estimates

Definition 4.14 (Local L^2-projection) *Let $p \geq 0$ and $V_{h,p}^d$ be the discontinuous finite element space defined in* (123)*. Then, by $P_{h,p}^d$ we denote the L^2-projection onto $V_{h,p}^d$, i.e. given a $u \in L^2(\Omega)$ we define $P_{h,p}^d u \in V_{h,p}^d$ by*

$$\int_\Omega \left(u - P_{h,p}^d u\right) v_h \, d\mathbf{x} = 0 \quad \forall v_h \in V_{h,p}^d. \tag{130}$$

We use the short notation $P_h u$ instead of $P_{h,p}^d u$ when it is clear which projection is meant.

Given a $\kappa \in \mathcal{T}_h$, we set $v_h \equiv 0$ on $\kappa' \in \mathcal{T}_h$ with $\kappa' \neq \kappa$ in (130) and see that P_h has following *local projection* property: For any $\kappa \in \mathcal{T}_h$ we have

$$\int_\kappa (u - P_h u) \, v_h \, \mathrm{d}\mathbf{x} = 0 \quad \forall v_h \in V_{h,p}^d. \tag{131}$$

P_h restricted to $\kappa \in \mathcal{T}_h$ is an $L^2(\kappa)$-projection which is why P_h is also called *local* L^2-projection.

Remark 4.15 *We note that $P_{h,p}^d u$ may in fact be discontinuous from element to element. Furthermore, the approximation error estimates for the L^2-projection are similar to the approximation estimates for the interpolation $I_{h,p}^c$ defined in Section 2.2.4. However, the L^2-projection additionally offers the local projection property given in* (131)*. This property will be used when deriving* a priori *error estimates for the discontinuous Galerkin discretization. We note, that in continuous finite element methods we used the interpolation operator $I_{h,p}^c$ for ensuring global continuity of $I_{h,p}^c u \in V_{h,p}^c$. There, we could have also used a* global L^2*-projection $P_{h,p}^c$ onto the continuous discrete functions with $P_{h,p}^c u \in V_{h,p}^c$. However, this operator does not have a local projection property.*

Analogous to the interpolation estimates in Section 2.2.4 we have following approximation estimates for the L^2-projection:

Corollary 4.16 (Local approximation estimates for the L^2-projection) *Let $p \geq 0$ and $P_h u := P_{h,p}^d u$ be the L^2-projection defined in Definition 4.14. Suppose $u|_\kappa$ in $H^{s_\kappa+1}(\kappa)$, $s_\kappa \geq 0$, for $\kappa \in \mathcal{T}_h$. Then*

$$\|u - P_h u\|_{H^m(\kappa)} \leq C h_\kappa^{t_\kappa+1-m} |u|_{H^{t_\kappa+1}(\kappa)}, \tag{132}$$

where $t_\kappa = \min(s_\kappa, p)$, $\kappa \in \mathcal{T}_h$.

Again, for sufficiently smooth functions $u \in H^{p+1}(\kappa)$, i.e. $s_\kappa \geq p$, this estimate simplifies to

$$\|u - P_h u\|_{H^m(\kappa)} \leq C h_\kappa^{p+1-m} |u|_{H^{p+1}(\kappa)}, \tag{133}$$

while for functions with a lower smoothness, i.e. $u \in H^{s_\kappa+1}(\kappa)$ with $s_\kappa < p$, we have

$$\|u - P_h u\|_{H^m(\kappa)} \leq C h_\kappa^{s_\kappa+1-m} |u|_{H^{s+1}(\kappa)}, \tag{134}$$

for $m \leq s_\kappa + 1$. Furthermore, an analogous estimate holds in the L^∞-norm:

$$\|u - P_h u\|_{L^\infty(\kappa)} \leq C h_\kappa^{p+1} |u|_{H^{p+1,\infty}(\kappa)}, \tag{135}$$

Furthermore, analogous to Theorem 2.14 we have

$$\|u - P_h u\|_{L^2(\partial\kappa)} \leq C h_\kappa^{p+1/2} |u|_{H^{p+1}(\kappa)}. \tag{136}$$

4.6 *A priori* error estimates

Theorem 4.17 (A priori error estimate, [13]) *Let $u \in H^{p+1}(\Omega)$ be the exact solution to the linear advection equation (99). Furthermore, let $u_h \in V_{h,p}^d$ be the solution to*

$$B_h(u_h, v_h) = F(v_h), \quad \forall v_h \in V_{h,p}^d,$$

where

$$\begin{aligned} B_h(u,v) &= -\int_\Omega (\mathbf{b}u)\cdot\nabla_h v \, d\mathbf{x} + \int_\Omega cuv \, d\mathbf{x} \\ &\quad + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} (\mathbf{b}\cdot\mathbf{n}\,\{u\} + b_0\,[u])\, v \, ds + \int_{\Gamma_+} \mathbf{b}\cdot\mathbf{n}\, uv \, ds, \\ F(v) &= \int_\Omega fv \, d\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, gv \, ds. \end{aligned}$$

Then, for $b_0 = \frac{1}{2}|\mathbf{b}\cdot\mathbf{n}|$, i.e. when using the upwind flux, *we have*

$$|||u - u_h|||_{b_0} \leq C h^{p+1/2} |u|_{H^{p+1}(\Omega)}, \tag{137}$$

and for $b_0 = 0$, i.e. when using the mean value flux, *we have*

$$|||u - u_h|||_{b_0} \leq C h^{p} |u|_{H^{p+1}(\Omega)}, \tag{138}$$

where

$$|||v|||_{b_0}^2 = \|c_0 v\|^2 + \sum_{e\in\Gamma_\mathcal{I}} \int_e b_0\,[v]^2 \, ds + \frac{1}{2}\int_\Gamma |\mathbf{b}\cdot\mathbf{n}|\, v^2 \, ds \tag{139}$$

is the DG norm as defined in Definition 4.9.

Proof: Let $e = u - u_h = \eta - \xi$ with $\eta = u - P_h u$ and $\xi = u_h - P_h u$ where $P_h := P_{h,p}^d$ is the L^2-projection onto $V_h := V_{h,p}^d$ as defined in Definition (4.14). Then, by triangle inequality,

$$|||e|||_{b_0} \leq |||\eta|||_{b_0} + |||\xi|||_{b_0}. \tag{140}$$

For the first term we use approximation results for $P_h u$:

$$\begin{aligned} \|c_0\eta\|_{L^2(\Omega)} &\leq C\|u - P_h u\|_{L^2(\Omega)} \leq C h^{p+1} |u|_{H^{p+1}(\Omega)}, \\ \left(\sum_{e\in\Gamma_\mathcal{I}} \int_e b_0\,[\eta]^2 \,\mathrm{d}s\right)^{1/2} &\leq C \sum_{e\in\Gamma_\mathcal{I}} \|u - P_h u\|_{L^2(e)} \leq C h^{p+1/2} |u|_{H^{p+1}(\Omega)}, \\ \left(\frac{1}{2}\int_{\Gamma_-} |\mathbf{b}\cdot\mathbf{n}|\,\eta^2 \,\mathrm{d}s\right)^{1/2} &\leq C\|u - P_h u\|_{L^2(\partial\Omega)} \leq C h^{p+1/2} |u|_{H^{p+1}(\Omega)}, \end{aligned} \tag{141}$$

see (133) for $m = 0$ and (136), and hence

$$|||\eta|||_{b_0} \leq h^{p+1/2}|u|_{H^{p+1}(\Omega)}. \tag{142}$$

The second term in (140) we rewrite as follows

$$|||\xi|||_{b_0}^2 = B_h(\xi,\xi) = B_h(\eta - e, \xi) = B_h(\eta,\xi), \tag{143}$$

where we used coercivity (120) of B_h and the Galerkin orthogonality property $B_h(e,\xi) = 0$ for $\xi \in V_h$. Using the definition of $B_h(\cdot,\cdot)$ we obtain

$$|||\xi|||_{b_0}^2 = \int_\Omega \eta\left(-\mathbf{b}\cdot\nabla_h\xi + c\xi\right)\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\left(\mathbf{b}\cdot\mathbf{n}\,\{\eta\} + b_0\,[\eta]\right)\xi\,\mathrm{d}s + \int_{\Gamma_+}\mathbf{b}\cdot\mathbf{n}\,\eta\xi\,\mathrm{d}s.$$

Next observe that $\nabla_h\xi \in V_h$, so that, by the definition of the projector $P_{h,0}^d$,

$$\int_\kappa \left(P_{h,0}^d\mathbf{b}\cdot\nabla\xi\right)\eta\,\mathrm{d}\mathbf{x} = 0.$$

Using this, together with the approximation estimate (135) for $p = 0$, the Cauchy-Schwarz inequality, the inverse inequality $\|\xi\|_{H^1(\kappa)} \leq Ch_\kappa^{m-1}\|\xi\|_{H^m(\kappa)}$ for $\xi \in V_h$, and the approximation estimate (133) for $m = 0$, we deduce that

$$\begin{aligned}
\int_\Omega \eta\left(-\mathbf{b}\cdot\nabla_h\xi + c\xi\right)\,\mathrm{d}\mathbf{x} &= \sum_{\kappa\in\mathcal{T}_h}\int_\kappa \eta\left(-\mathbf{b}\cdot\nabla\xi + c\xi\right)\,\mathrm{d}\mathbf{x}\\
&= \sum_{\kappa\in\mathcal{T}_h}\int_\kappa \eta\left(\left(P_{h,0}^d\mathbf{b} - \mathbf{b}\right)\cdot\nabla\xi + c\xi\right)\,\mathrm{d}\mathbf{x}\\
&\leq C\|\eta\|_{L^2(\Omega)}\left(\sum_{\kappa\in\mathcal{T}_h}\left(\|P_h^0\mathbf{b} - \mathbf{b}\|_{L^\infty(\kappa)}|\xi|_{H^1(\kappa)} + \|\xi\|_{L^2(\kappa)}\right)^2\right)^{\frac{1}{2}}\\
&\leq C\|\eta\|_{L^2(\Omega)}\left(\sum_{\kappa\in\mathcal{T}_h}\left(h_\kappa|\mathbf{b}|_{H^{1,\infty}(\kappa)}h_\kappa^{-1}\|\xi\|_{L^2(\kappa)} + \|\xi\|_{L^2(\kappa)}\right)^2\right)^{\frac{1}{2}}\\
&\leq C\|\eta\|_{L^2(\Omega)}\|\xi\|_{L^2(\Omega)}\\
&\leq Ch^{p+1}|u|_{H^{p+1}(\Omega)}\,|||\xi|||_{b_0}.
\end{aligned} \tag{144}$$

Furthermore, using Cauchy-Schwarz inequality and approximation estimate (136) we find

$$\begin{aligned}
\int_{\Gamma_+}\mathbf{b}\cdot\mathbf{n}\,\eta\xi\,\mathrm{d}s &\leq \left(\int_{\Gamma_+}|\mathbf{b}\cdot\mathbf{n}|\,\eta^2\,\mathrm{d}s\right)^{\frac{1}{2}}\left(\int_{\Gamma_+}|\mathbf{b}\cdot\mathbf{n}|\,\xi^2\,\mathrm{d}s\right)^{\frac{1}{2}}\\
&\leq Ch^{p+1/2}|u|_{H^{p+1}(\Omega)}|||\xi|||_{b_0}.
\end{aligned} \tag{145}$$

Finally, we have

$$\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\left(\mathbf{b}\cdot\mathbf{n}\,\{\eta\} + b_0\,[\eta]\right)\xi\,\mathrm{d}s = \sum_{e\in\Gamma_\mathcal{I}}\int_e\left(\mathbf{b}\cdot\mathbf{n}\,\{\eta\} + b_0\,[\eta]\right)[\xi]\,\mathrm{d}s.$$

If $b_0 = \frac{1}{2}|\mathbf{b}\cdot\mathbf{n}|$ we have $\mathbf{b}\cdot\mathbf{n} \leq |\mathbf{b}\cdot\mathbf{n}| = 2b_0$ and obtain

$$\begin{aligned}\int_e (\mathbf{b}\cdot\mathbf{n}\{\eta\} + b_0\,[\eta])\,[\xi]\,\mathrm{d}s &\leq \int_e b_0^{1/2}\,(2\{\eta\} + [\eta])\,b_0^{1/2}|[\xi]|\,\mathrm{d}s \\ &\leq \left(\int_e b_0\,(2\{\eta\} + [\eta])^2\,\mathrm{d}s\right)^{\frac{1}{2}} \left(\int_e b_0[\xi]^2\,\mathrm{d}s\right)^{\frac{1}{2}} \\ &\leq Ch_\kappa^{p+1/2}|u|_{H^{p+1}(\kappa)} \left(\int_e b_0[\xi]^2\,\mathrm{d}s\right)^{\frac{1}{2}},\end{aligned}$$

and hence

$$\begin{aligned}\sum_{e\in\Gamma_\mathcal{I}}\int_e (\mathbf{b}\cdot\mathbf{n}\{\eta\} + b_0\,[\eta])\,[\xi]\,\mathrm{d}s &\leq Ch^{p+1/2}|u|_{H^{p+1}(\Omega)} \left(\sum_{e\in\Gamma_\mathcal{I}}\int_e b_0[\xi]^2\,\mathrm{d}s\right)^{\frac{1}{2}} \\ &\leq Ch^{p+1/2}|u|_{H^{p+1}(\Omega)}|||\xi|||_{b_0}.\end{aligned} \tag{146}$$

However, if $b_0 = 0$, the norm $|||\xi|||_{b_0}$ does not include $\int_e [\xi]^2\,\mathrm{d}s$ which is why we cannot bound $\int_e \mathbf{b}\cdot\mathbf{n}\{\eta\}[\xi]\,\mathrm{d}s$ in terms of $|||\xi|||_{b_0}$. Thereby, we are forced to use the inverse inequality

$$\|[\xi]\|_{L^2(e)} \leq Ch_\kappa^{-\frac{1}{2}}\|\xi\|_{L^2(\kappa)},$$

to bound the $L^2(e)$-norm by the $L^2(\kappa)$-norm. Hence, instead of (146) we obtain

$$\int_e \mathbf{b}\cdot\mathbf{n}\{\eta\}[\xi]\,\mathrm{d}s \leq C\|[\eta]\|_{L^2(e)}\|[\xi]\|_{L^2(e)} \leq Ch^{p+1/2}|u|_{H^{p+1}(\Omega)}h_\kappa^{-\frac{1}{2}}\|\xi\|_{L^2(\kappa)},$$

and hence,

$$\begin{aligned}\sum_{e\in\Gamma_\mathcal{I}}\int_e \mathbf{b}\cdot\mathbf{n}\{\eta\}[\xi]\,\mathrm{d}s &\leq Ch^p|u|_{H^{p+1}(\Omega)}\|\xi\|_{L^2(\kappa)} \\ &\leq Ch^p|u|_{H^{p+1}(\Omega)}|||\xi|||_{b_0}.\end{aligned} \tag{147}$$

Combining (144), (145) and (146) we obtain (137), while (144), (145) and (147) gives (138). □

From the proof above we see that it is essential whether the term $\sum_{e\in\Gamma_\mathcal{I}}\int_e b_0\,[v]^2\,\mathrm{d}s$ is included in the DG norm with $b_0 \neq 0$ or with $b_0 = 0$. In the former case we can bound the interior face terms

$$\sum_{e\in\Gamma_\mathcal{I}}\int_e (\mathbf{b}\cdot\mathbf{n}\{\eta\} + b_0\,[\eta])\,[\xi]\,\mathrm{d}s \tag{148}$$

in terms of $|||\xi|||_{b_0}$ whereas in the latter case we are forced to use the inverse inequality due to which we loose half an order of h. Recalling the discussion in Remark 4.10 we see that the stability of the discretization, in particular of the interior face terms, is connected to the order of convergence. For sufficiently smooth solutions, $u \in H^{p+1}(\Omega)$, the discretization based on the upwind flux is of order $\mathcal{O}(h^{p+1/2})$ and the discretization based on the mean value flux is of the order $\mathcal{O}(h^p)$. In contrast to that the order of convergence will be reduced for solutions with a lower smoothness. In fact, for $u \in H^{s+1}(\Omega)$ with $s < p$ the estimates (137) and (138) are replaced by the estimates

$$|||u - u_h|||_{b_0} \leq Ch^{s+1/2}|u|_{H^{s+1}(\Omega)}, \tag{149}$$

for the upwind flux, and by

$$|||u - u_h|||_{b_0} \leq Ch^s|u|_{H^{s+1}(\Omega)}, \tag{150}$$

for the mean value flux, respectively. In particular, for $u \in H^1(\Omega)$ we see that the discontinuous Galerkin solution based on the mean value flux does not converge under $h \to 0$. In fact, for $u \in H^1(\Omega)$, the discretization based on the mean value flux is unstable, whereas the upwind flux yields a stable discretization.

We recall that the difference in stability and the order of convergence of the discontinuous Galerkin discretization based on the mean value as compared to the upwind flux corresponds to the difference in the standard (continuous) Galerkin method as compared to the streamline diffusion method discussed in Sections 3.2 and 3.3, respectively.

Finally, we note that estimate (137) which is suboptimal by $h^{1/2}$ as compared to the $\mathcal{O}(h^{p+1})$ approximation order of $V_{h,p}^d$, Peterson [37] confirmed by considering so–called Peterson meshes, that $O(h^{p+1/2})$ is actually a sharp estimate.

4.7 The discontinuous Galerkin discretization based on upwind

As shown in the previous section the discontinuous Galerkin discretization for the linear advection equation based on the upwind flux is stable whereas that based on the mean value flux is unstable. Therefore, in this subsection we concentrate on the discretization based on the upwind flux while ignoring the discretization based on the mean value flux.

First, we recall the discontinuous Galerkin discretization (124) for the linear advection: find $u_h \in V_{h,p}^d$ such that

$$-\int_\Omega (\mathbf{b}u_h)\cdot\nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_\Omega c u_h v_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathcal{H}(u_h^+, u_h^-, \mathbf{n}) v_h^+ \,\mathrm{d}s + \int_{\Gamma_+} \mathbf{b}\cdot\mathbf{n}\, u_h v_h \,\mathrm{d}s$$
$$= \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, g v_h \,\mathrm{d}s \quad \forall v_h \in V_{h,p}^d.$$

As shown in the previous sections the numerical flux function must be consistent and conservative, see Definition 4.5. The upwind flux

$$\mathcal{H}_{\mathrm{uw}}(u^+, u^-, \mathbf{n}) = \begin{cases} \mathbf{b}\cdot\mathbf{n}\, u^-, & \text{for } (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) < 0, \text{ i.e. } \mathbf{x} \in \partial\kappa_-, \\ \mathbf{b}\cdot\mathbf{n}\, u^+, & \text{for } (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) \geq 0, \text{ i.e. } \mathbf{x} \in \partial\kappa_+, \end{cases}$$

is consistent and conservative and we have shown in previous sections that it yields a stable scheme. Using this flux the discontinuous Galerkin discretization of the linear advection equation is given by: find $u_h \in V_{h,p}^d$ such that

$$-\int_\Omega (\mathbf{b}u_h)\cdot\nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_\Omega c u_h v_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_-\setminus\Gamma} \mathbf{b}\cdot\mathbf{n}\, u_h^- v_h^+ \,\mathrm{d}s + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_+} \mathbf{b}\cdot\mathbf{n}\, u_h^+ v_h^+ \,\mathrm{d}s$$
$$= \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, g v_h \,\mathrm{d}s \quad \forall v_h \in V_{h,p}^d. \quad (151)$$

Integrating back by parts on each element κ we obtain following equivalent form of the discretization: find $u_h \in V_{h,p}^d$ such that

$$\int_\Omega \left(\nabla_h\cdot(\mathbf{b}u_h) + c u_h\right) v_h \,\mathrm{d}\mathbf{x} - \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_-\setminus\Gamma} \mathbf{b}\cdot\mathbf{n}\,(u_h^+ - u_h^-) v_h^+ \,\mathrm{d}s - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, u_h v_h \,\mathrm{d}s$$
$$= \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b}\cdot\mathbf{n}\, g v_h \,\mathrm{d}s \quad \forall v_h \in V_{h,p}^d. \quad (152)$$

In contrast to the discontinuous Galerkin discretization of viscous terms where several different discretization schemes based on different choices of numerical fluxes can be derived, see Section 5, the discretization in (152) is known as *the* (one and only "reasonable") discontinuous Galerkin discretization of the linear advection equation.

4.7.1 The importance of the inter-element jump terms

We note that the discrete functions $u_h \in V_{h,p}^d$ may be discontinuous on interior faces between neighboring elements. Originating from the upwind flux, the interior face terms

$$\sum_{\kappa \in \mathcal{T}_h} \int_{\partial \kappa_- \setminus \Gamma} \mathbf{b} \cdot \mathbf{n} \, (u_h^+ - u_h^-) v_h^+ \, \mathrm{d}s \tag{153}$$

include the jump $[u_h] = u_h^+ - u_h^-$ of the discrete solution and impose continuity of the discrete solution in a weak sense. If these inter-element jump terms were neglected the discrete problem would – due to the discontinuity of the ansatz and test functions – decouple into one local problem per element with no data coupling to neighboring elements and hence to an inconsistent discretization.

For getting further insight into the importance of these jump terms, let us, for a moment, evaluate the bilinear form in (152) for functions $u_h, v_h \in V_{h,p}^c \subset C^0(\Omega)$, i.e. the discrete space $V_{h,p}^c$ of continuous piecewise polynomial functions given in Definition 2.6. Then the jump terms vanish and we obtain: find $u_h \in V_{h,p}^c$ such that

$$\int_\Omega (\nabla_h \cdot (\mathbf{b} u_h) + c u_h) \, v_h \, \mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n} \, u_h v_h \, \mathrm{d}s = \int_\Omega f v_h \, \mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n} \, g v_h \, \mathrm{d}s, \tag{154}$$

for all $v_h \in V_{h,p}^c$. This is the standard Galerkin discretization with weakly imposed boundary conditions of the linear advection equation based on continuous finite element methods as introduced in Section 3.1.2. We recall that this discretization is unstable and of order $\mathcal{O}(h^p)$. In contrast to that we have seen that the discontinuous Galerkin discretization (152) based on the upwind flux is stable and of order $\mathcal{O}(h^{p+1/2})$. From this we see, that allowing inter-element discontinuities and imposing continuity in a weak sense has a stabilizing effect on the discretization and yields an improved order of convergence as compared to the standard (continuous) Galerkin discretization.

Finally, given a $\kappa \in \mathcal{T}_h$ and setting $v_h \equiv 0$ on all $\kappa' \neq \kappa$, $\kappa' \in \mathcal{T}_h$, in (152) we obtain

$$\int_\kappa (\nabla \cdot (\mathbf{b} u_h) + c u_h) \, v_h \, \mathrm{d}\mathbf{x} - \int_{\partial \kappa_-} \mathbf{b} \cdot \mathbf{n} \, u_h^+ v_h^+ \, \mathrm{d}s = \int_\kappa f v_h \, \mathrm{d}\mathbf{x} - \int_{\partial \kappa_-} \mathbf{b} \cdot \mathbf{n} \, u_h^- v_h^+ \, \mathrm{d}s, \tag{155}$$

where u_h^- is replaced by g on $\partial \kappa_- \cap \Gamma_-$. This corresponds to the standard Galerkin discretization with weakly imposed boundary conditions, see (154). However, in (155) we consider a single element κ instead of the whole domain Ω. Furthermore, on $\partial \kappa_- \setminus \Gamma -$ the discretization (155) includes the exterior trace u_h^- instead of the boundary function g as in (154). In fact, u_h^-, i.e. the value of u_h on a neighboring element, can be considered as boundary function to the linear advection problem (99) localized on κ. In summary, using the upwind flux in the discontinuous Galerkin discretization of the linear advection equation corresponds to a weak imposition of boundary conditions on each element $\kappa \in \mathcal{T}_h$.

4.7.2 The global and local conservation property

We recall from Lemma 4.7 that a discretization based on any conservative flux $\mathcal{H}$ has a global conservation property. The upwind flux is conservative, hence the discretization based on the upwind flux, see (151) or (152), has a global conservation property.

Ignoring Lemma 4.7 for a moment we want to (re)show the global conservation property of the discretization based on the upwind flux. To this end, we rewrite (151) with $c \equiv 0$ in terms of interior faces $e \in \Gamma_{\mathcal{I}}$. For any two neighboring elements κ and κ' with $e := \partial \kappa_+ \cap \partial \kappa'_- \neq \emptyset$ we

rewrite $\mathbf{b}\cdot\mathbf{n}_{\kappa'}\, u^-_{\kappa'} v^+_{\kappa'}$ on $\partial\kappa'_-$ as $\mathbf{b}\cdot(-\mathbf{n}_\kappa)\, u^+_\kappa v^-_\kappa$ on $\partial\kappa_+$ and obtain

$$-\int_\Omega (\mathbf{b}u_h)\cdot\nabla_h v_h \,\mathrm{d}\mathbf{x} + \sum_{e\subset\Gamma_{\mathcal{I}}}\int_e \mathbf{b}\cdot\mathbf{n}\, u_h^+ \left(v_h^+ - v_h^-\right)\,\mathrm{d}s + \int_{\Gamma_+}\mathbf{b}\cdot\mathbf{n}\, u_h^+ v_h^+\,\mathrm{d}s$$
$$= \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-}\mathbf{b}\cdot\mathbf{n}\, g v_h \,\mathrm{d}s \quad \forall v_h \in V^d_{h,p}. \tag{156}$$

Setting $v_h \equiv 1$ on Ω we obtain the *global conservation property*:

$$\int_{\Gamma_-}\mathbf{b}\cdot\mathbf{n}\, g\,\mathrm{d}s + \int_{\Gamma_+}\mathbf{b}\cdot\mathbf{n}\, u_h^+\,\mathrm{d}s = \int_\Omega f\,\mathrm{d}\mathbf{x}. \tag{157}$$

Furthermore, given a $\kappa\in\mathcal{T}_h$ and setting $v_h\equiv 1$ on κ and $v_h\equiv 0$ on all $\kappa'\neq\kappa$, $\kappa'\in\mathcal{T}_h$, in (151) with $c\equiv 0$ or in (156) we obtain the *local conservation property*:

$$\int_{\partial\kappa_-}\mathbf{b}\cdot\mathbf{n}\, u_h^-\,\mathrm{d}s + \int_{\partial\kappa_+}\mathbf{b}\cdot\mathbf{n}\, u_h^+\,\mathrm{d}s = \int_\kappa f\,\mathrm{d}\mathbf{x}, \tag{158}$$

where u_h^- is replaced by g on $\kappa_-\cap\Gamma_-$.

4.7.3 Consistency

We recall from Lemma 4.7 that a discretization based on any consistent flux $\mathcal{H}$ is consistent. The upwind flux is consistent, hence the discretization based on the upwind flux, see (151) or (152), is consistent. In the following we present another way to show that the discretization based on the upwind flux is consistent. From (152) we see that the discrete solution $u_h\in V^d_{h,p}$ satisfies the *primal residual form*:

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa R(u_h)v\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} r(u_h)v\,\mathrm{d}s + \int_\Gamma r_\Gamma(u_h)v\,\mathrm{d}s = 0 \quad \forall v\in V^d_{h,p}, \tag{159}$$

where $R(u_h)$, $r(u_h)$ and $r_\Gamma(u_h)$ denote the element, interior face and boundary residuals, respectively, given by

$$\begin{aligned} R(u_h) &= f - \nabla_h\cdot(\mathbf{b}u_h) - cu_h && \text{in } \kappa,\kappa\in\mathcal{T}_h,\\ r(u_h) &= \mathbf{b}\cdot\mathbf{n}\,(u_h^+ - u_h^-) && \text{on } \partial\kappa_-\setminus\Gamma, \kappa\in\mathcal{T}_h,\\ r_\Gamma(u_h) &= \mathbf{b}\cdot\mathbf{n}(u_h - g) && \text{on } \Gamma_-, \end{aligned}$$

and $r_\Gamma(u_h)\equiv 0$ on Γ_+. We see that the exact solution $u\in H^{1,\mathbf{b}}(\Omega)$ to (99) satisfies

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa R(u)v\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} r(u)v\,\mathrm{d}s + \int_\Gamma r_\Gamma(u)v\,\mathrm{d}s = 0 \quad \forall v\in H^{1,\mathbf{b}}(\mathcal{T}_h), \tag{160}$$

with $R(u)=0$, $r(u)=0$ and $r_\Gamma(u)=0$. Thereby, we have consistency of the discretization, i.e. the exact solution $u\in H^{1,\mathbf{b}}(\Omega)$ to (99) satisfies

$$B_h(u,v) = F(v) \quad \forall v\in H^{1,\mathbf{b}}(\mathcal{T}_h). \tag{161}$$

5 Higher order DG discretizations of Poisson's equation

In the following we consider the elliptic model problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D, \qquad \mathbf{n}\cdot\nabla u = g_N \quad \text{on } \Gamma_N, \tag{162}$$

where $f \in L^2(\Omega)$, $g_D \in L^2(\Gamma_D)$ and $g_N \in L^2(\Gamma_N)$ are given functions. We assume that Γ_D and Γ_N are disjoint subsets with union Γ, that is $\Gamma_D \cup \Gamma_N = \Gamma$ and $\Gamma_D \cap \Gamma_N = \emptyset$. Furthermore, we assume that $\Gamma_D \neq \emptyset$. Problem (162) represents the general Dirichlet-Neumann problem of Poisson's equation. In case of $\Gamma_D = \Gamma$ and $g_D = 0$ this represents the Dirichlet problem with homogeneous boundary conditions given by

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma. \tag{163}$$

Let $H^m(\mathcal{T}_h)$ be the *broken* Sobolev space consisting of functions $v \in L^2(\Omega)$ whose restriction to each element $\kappa \in \mathcal{T}_h$ belongs to the Sobolev space $H^m(\kappa)$. Analogously, by $[H^m(\mathcal{T}_h)]^d$ we denote the d-vector valued broken Sobolov space.

5.1 The system and primal flux formulation

In this and the following sections we recall the common framework for deriving DG discretizations of the homogeneous Dirichlet problem (163) from [2] and apply it to the Dirichlet-Neumann problem (162). We begin by rewriting (162) as a first-order system as follows

$$\sigma = \nabla u, \qquad -\nabla\cdot\sigma = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D, \qquad \mathbf{n}\cdot\nabla u = g_N \quad \text{on } \Gamma_N. \tag{164}$$

Assuming $u \in H^2(\mathcal{T}_h)$ and $\sigma \in [H^1(\mathcal{T}_h)]^d$ we multiply the first and second equation by test functions $\tau \in [H^1(\mathcal{T}_h)]^d$ and $v \in H^1(\mathcal{T}_h)$, respectively, integrate over an element $\kappa \in \mathcal{T}_h$, and integrate by parts. Thus

$$\begin{aligned} \int_\kappa \sigma\cdot\tau \,\mathrm{d}\mathbf{x} &= -\int_\kappa u\nabla\cdot\tau \,\mathrm{d}\mathbf{x} + \int_{\partial\kappa} u\,\tau\cdot\mathbf{n}\,\mathrm{d}s, \\ \int_\kappa \sigma\cdot\nabla v \,\mathrm{d}\mathbf{x} &= \int_\kappa f v \,\mathrm{d}\mathbf{x} + \int_{\partial\kappa} \sigma\cdot\mathbf{n}\, v \,\mathrm{d}s, \end{aligned} \tag{165}$$

where $\mathbf{n}$ is the unit outward normal vector to $\partial\kappa$. Then, we sum (165) over all elements $\kappa \in \mathcal{T}_h$. As u and σ may be discontinuous across inter-element faces $\partial\kappa \setminus \Gamma$, $\kappa \in \mathcal{T}_h$, they must be replaced by *numerical flux functions* $\hat{u}$ and $\hat{\sigma}$ which are approximations to u and $\sigma = \nabla u$, respectively. Thus we obtain following *system flux formulation*: find $u \in H^2(\mathcal{T}_h)$ and $\sigma \in [H^1(\mathcal{T}_h)]^d$ such that

$$\int_\Omega \sigma\cdot\tau \,\mathrm{d}\mathbf{x} = -\int_\Omega u\nabla_h\cdot\tau \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa} \hat{u}\tau\cdot\mathbf{n}\,\mathrm{d}s \quad \forall\tau \in [H^1(\mathcal{T}_h)]^d, \tag{166}$$

$$\int_\Omega \sigma\cdot\nabla_h v \,\mathrm{d}\mathbf{x} = \int_\Omega f v \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa} \hat{\sigma}\cdot\mathbf{n}\, v \,\mathrm{d}s \qquad \forall v \in H^1(\mathcal{T}_h). \tag{167}$$

Here, $\hat{u} : H^1(\mathcal{T}_h) \to T(\mathcal{T}_h)$ is a scalar numerical flux function, and $\hat{\sigma} : H^2(\mathcal{T}_h)\times[H^1(\mathcal{T}_h)]^d \to [T(\mathcal{T}_h)]^d$ is a vector-valued numerical flux function. Depending on the particular choice of the numerical flux functions $\hat{u}$ and $\hat{\sigma}$ several different DG discretizations can be derived, each with specific properties with respect to stability and accuracy.

Definition 5.1 *We say that the numerical fluxes $\hat{u}$ and $\hat{\sigma}$ are consistent if*

$$\hat{u}(v) = v, \qquad \hat{\sigma}(v,\nabla v) = \nabla v, \quad \text{on } \Gamma_{\mathcal{I}} \cup \Gamma, \tag{168}$$

whenever v is a smooth function satisfying the Dirichlet boundary conditions. Furthermore, we say that $\hat{u}$ and $\hat{\sigma}$ are conservative if they are single-valued *on $\Gamma_{\mathcal{I}} \cup \Gamma$.*

The term *conservative* comes from the following useful property, which holds whenever the vector flux $\hat{\sigma}$ is single-valued: If S is the union of any collection of elements, we obtain

$$\begin{aligned}\int_S \sigma \cdot \nabla_h v \,\mathrm{d}\mathbf{x} &= \int_S f v \,\mathrm{d}\mathbf{x} + \sum_{\kappa \subset S} \int_{\partial\kappa \setminus \partial S} \hat{\sigma} \cdot \mathbf{n}\, v \,\mathrm{d}s + \int_{\partial S} \hat{\sigma} \cdot \mathbf{n}\, v \,\mathrm{d}s \\ &= \int_S f v \,\mathrm{d}\mathbf{x} + \sum_{e \subset \Gamma_{\mathcal{I}_S}} \int_e \hat{\sigma} \cdot \mathbf{n}^+ \,(v^+ - v^-) \,\mathrm{d}s + \int_{\partial S} \hat{\sigma} \cdot \mathbf{n}\, v \,\mathrm{d}s,\end{aligned} \tag{169}$$

for $v \equiv 0$ on $\Omega \setminus S$, where $\Gamma_{\mathcal{I}_S}$ are the faces interior to S. Taking $v \equiv 1$ on S we then have

$$\int_S f v \,\mathrm{d}\mathbf{x} + \int_{\partial S} \hat{\sigma} \cdot \mathbf{n}\, v \,\mathrm{d}s = 0, \tag{170}$$

i.e. a global and local conservation property similar to conservation properties of the DG discretization of the linear advection in Section 4.7.

Equations (166) and (167) represent a first order system in u and σ with $(d+1)$ as many unknowns as the original (scalar) problem in u. In order to reduce the problem size, the auxiliary variable σ is usually eliminated to gain a so-called *primal formulation* involving only the primal variable u. To this end, we perform a second integration by parts on each element κ in (166) and set $\tau = \nabla_h v$ which gives us

$$\int_\Omega \sigma \cdot \nabla_h v \,\mathrm{d}\mathbf{x} = \int_\Omega \nabla_h u \cdot \nabla_h v \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} (\hat{u} - u)\mathbf{n} \cdot \nabla_h v \,\mathrm{d}s. \tag{171}$$

Substituting (171) into (167) we obtain the *primal flux formulation*: find $u \in H^2(\mathcal{T}_h)$ such that

$$\hat{B}_h(u, v) = \int_\Omega f v \,\mathrm{d}\mathbf{x} \quad \forall v \in H^2(\mathcal{T}_h),$$

where the bilinear form $\hat{B}_h(\cdot,\cdot) : H^2(\mathcal{T}_h) \times H^2(\mathcal{T}_h) \to \mathbb{R}$ is defined by

$$\hat{B}_h(u, v) = \int_\Omega \nabla_h u \cdot \nabla_h v \,\mathrm{d}\mathbf{x} - \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} \hat{\sigma} \cdot \mathbf{n} v \,\mathrm{d}s + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} (\hat{u} - u)\mathbf{n} \cdot \nabla_h v \,\mathrm{d}s. \tag{172}$$

This bilinear form is denoted by $\hat{B}_h$ (and not B_h) as it includes the (still unspecified) numerical fluxes $\hat{u}$ and $\hat{\sigma}$. Furthermore, $\hat{B}_h$ includes – through the specification of $\hat{u}$ and $\hat{\sigma}$ on the boundary — all boundary data terms.

Finally, we note that $\hat{B}_h$ in (172) is an *element-based form*, i.e. it is given in terms of $\sum_\kappa \int_{\partial\kappa}$. This means that each interior face $e = \Gamma_{\mathcal{I}}$ occurs twice in the sum over all elements κ (once in $\int_{\partial\kappa}$ and once in $\int_{\partial\kappa'}$ for $\kappa' \neq \kappa$ and $e = \partial\kappa \cap \partial\kappa' \neq 0$). In the following, we transfer the element-based form into a *face-based form*, i.e. we rewrite $\hat{B}_h$ in terms of $\int_{\Gamma_{\mathcal{I}}}$ where each interior face occurs only once. However, before doing so, we introduce some more notation.

Definition 5.2 *Let $e \in \Gamma_{\mathcal{I}}$ be an interior edge between two adjacent elements κ^+ and κ^- with unit outward normal vectors, $\mathbf{n}^+, \mathbf{n}^- \in \mathbb{R}^d$, respectively. Let $q \in T(\mathcal{T}_h)$ and $\boldsymbol{\phi} \in [T(\mathcal{T}_h)]^d$ be the traces of a scalar and a vector valued function, respectively. Then, we define the mean value and jump operators, $\{\!\{\cdot\}\!\}$ and $[\![\cdot]\!]$, as follows*

$$\begin{aligned} \{\!\{q\}\!\} &= \tfrac{1}{2}(q^+ + q^-), & [\![q]\!] &= q^+\mathbf{n}^+ + q^-\mathbf{n}^-, \\ \{\!\{\boldsymbol{\phi}\}\!\} &= \tfrac{1}{2}(\boldsymbol{\phi}^+ + \boldsymbol{\phi}^-), & [\![\boldsymbol{\phi}]\!] &= \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ + \boldsymbol{\phi}^- \cdot \mathbf{n}^-. \end{aligned}$$

Definition 5.3 *On boundary edges $e \in \Gamma$ the mean value and jump operators are defined by*

$$\begin{aligned} \{\!\{q\}\!\} &= q^+, & [\![q]\!] &= q^+ \mathbf{n}^+, \\ \{\!\{\boldsymbol{\phi}\}\!\} &= \boldsymbol{\phi}^+, & [\![\boldsymbol{\phi}]\!] &= \boldsymbol{\phi}^+ \cdot \mathbf{n}^+. \end{aligned}$$

Based on these notations, we can show following result which will frequently be used to transfer between element-based and face-based forms.

Lemma 5.4 *Again, let $q \in T(\mathcal{T}_h)$ and $\boldsymbol{\phi} \in [T(\mathcal{T}_h)]^d$, then*

$$\sum_\kappa \int_{\partial\kappa\setminus\Gamma} \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+ \, ds = \int_{\Gamma_\mathcal{I}} \{\!\{\boldsymbol{\phi}\}\!\} \cdot [\![q]\!] \, ds + \int_{\Gamma_\mathcal{I}} [\![\boldsymbol{\phi}]\!] \, \{\!\{q\}\!\} \, ds, \tag{173}$$

$$\sum_\kappa \int_{\partial\kappa} \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+ \, ds = \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\boldsymbol{\phi}\}\!\} \cdot [\![q]\!] \, ds + \int_{\Gamma_\mathcal{I}} [\![\boldsymbol{\phi}]\!] \, \{\!\{q\}\!\} \, ds. \tag{174}$$

Proof: On $\Gamma_\mathcal{I}$ we have

$$\begin{aligned} \{\!\{\boldsymbol{\phi}\}\!\} \cdot [\![q]\!] + [\![\boldsymbol{\phi}]\!] \, \{\!\{q\}\!\} =& \tfrac{1}{2}(\boldsymbol{\phi}^+ + \boldsymbol{\phi}^-) \cdot (q^+\mathbf{n}^+ + q^-\mathbf{n}^-) + \\ & \tfrac{1}{2}(\boldsymbol{\phi}^+ \cdot \mathbf{n}^+ + \boldsymbol{\phi}^- \cdot \mathbf{n}^-)(q^+ + q^-) \\ =& \tfrac{1}{2}(\boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+ + \boldsymbol{\phi}^- \cdot \mathbf{n}^+ \, q^+ + \boldsymbol{\phi}^+ \cdot \mathbf{n}^- \, q^- + \boldsymbol{\phi}^- \cdot \mathbf{n}^- \, q^-) + \\ & \tfrac{1}{2}(\boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+ + \boldsymbol{\phi}^- \cdot \mathbf{n}^- \, q^+ + \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^- + \boldsymbol{\phi}^- \cdot \mathbf{n}^- \, q^-) \\ =& \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+ + \boldsymbol{\phi}^- \cdot \mathbf{n}^- \, q^- \end{aligned}$$

using $\mathbf{n}^- = -\mathbf{n}^+$ in the last identity. On Γ we have $\{\!\{\boldsymbol{\phi}\}\!\} \cdot [\![q]\!] = \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+$. □

Using (174) and the Gauss integral formula we obtain following result.

Corollary 5.5 *Let $v \in H^1(\mathcal{T}_h)$ and $\tau \in [H^1(\mathcal{T}_h)]^d$, then*

$$\int_\Omega \tau \cdot \nabla_h v \, d\mathbf{x} = -\int_\Omega \nabla_h \cdot \tau \, v \, d\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\tau\}\!\} \cdot [\![v]\!] \, ds + \int_{\Gamma_\mathcal{I}} [\![\tau]\!] \, \{\!\{v\}\!\} \, ds. \tag{175}$$

Proof: Using the Gauss integral formula on each $\kappa \in \mathcal{T}_h$,

$$\int_\kappa \nabla \cdot \psi \, \mathrm{d}\mathbf{x} = \int_{\partial\kappa} \psi \cdot \mathbf{n} \, \mathrm{d}s, \tag{176}$$

for $\psi := \tau v \in [H^1(\mathcal{T}_h)]^d$, and summing over all $\kappa \in \mathcal{T}_h$ we obtain

$$\int_\Omega \nabla_h \cdot \tau v \, \mathrm{d}\mathbf{x} + \int_\Omega \tau \cdot \nabla_h v \, \mathrm{d}\mathbf{x} = \sum_\kappa \int_{\partial\kappa} \tau \cdot \mathbf{n} \, v \, \mathrm{d}s = \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\tau\}\!\} \cdot [\![v]\!] \, \mathrm{d}s + \int_{\Gamma_\mathcal{I}} [\![\tau]\!] \, \{\!\{v\}\!\} \, \mathrm{d}s, \tag{177}$$

which shows (175). □

We now proceed in transferring the element-based form (172) into a face-based form. To this end, we use equation (174) twice (once for $\boldsymbol{\phi} = \hat{\sigma}$ and $q = v$, and once for $\boldsymbol{\phi} = \nabla_h v$ and $q = \hat{u} - u$), and rewrite (172) as follows

$$\begin{aligned} \hat{B}_h(u,v) = & \int_\Omega \nabla_h u \cdot \nabla_h v \, \mathrm{d}\mathbf{x} - \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\hat{\sigma}\}\!\} \cdot [\![v]\!] \, \mathrm{d}s - \int_{\Gamma_\mathcal{I}} [\![\hat{\sigma}]\!] \, \{\!\{v\}\!\} \, \mathrm{d}s \\ & + \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\nabla_h v\}\!\} \cdot [\![\hat{u} - u]\!] \, \mathrm{d}s + \int_{\Gamma_\mathcal{I}} [\![\nabla_h v]\!] \, \{\!\{\hat{u} - u\}\!\} \, \mathrm{d}s, \end{aligned}$$

which results in following face-based *primal flux form*,

$$\begin{aligned} \hat{B}_h(u,v) = & \int_\Omega \nabla_h u \cdot \nabla_h v \, \mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} ([\![\hat{u} - u]\!] \cdot \{\!\{\nabla_h v\}\!\} - \{\!\{\hat{\sigma}\}\!\} \cdot [\![v]\!]) \, \mathrm{d}s \\ & + \int_{\Gamma_\mathcal{I}} (\{\!\{\hat{u} - u\}\!\} \, [\![\nabla_h v]\!] - [\![\hat{\sigma}]\!] \, \{\!\{v\}\!\}) \, \mathrm{d}s. \end{aligned} \tag{178}$$

5.2 The DG discretization: Consistency and adjoint consistency

Let $V_h := V_{h,p}^d \subset H^2(\mathcal{T}_h)$ be the space of discontinuous piecewise polynomials of degree p as defined in (123). Then the discontinuous Galerkin discretization in face-based flux formulation is given by: find $u_h \in V_h$ such that

$$\hat{B}_h(u_h, v_h) = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} \quad \forall v_h \in V_h. \tag{179}$$

where $\hat{B}_h$ is as defined in (178) and the fluxes $\hat{u}$ and $\hat{\sigma}$ are still unspecified.

Definition 5.6 (Consistency) *Let $u \in H^2(\Omega)$ be the exact solution to problem* (162)*. Then, the discretization* (179) *of* (162) *is* consistent *if and only if*

$$\hat{B}_h(u, v) = \int_\Omega f v \,\mathrm{d}\mathbf{x} \quad \forall v \in H^2(\mathcal{T}_h). \tag{180}$$

Theorem 5.7 *Let $\hat{B}_h(\cdot,\cdot)$ be given by* (178)*. Then the discretization* (179) *of the homogeneous Dirichlet problem* (163) *is* consistent *if and only if the* numerical fluxes *$\hat{u}$ and $\hat{\sigma}$ are consistent,*

$$\hat{u}(v) = v, \qquad \hat{\sigma}(v, \nabla v) = \nabla v \quad \text{on } \Gamma_\mathcal{I} \cup \Gamma, \tag{181}$$

Proof: Let u be the solution to problem (162). Setting $\tau = \nabla u$ in (175) we obtain

$$\int_\Omega \nabla u \cdot \nabla_h v \,\mathrm{d}\mathbf{x} = -\int_\Omega \Delta u v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\nabla u\}\!\} \cdot [\![v]\!] \,\mathrm{d}s + \int_{\Gamma_\mathcal{I}} [\![\nabla u]\!] \{\!\{v\}\!\} \,\mathrm{d}s, \tag{182}$$

for $v \in H^2(\mathcal{T}_h)$. Substituting this into $\hat{B}_h(u, v)$, cf. (178), gives

$$\begin{aligned}\hat{B}_h(u, v) = &-\int_\Omega \Delta_h u v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\nabla_h u\}\!\} \cdot [\![v]\!] \,\mathrm{d}s + \int_{\Gamma_\mathcal{I}} [\![\nabla_h u]\!] \{\!\{v\}\!\} \,\mathrm{d}s \\ &+ \int_{\Gamma_\mathcal{I}\cup\Gamma} \left([\![\hat{u} - u]\!] \cdot \{\!\{\nabla_h v\}\!\} - \{\!\{\hat{\sigma}\}\!\} \cdot [\![v]\!]\right) \mathrm{d}s + \int_{\Gamma_\mathcal{I}} \left(\{\!\{\hat{u} - u\}\!\} [\![\nabla_h v]\!] - [\![\hat{\sigma}]\!] \{\!\{v\}\!\}\right) \mathrm{d}s,\end{aligned}$$

where $\hat{u} = \hat{u}(u)$ and $\hat{\sigma} = \hat{\sigma}(u, \nabla u)$. Using $\{\!\{u\}\!\} = u$, $[\![u]\!] = 0$, $\{\!\{\nabla_h u\}\!\} = \nabla u$, $[\![\nabla_h u]\!] = 0$, and $-\Delta u = f$, we obtain

$$\begin{aligned}\hat{B}_h(u, v) = &\int_\Omega f v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} \left([\![\hat{u}]\!] \cdot \{\!\{\nabla_h v\}\!\} + (\nabla u - \{\!\{\hat{\sigma}\}\!\}) \cdot [\![v]\!]\right) \mathrm{d}s \\ &+ \int_{\Gamma_\mathcal{I}} \left((\{\!\{\hat{u}\}\!\} - u) [\![\nabla_h v]\!] - [\![\hat{\sigma}]\!] \{\!\{v\}\!\}\right) \mathrm{d}s.\end{aligned} \tag{183}$$

If the numerical flux $\hat{u}$ is consistent, i.e. $\hat{u}(u) = u$ on $\Gamma_\mathcal{I} \cup \Gamma$, then $[\![\hat{u}]\!] = 0$ and $\{\!\{\hat{u}\}\!\} = u$ on $\Gamma_\mathcal{I} \cup \Gamma$. If the numerical flux $\hat{\sigma}$ is also consistent, i.e. $\hat{\sigma}(u) = \nabla u$, then $[\![\hat{\sigma}]\!] = 0$ and $\{\!\{\hat{\sigma}\}\!\} = \nabla u$ on $\Gamma_\mathcal{I} \cup \Gamma$. Inserting these relations in (183) we obtain

$$\hat{B}_h(u, v) = \int_\Omega f v \,\mathrm{d}\mathbf{x}. \tag{184}$$

Hence, the primal formulation is consistent. This argument can easily be reversed. Assuming that the primal formulation is consistent, i.e. Equation (184) holds, then, in view of (183), we recognize that consistency (181) of the numerical fluxes is necessary. □

Remark 5.8 *Theorem 5.7 can be generalized to the Dirichlet-Neumann problem* (162)*: The discretization* (179) *of* (162) *is* consistent *if and only if*

$$\hat{u}(v) = v, \qquad \hat{\sigma}(v, \nabla v) = \nabla v \quad \text{on } \Gamma_\mathcal{I} \cup \Gamma_D, \qquad \mathbf{n} \cdot \hat{\sigma}(v, \nabla v) = \mathbf{n} \cdot \nabla v \quad \text{on } \Gamma_N. \tag{185}$$

Corollary 5.9 (Galerkin orthogonality) *Let* (179) *be a consistent discretization. Then, the error* $u - u_h$ *is orthogonal (with respect to* $\hat{B}_h(\cdot,\cdot)$*) to the discrete function space* V_h*, i.e.*

$$\hat{B}_h(u - u_h, v_h) = 0 \quad \forall v_h \in V_h. \tag{186}$$

Proof: Subtract (179) from (180) for $v := v_h \in V_h \subset H^2(\mathcal{T}_h)$. □

As we will see later, in addition to the consistency of a discretization also the so-called *adjoint consistency* of the discretization is of interest. Whereas optimal estimates in the energy norm (H^1-norm) can be derived for consistent discretizations, the derivation of optimal error estimates in the L^2-norm requires the application of a duality argument which requires the discretization to be adjoint consistent. In that sense, adjoint consistency represents an additional measure of quality of a discretization.

Definition 5.10 (Adjoint consistency) *Given a function* $j_\Omega \in L^2(\Omega)$*, let* $z \in H^2(\Omega)$ *be the exact solution to the dual or* adjoint problem

$$-\Delta z = j_\Omega \quad \text{in } \Omega, \qquad\qquad z = 0 \quad \text{on } \Gamma. \tag{187}$$

Then, the discretization (179) *of the homogeneous Dirichlet problem* (163) *is* adjoint consistent *if and only if*

$$\hat{B}_h(v, z) = \int_\Omega j_\Omega v \, d\mathbf{x} \quad \forall v \in H^2(\mathcal{T}_h). \tag{188}$$

Remark 5.11 *Related to the adjoint problem* (187) *we note that*

- *The differential operator in* (187) *is the adjoint operator* L^* *to the differential operator* L *of the primal problem. As the Laplace operator is self-adjoint, the adjoint problem to Poisson's equation is again Poisson's equation.*
- *The right hand side* j_Ω *in* (187) *may be any arbitrary (but fixed) function in* $L^2(\Omega)$*. Depending on the "purpose" of the adjoint problem the function* j_Ω *may be chosen appropriately. For example, when deriving* a priori *error estimates in* $L^2(\Omega)$ *we will use* $j_\Omega = e = u - u_h$ *which gives* $\|e\|^2 = \hat{B}_h(e, z)$*. An adjoint problem like in* (187) *will also be required when deriving error estimates with respect to target quantities of the form* $J(v) = \int_\Omega j_\Omega v \, d\mathbf{x}$ *which gives* $J(e) = \hat{B}_h(e, z)$*.*

Theorem 5.12 *Let* $\hat{B}_h(\cdot,\cdot)$ *be given by* (178)*. Then the discretization* (179) *of the homogeneous Dirichlet problem* (163) *is*

adjoint consistent

if and only if the numerical fluxes $\hat{u}$ *and* $\hat{\sigma}$ *are conservative, i.e.*

$$[\![\hat{u}]\!] = 0, \qquad \text{and} \qquad [\![\hat{\sigma}]\!] = 0. \tag{189}$$

Proof: Let $z \in H^2(\Omega)$ be the solution to (187), we set $\tau = \nabla z$ and $v = w$ in (175) and obtain

$$\int_\Omega \nabla z \cdot \nabla_h w \, \mathrm{d}\mathbf{x} = \int_\Omega j_\Omega w \, \mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I} \cup \Gamma} \{\!\{\nabla z\}\!\} \cdot [\![w]\!] \, \mathrm{d}s + \int_{\Gamma_\mathcal{I}} [\![\nabla z]\!] \, \{\!\{w\}\!\} \, \mathrm{d}s. \tag{190}$$

Substituting u by w and v by z in (178) and using (190) results in

$$\begin{aligned}\hat{B}_h(w, z) = &\int_\Omega j_\Omega w \, \mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I} \cup \Gamma} \{\!\{\nabla z\}\!\} \cdot [\![w]\!] \, \mathrm{d}s + \int_{\Gamma_\mathcal{I}} [\![\nabla z]\!] \, \{\!\{w\}\!\} \, \mathrm{d}s \\ &+ \int_{\Gamma_\mathcal{I} \cup \Gamma} \left([\![\hat{u}(w) - w]\!] \cdot \{\!\{\nabla z\}\!\} - \{\!\{\hat{\sigma}(w)\}\!\} \cdot [\![z]\!] \right) \mathrm{d}s \\ &+ \int_{\Gamma_\mathcal{I}} \left(\{\!\{\hat{u}(w) - w\}\!\} \, [\![\nabla z]\!] - [\![\hat{\sigma}(w)]\!] \, \{\!\{z\}\!\} \right) \mathrm{d}s,\end{aligned}$$

which reduces to following face-based dual form

$$\hat{B}_h(w,z) = \int_\Omega j_\Omega w \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} (\llbracket \hat{u}(w) \rrbracket \cdot \{\!\{\nabla z\}\!\} - \{\!\{\hat{\sigma}(w)\}\!\} \cdot \llbracket z \rrbracket) \,\mathrm{d}s + \int_{\Gamma_\mathcal{I}} (\{\!\{\hat{u}(w)\}\!\} \llbracket \nabla z \rrbracket - \llbracket \hat{\sigma}(w) \rrbracket \{\!\{z\}\!\}) \,\mathrm{d}s.$$

As $z \in H^2(\Omega)$ we have $\{\!\{z\}\!\} = z$, $\llbracket z \rrbracket = 0$, $\{\!\{\nabla z\}\!\} = \nabla z$ and $\llbracket \nabla z \rrbracket = 0$. Therefore,

$$\hat{B}_h(w,z) = \int_\Omega j_\Omega w \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} \llbracket \hat{u}(w) \rrbracket \cdot \nabla z \,\mathrm{d}s - \int_{\Gamma_\mathcal{I}} \llbracket \hat{\sigma}(w) \rrbracket \, z \,\mathrm{d}s, \tag{191}$$

which reduces to $\hat{B}_h(w,z) = \int_\Omega j_\Omega w \,\mathrm{d}\mathbf{x}$ if and only if (189) holds. □

Remark 5.13 *We note that Definition 5.10 gives the adjoint consistency property for the Dirichlet problem with homogeneous boundary conditions* (163)*. The extension of this to the Dirichlet-Neumann problem* (162) *is more involved and will be given in Section 6 within a general framework for analyzing consistency and adjoint consistency.*

5.3 Derivation of various DG discretization methods

We recall the DG discretization (179) together with (178): find $u_h \in V_h$ such that

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} (\llbracket \hat{u}_h - u_h \rrbracket \cdot \{\!\{\nabla_h v_h\}\!\} - \{\!\{\hat{\sigma}_h\}\!\} \cdot \llbracket v_h \rrbracket) \,\mathrm{d}s + \int_{\Gamma_\mathcal{I}} (\{\!\{\hat{u}_h - u_h\}\!\} \llbracket \nabla_h v_h \rrbracket - \llbracket \hat{\sigma}_h \rrbracket \{\!\{v_h\}\!\}) \,\mathrm{d}s = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} \quad \forall v_h \in V_h, \tag{192}$$

where $\hat{u}_h := \hat{u}(u_h)$ and $\hat{\sigma}_h := \hat{\sigma}(u_h, \nabla u_h)$. Here the numerical flux functions $\hat{u}$ and $\hat{\sigma}$ are still unspecified. (192) results in a consistent discretization of Problem (162) provided the fluxes $\hat{u}$ and $\hat{\sigma}$ are consistent, i.e. $\hat{u}(u) = u$ and $\hat{\sigma}(u, \nabla u) = \nabla u$. Depending on the specific choice of $\hat{u}$ and $\hat{\sigma}$ several different DG methods can be derived, each with specific properties with respect to stability and accuracy. Before continuing with the derivation of specific DG discretizations we first collect some elementary relations of the mean value and jump operators:

Lemma 5.14 *Let $\{\!\{\cdot\}\!\}$ and $\llbracket \cdot \rrbracket$ be the mean value and jump operators defined in Definition 5.2. Furthermore, let $q \in T(\mathcal{T}_h)$ and $\boldsymbol{\phi} \in [T(\mathcal{T}_h)]^d$, then*

$$\{\!\{\{\!\{q\}\!\}\}\!\} = \{\!\{q\}\!\}, \qquad \{\!\{\llbracket q \rrbracket\}\!\} = \llbracket q \rrbracket, \qquad \llbracket\llbracket q \rrbracket\rrbracket = 0, \qquad \llbracket\{\!\{q\}\!\}\rrbracket = 0, \tag{193}$$

$$\{\!\{\{\!\{\boldsymbol{\phi}\}\!\}\}\!\} = \{\!\{\boldsymbol{\phi}\}\!\}, \qquad \{\!\{\llbracket \boldsymbol{\phi} \rrbracket\}\!\} = \llbracket \boldsymbol{\phi} \rrbracket, \qquad \llbracket\llbracket \boldsymbol{\phi} \rrbracket\rrbracket = 0, \qquad \llbracket\{\!\{\boldsymbol{\phi}\}\!\}\rrbracket = 0. \tag{194}$$

5.3.1 The SIPG and NIPG methods and the method of Baumann-Oden

The symmetric interior penalty method (SIPG): Let the fluxes $\hat{u}_h$ and $\hat{\sigma}_h$ be given by

$$\hat{u}_h = \{\!\{u_h\}\!\}, \qquad \hat{\sigma}_h = \{\!\{\nabla_h u_h\}\!\} - \boldsymbol{\delta}^{\mathrm{ip}}(u_h) \qquad \text{on } \Gamma_\mathcal{I}, \tag{195}$$

$$\hat{u}_h = g_D, \qquad \hat{\sigma}_h = \nabla_h u_h - \boldsymbol{\delta}^{\mathrm{ip}}_\Gamma(u_h) \qquad \text{on } \Gamma_D, \tag{196}$$

$$\hat{u}_h = u_h, \qquad \hat{\sigma}_h = g_N \mathbf{n} \qquad \text{on } \Gamma_N, \tag{197}$$

where

$$\boldsymbol{\delta}^{\mathrm{ip}}(u_h) = \delta \llbracket u_h \rrbracket = C_{\mathrm{IP}} \frac{p^2}{h} \llbracket u_h \rrbracket \qquad \text{on } \Gamma_\mathcal{I}, \tag{198}$$

$$\boldsymbol{\delta}^{\mathrm{ip}}_\Gamma(u_h) = \delta (u_h - g_D)\mathbf{n} = C_{\mathrm{IP}} \frac{p^2}{h} (u_h - g_D)\mathbf{n} \qquad \text{on } \Gamma_D. \tag{199}$$

Using Lemma 5.14 we obtain $[\![\hat{u}_h]\!] = [\![\{\!\{u_h\}\!\}]\!] = 0$, $\{\!\{\hat{u}_h\}\!\} = \{\!\{\{\!\{u_h\}\!\}\!\}\!\} = \{\!\{u_h\}\!\}$, $\{\!\{\hat{\sigma}_h\}\!\} = \{\!\{\{\!\{\nabla_h u_h\}\!\}\!\} - \{\!\{\boldsymbol{\delta}^{\mathrm{ip}}(u_h)\}\!\} = \{\!\{\nabla_h u_h\}\!\} - \boldsymbol{\delta}^{\mathrm{ip}}(u_h)$, and $[\![\hat{\sigma}_h]\!] = [\![\{\!\{\nabla_h u_h\}\!\}]\!] - [\![\boldsymbol{\delta}^{\mathrm{ip}}(u_h)]\!] = 0$ on $\Gamma_{\mathcal{I}}$, and (192) reduces to the *symmetric interior penalty method* given by: find $u_h \in V_h$ such that

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} (-[\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\} - \{\!\{\nabla_h u_h\}\!\} \cdot [\![v_h]\!])\,\mathrm{d}s + \int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} \delta [\![u_h]\!] \cdot [\![v_h]\!]\,\mathrm{d}s = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_D} g_D\, \mathbf{n}\cdot\nabla_h v_h \,\mathrm{d}s + \int_{\Gamma_D} \delta g_D v_h \,\mathrm{d}s + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s \quad (200)$$

for all $v_h \in V_h$.

The non-symmetric interior penalty method (NIPG): Let the fluxes $\hat{u}_h$, $\hat{\sigma}_h$ be given by

$$\hat{u}_h = \{\!\{u_h\}\!\} + \mathbf{n}^+ \cdot [\![u_h]\!], \qquad \hat{\sigma}_h = \{\!\{\nabla_h u_h\}\!\} - \boldsymbol{\delta}^{\mathrm{ip}}(u_h) \qquad \text{on } \Gamma_{\mathcal{I}}, \quad (201)$$

$$\hat{u}_h = 2u_h - g_D, \qquad \hat{\sigma}_h = \nabla_h u_h - \boldsymbol{\delta}^{\mathrm{ip}}_{\Gamma}(u_h) \qquad \text{on } \Gamma_D, \quad (202)$$

and by (197). We use $\mathbf{n}^+ \cdot [\![u_h]\!] = \mathbf{n}^+ \cdot (u_h^+\mathbf{n}^+ + u_h^-\mathbf{n}^-) = u_h^+ - u_h^-$ and $[\![\{\!\{u_h\}\!\}]\!] = 0$, and obtain

$$[\![\hat{u}_h]\!] = [\![u_h^+ - u_h^-]\!] = (u_h^+ - u_h^-)\mathbf{n}^+ + (u_h^- - u_h^+)\mathbf{n}^- = 2(u_h^+\mathbf{n}^+ + u_h^-\mathbf{n}^-) = 2[\![u_h]\!],$$
$$\{\!\{\hat{u}_h\}\!\} = \{\!\{\{\!\{u_h\}\!\}\!\}\!\} + \{\!\{\mathbf{n}\cdot[\![u_h]\!]\}\!\} = \{\!\{u_h\}\!\} + \tfrac{1}{2}(u_h^+ - u_h^- + u_h^- - u_h^+)) = \{\!\{u_h\}\!\}.$$

Then, (192) reduces to the *non-symmetric interior penalty method*: find $u_h \in V_h$ such that

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} ([\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\} - \{\!\{\nabla_h u_h\}\!\} \cdot [\![v_h]\!])\,\mathrm{d}s + \int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} \delta [\![u_h]\!] \cdot [\![v_h]\!]\,\mathrm{d}s = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} g_D\, \mathbf{n}\cdot\nabla_h v_h \,\mathrm{d}s + \int_{\Gamma_D} \delta g_D v_h \,\mathrm{d}s + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s \quad (203)$$

for all $v_h \in V_h$. We note, that the only difference of this discretization to the SIPG discretization in (200) is the sign of the $\int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} [\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\}\,\mathrm{d}s$ and $\int_{\Gamma_D} g_D\, \mathbf{n}\cdot\nabla_h v_h \,\mathrm{d}s$ term.

The method of Baumann-Oden (BO): Let the fluxes $\hat{u}_h$ and $\hat{\sigma}_h$ be given by

$$\hat{u}_h = \{\!\{u_h\}\!\} + \mathbf{n}^+ \cdot [\![u_h]\!], \qquad \hat{\sigma}_h = \{\!\{\nabla_h u_h\}\!\} \qquad \text{on } \Gamma_{\mathcal{I}}, \quad (204)$$

$$\hat{u}_h = 2u_h - g_D, \qquad \hat{\sigma}_h = \nabla_h u_h \qquad \text{on } \Gamma_D, \quad (205)$$

and by (197). Then we obtain the method by Baumann and Oden: find $u_h \in V_h$ such that

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} ([\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\} - \{\!\{\nabla_h u_h\}\!\} \cdot [\![v_h]\!])\,\mathrm{d}s = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} g_D\, \mathbf{n}\cdot\nabla v_h \,\mathrm{d}s + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s \quad (206)$$

for all $v_h \in V_h$. We note, that this discretization can be obtained from the NIPG discretization in (203) simply by ignoring the interior penalty term $\boldsymbol{\delta}^{\mathrm{ip}}(u_h) = \delta[\![u_h]\!]$. However, we will show later, that the method of Baumann-Oden is unstable whereas the NIPG discretization is stable due to the stabilizing effect of the interior penalty term.

Unified description for SIPG, NIPG and Baumann-Oden: The discretizations derived above can be written in unified form as follows: find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_h, \tag{207}$$

where

$$\begin{aligned} B_h(u,v) &= \int_\Omega \nabla_h u \cdot \nabla_h v \,\mathrm{d}\mathbf{x} \\ &\quad + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} (\theta[\![u]\!]\cdot\{\!\{\nabla_h v\}\!\} - \{\!\{\nabla_h u\}\!\}\cdot[\![v]\!])\,\mathrm{d}s + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[\![u]\!]\cdot[\![v]\!]\,\mathrm{d}s, \\ F_h(v) &= \int_\Omega fv\,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} \theta g_D\,\mathbf{n}\cdot\nabla v\,\mathrm{d}s + \int_{\Gamma_D}\delta g_D v\,\mathrm{d}s + \int_{\Gamma_N} g_N v\,\mathrm{d}s, \end{aligned} \tag{208}$$

and the constants θ and δ are given by

$$\begin{array}{lll} \text{SIPG}: & \theta = -1, & \delta > 0, \\ \text{NIPG}: & \theta = 1, & \delta > 0, \\ \text{Baumann-Oden}: & \theta = 1, & \delta = 0. \end{array} \tag{209}$$

We note that in the primal form $\hat{B}_h$ as defined in (172) or (178) the numerical fluxes $\hat{u}$ and $\hat{\sigma}$ were still unspecified. Furthermore, $\hat{B}_h$ (implicitly) included all boundary data functions. In contrast to that B_h as given in (208) includes *no* boundary data. Instead, all boundary data terms have been moved to the right hand side and are now included in $F_h(v_h)$.

5.3.2 The original DG discretization of Bassi and Rebay (BR1)

Let us choose the fluxes $\hat{u}_h$ and $\hat{\sigma}_h$ to be given by

$$\hat{u}_h = \{\!\{u_h\}\!\}, \qquad \hat{\sigma}_h = \{\!\{\nabla_h u_h\}\!\} - \boldsymbol{\delta}^{\mathrm{br1}}(u_h) \qquad \text{on } \Gamma_\mathcal{I}, \tag{210}$$

$$\hat{u}_h = g_D, \qquad \hat{\sigma}_h = \nabla_h u_h - \boldsymbol{\delta}_\Gamma^{\mathrm{br1}}(u_h) \qquad \text{on } \Gamma_D, \tag{211}$$

$$\hat{u}_h = u_h, \qquad \hat{\sigma}_h = g_N\mathbf{n} \qquad \text{on } \Gamma_N, \tag{212}$$

with

$$\boldsymbol{\delta}^{\mathrm{br1}}(u_h) = \boldsymbol{\delta}_\Gamma^{\mathrm{br1}}(u_h) = -\{\!\{\mathbf{L}_{g_D}(u_h)\}\!\}, \tag{213}$$

where the so-called *global lifting operator* including Dirichlet boundary values,

$$\mathbf{L}_{g_D} : T(\mathcal{T}_h) \to \underline{\Sigma}_{h,p}^d := [V_{h,p}^d]^d \subset [H^1(\mathcal{T}_h)]^d,$$

is a vector-valued affine operator defined by: Let $\mathbf{L}_{g_D}(w) \in \underline{\Sigma}_{h,p}^d$ be the solution to

$$\int_\Omega \mathbf{L}_{g_D}(w)\cdot\tau\,\mathrm{d}\mathbf{x} = -\int_{\Gamma_\mathcal{I}} [\![w]\!]\cdot\{\!\{\tau\}\!\}\,\mathrm{d}s - \int_{\Gamma_D}(w-g_D)\mathbf{n}\cdot\tau\,\mathrm{d}s \quad \forall\tau\in\underline{\Sigma}_{h,p}^d. \tag{214}$$

Here, $T(\mathcal{T}_h) := L^2(\Gamma_\mathcal{I}\cup\Gamma)$ denotes the space of traces of functions $v \in H^1(\mathcal{T}_h)$. Furthermore, we consider the *global lifting operator* $\mathbf{L}_0$ [1] with homogeneous Dirichlet boundary values, which is the vector-valued linear operator given by: Let $\mathbf{L}_0(w) \in \underline{\Sigma}_{h,p}^d$ be the solution to

$$\int_\Omega \mathbf{L}_0(w)\cdot\tau\,\mathrm{d}\mathbf{x} = -\int_{\Gamma_\mathcal{I}\cup\Gamma} [\![w]\!]\cdot\{\!\{\tau\}\!\}\,\mathrm{d}s \quad \forall\tau\in\underline{\Sigma}_{h,p}^d. \tag{215}$$

[1] We note that in some publications, the global lifting operator is defined as $\mathbf{l}_0 : [T(\mathcal{T}_h)]^d \to \underline{\Sigma}_{h,p}^d$ with

$$\int_\Omega \mathbf{l}_0(\boldsymbol{\phi})\cdot\tau\,\mathrm{d}\mathbf{x} = -\int_{\Gamma_\mathcal{I}\cup\Gamma} \boldsymbol{\phi}\cdot\{\!\{\tau\}\!\}\,\mathrm{d}s \quad \forall\tau\in\underline{\Sigma}_{h,p}^d,$$

for which we then have $\mathbf{L}_0(w) = \mathbf{l}_0([\![w]\!])$.

In view of (214) and (215), we have

$$\int_\Omega \mathbf{L}_{g_D}(w) \cdot \tau \,\mathrm{d}\mathbf{x} = \int_\Omega \mathbf{L}_0(w) \cdot \tau \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} g_D \,\mathbf{n} \cdot \tau \,\mathrm{d}s. \tag{216}$$

Using the numerical fluxes $\hat{u}_h$ and $\hat{\sigma}_h$ as given in (210)-(212) the DG discretization (192) reduces to: find $u_h \in V_h$ such that

$$\begin{aligned}\int_\Omega \nabla_h u_h \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I} \cup \Gamma_D} (-[\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\} - \{\!\{\nabla_h u_h\}\!\} \cdot [\![v_h]\!]) \,\mathrm{d}s - \int_{\Gamma_\mathcal{I} \cup \Gamma_D} \{\!\{\mathbf{L}_{g_D}(u_h)\}\!\} \cdot [\![v_h]\!] \,\mathrm{d}s \\ = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_D} g_D \,\mathbf{n} \cdot \nabla_h v_h \,\mathrm{d}s + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s \qquad \forall v_h \in V_h. \end{aligned} \tag{217}$$

Using the definition (214) for $\mathbf{L}_{g_D}$ we can rewrite

$$\int_\Omega \mathbf{L}_{g_D}(u_h) \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} = -\int_{\Gamma_\mathcal{I} \cup \Gamma} [\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\} \,\mathrm{d}s + \int_{\Gamma_D} g_D \,\mathbf{n} \cdot \nabla_h v_h \,\mathrm{d}s. \tag{218}$$

Furthermore, using the relation (215) of $\mathbf{L}_0$ we can rewrite

$$\int_\Omega \nabla_h u_h \cdot \mathbf{L}_0(v_h) \,\mathrm{d}\mathbf{x} = -\int_{\Gamma_\mathcal{I} \cup \Gamma} \{\!\{\nabla_h u_h\}\!\} \cdot [\![v_h]\!] \,\mathrm{d}s, \tag{219}$$

$$\int_\Omega \mathbf{L}_{g_D}(u_h) \cdot \mathbf{L}_0(v_h) \,\mathrm{d}\mathbf{x} = -\int_{\Gamma_\mathcal{I} \cup \Gamma} \{\!\{\mathbf{L}_{g_D}(u_h)\}\!\} \cdot [\![v_h]\!] \,\mathrm{d}s. \tag{220}$$

Substituting these relations into (217) we obtain the discretization: find $u_h \in V_h$ such that

$$\int_\Omega (\nabla_h u_h + \mathbf{L}_{g_D}(u_h)) \cdot (\nabla_h v_h + \mathbf{L}_0(v_h)) \,\mathrm{d}\mathbf{x} = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s \qquad \forall v_h \in V_h, \tag{221}$$

or equivalently, using (216): find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F_h(v_h) \qquad \forall v_h \in V_h, \tag{222}$$

where

$$\begin{aligned} B_h(u, v) &= \int_\Omega (\nabla_h u + \mathbf{L}_0(u)) \cdot (\nabla_h v + \mathbf{L}_0(v)) \,\mathrm{d}\mathbf{x}, \\ F_h(v) &= \int_\Omega f v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N v \,\mathrm{d}s - \int_{\Gamma_D} g_D \,\mathbf{n} \cdot (\nabla_h v + \mathbf{L}_0(v)) \,\mathrm{d}s. \end{aligned} \tag{223}$$

This is the original method of Bassi and Rebay introduced in [5] for which, however, several problems have been observed: In contrast to most other DG discretizations where an element communicates with its direct neighboring elements only, the stencil of the BR1 discretization is considerably larger as it includes also neighbors of neighbors. Furthermore, this discretization is unstable. In fact, we obtain

$$B_h(v, v) = \|\nabla_h v + \mathbf{L}_0(v)\|^2_{L^2(\Omega)},$$

which vanishes on the set

$$Z := \{v \in V_h : \nabla_h v + \mathbf{L}_0(v) = 0\},$$

where $Z \setminus \{0\}$ can, in general, be nonempty. This discretization is called the *BR1 discretization* in order to distinguish it from the modification of Bassi and Rebay, the so-called *BR2 discretization*, which we will introduce in the following subsection.

5.3.3 The modified DG discretization of Bassi and Rebay (BR2)

Let us choose the fluxes $\hat{u}_h$ and $\hat{\sigma}_h$ to be given by

$$\hat{u}_h = \{\!\{u_h\}\!\}, \qquad \hat{\sigma}_h = \{\!\{\nabla_h u_h\}\!\} - \boldsymbol{\delta}^{\mathrm{br2}}(u_h) \qquad \text{on } \Gamma_{\mathcal{I}}, \tag{224}$$

$$\hat{u}_h = g_D, \qquad \hat{\sigma}_h = \nabla_h u_h - \boldsymbol{\delta}^{\mathrm{br2}}_{\Gamma}(u_h) \qquad \text{on } \Gamma_D, \tag{225}$$

$$\hat{u}_h = u_h, \qquad \hat{\sigma}_h = g_N \mathbf{n} \qquad \text{on } \Gamma_N, \tag{226}$$

with

$$\boldsymbol{\delta}^{\mathrm{br2}}(u_h) = \boldsymbol{\delta}^{\mathrm{br2}}_{\Gamma}(u_h) = -C_{\mathrm{BR2}} \{\!\{\mathbf{L}^e_{g_D}(u_h)\}\!\} \quad \text{for } e \subset \Gamma_{\mathcal{I}} \cup \Gamma_D, \tag{227}$$

where the so-called *local lifting operator* including Dirichlet boundary conditions, $\mathbf{L}^e_{g_D} : L^2(e) \to \underline{\Sigma}^d_{h,p}$, is a vector-valued affine operator defined by: $\mathbf{L}^e_{g_D}(w) \in \underline{\Sigma}^d_{h,p}$ is the solution to

$$\begin{aligned} \int_\Omega \mathbf{L}^e_{g_D}(w) \cdot \tau \,\mathrm{d}\mathbf{x} &= -\int_e (w - g_D)\, \mathbf{n} \cdot \tau \,\mathrm{d}s \quad \forall \tau \in \underline{\Sigma}^d_{h,p}, && \text{for } e \subset \Gamma_D \\ \int_\Omega \mathbf{L}^e_{g_D}(w) \cdot \tau \,\mathrm{d}\mathbf{x} &= -\int_e [\![w]\!] \cdot \{\!\{\tau\}\!\} \,\mathrm{d}s \quad \forall \tau \in \underline{\Sigma}^d_{h,p}, && \text{on } e \subset \Gamma_{\mathcal{I}}, \end{aligned} \tag{228}$$

and $\mathbf{L}^e_{g_D}(w)$ is defined to be zero for $e \subset \Gamma_N$. The local lifting operator with homogeneous Dirichlet boundary conditions $\mathbf{L}^e_0$ is defined accordingly. In particular, for $e \subset \Gamma_D$ we have

$$\int_\Omega \mathbf{L}^e_{g_D}(w) \cdot \tau \,\mathrm{d}\mathbf{x} = \int_\Omega \mathbf{L}^e_0(w) \cdot \tau \,\mathrm{d}\mathbf{x} + \int_e g_D\, \mathbf{n} \cdot \tau \,\mathrm{d}s \quad \forall \tau \in \underline{\Sigma}^d_{h,p}. \tag{229}$$

We note, that $\mathbf{L}^e_{g_D}(w)$ has support (i.e. is non-equal zero) only on the (one or two) elements sharing the edge e. Furthermore, $\mathbf{L}^e_{g_D}(w)$ does not depend on g_D on interior edges $e \subset \Gamma_{\mathcal{I}}$. Using the numerical fluxes $\hat{u}_h$ and $\hat{\sigma}_h$ as given in (224)-(226) the DG discretization (192) reduces to: find $u_h \in V_h$ such that

$$\begin{aligned} \int_\Omega \nabla_h u_h \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} \left(-[\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\} - \{\!\{\nabla_h u_h\}\!\} \cdot [\![v_h]\!]\right) \mathrm{d}s - \int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} C_{\mathrm{BR2}} \{\!\{\mathbf{L}^e_{g_D}(u_h)\}\!\} \cdot [\![v_h]\!] \,\mathrm{d}s \\ = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_D} g_D\, \mathbf{n} \cdot \nabla_h v_h \,\mathrm{d}s + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s \qquad \forall v_h \in V_h. \end{aligned} \tag{230}$$

Using the definition (228) of $\mathbf{L}^e_{g_D}$ we can rewrite

$$\sum_{e \subset \Gamma_{\mathcal{I}} \cup \Gamma_D} C_{\mathrm{BR2}} \int_\Omega \mathbf{L}^e_{g_D}(u_h) \cdot \mathbf{L}^e_0(v_h) \,\mathrm{d}\mathbf{x} = -\int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} C_{\mathrm{BR2}} \{\!\{\mathbf{L}^e_{g_D}(u_h)\}\!\} \cdot [\![v_h]\!] \,\mathrm{d}s. \tag{231}$$

Substituting this, (218) and (219) into (230) yields: find $u_h \in V_h$ such that

$$\begin{aligned} \int_\Omega \left(\nabla_h u_h \cdot \nabla_h v_h + \mathbf{L}_{g_D}(u_h) \cdot \nabla_h v_h + \nabla_h u_h \cdot \mathbf{L}_0(v_h)\right) \mathrm{d}\mathbf{x} \\ + \sum_{e \subset \Gamma_{\mathcal{I}} \cup \Gamma_D} C_{\mathrm{BR2}} \int_\Omega \mathbf{L}^e_{g_D}(u_h) \cdot \mathbf{L}^e_0(v_h) \,\mathrm{d}\mathbf{x} = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s \qquad \forall v_h \in V_h, \end{aligned} \tag{232}$$

or equivalently, using (216) and (229): find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F_h(v_h) \qquad \forall v_h \in V_h, \tag{233}$$

where

$$B_h(u,v) = \int_\Omega \left(\nabla_h u \cdot \nabla_h v + \mathbf{L}_0(u) \cdot \nabla_h v + \nabla_h u \cdot \mathbf{L}_0(v)\right) \,\mathrm{d}\mathbf{x} + \sum_{e \subset \Gamma_{\mathcal{I}} \cup \Gamma_D} C_{\mathrm{BR2}} \int_\Omega \mathbf{L}_0^e(u) \cdot \mathbf{L}_0^e(v) \,\mathrm{d}\mathbf{x},$$

$$F_h(v) = \int_\Omega f v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N v \,\mathrm{d}s - \int_{\Gamma_D} g_D \,\mathbf{n} \cdot \left(\nabla_h v + C_{\mathrm{BR2}} \mathbf{L}_0^e(v)\right) \,\mathrm{d}s. \tag{234}$$

We note, that (232) can be obtained also by replacing $\int_\Omega \mathbf{L}_{g_D}(u_h)\mathbf{L}_0(v_h)\,\mathrm{d}\mathbf{x}$ in (221) by (231). The BR2 discretization has several advantages over the BR1 scheme. The stencil of BR2 scheme includes only first neighbors instead of additional second neighbors as does the BR1 scheme. Furthermore, as we will show later, the BR2 discretization is stable, provided C_{BR2} is larger than the number of neighboring elements, i.e. $C_{\mathrm{BR2}} > 3$ for triangular elements and $C_{\mathrm{BR2}} > 4$ for quadrilateral elements. Finally, the BR2 discretization differs from the SIPG discretization in Subsection 5.3.1 only in the definition of the stabilization/penalization term $\boldsymbol{\delta}^{\mathrm{br2}}(u_h)$ versus $\boldsymbol{\delta}^{\mathrm{ip}}(u_h)$. We will show later, that the BR2 scheme has the advantage that a lower bound of the constant C_{BR2} is known (number of neighboring elements), whereas the constant C_{IP} occurring in the symmetric interior penalty term must be larger than a constant $C_{\mathrm{IP}}^0 > 0$ which in general is not known. However, due to the use of lifting operators the BR2 discretization is significantly more complicated and more computing time expensive than the SIPG discretization.

5.4 Consistency, adjoint consistency, continuity and coercivity

Corollary 5.15 (Consistency) *SIPG, NIPG, Baumann-Oden, BR1 and BR2 are consistent discretizations, i.e. the exact solution $u \in H^2(\Omega)$ to* (162) *satisfies*

$$B_h(u,v) = F_h(v) \quad \forall v \in H^2(\mathcal{T}_h). \tag{235}$$

Proof: We have $\hat{u}(u) = u$. Furthermore, we have $\boldsymbol{\delta}^{\mathrm{ip}}(u) = \boldsymbol{\delta}^{\mathrm{br1}}(u) = \boldsymbol{\delta}^{\mathrm{br2}}(u) = 0$ for $u \in H^2(\Omega)$ which is continuous. Thereby $\hat{\sigma}(u, \nabla u) = \nabla u$. Finally, we have $\mathbf{n} \cdot \hat{\sigma}(u, \nabla u) = g_N = \mathbf{n} \cdot \nabla u \quad \text{on } \Gamma_N$ and hence consistency by Theorem 5.7 and Remark 5.8. □

Corollary 5.16 (Galerkin orthogonality) *We have the Galerkin orthogonality*

$$B_h(u - u_h, v_h) = 0 \quad \forall v_h \in V_h. \tag{236}$$

Proof: Subtract (207) from (235) for $v_h \in V_h \subset H^2(\mathcal{T}_h)$. □

Corollary 5.17 (Adjoint consistency) *SIPG, BR1 and BR2 for the homogeneous Dirichlet problem* (163) *are adjoint consistent. NIPG and Baumann-Oden are adjoint* in*consistent.*

Proof: As shown in Subsection 5.3.1 we have $[\![\hat{\sigma}_h]\!] = 0$ and $[\![\hat{u}_h]\!] = 0$ for the SIPG method and similarly for BR1 and BR2. However, for NIPG and Baumann-Oden we have $[\![\hat{\sigma}_h]\!] = 0$ and $[\![\hat{u}_h]\!] = 2[\![u_h]\!]$. Adjoint (in)consistency now follows from Theorem 5.12. □

In the following we show that the bilinear form corresponding to the method by Baumann-Oden, and the bilinear forms of the SIPG, the NIPG and the BR2 methods are continuous.

Lemma 5.18 (Continuity of Baumann-Oden) *Let B_h be given by*

$$B_h(u,v) = \int_\Omega \nabla_h u \cdot \nabla_h v \,d\mathbf{x} + \int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} \left(\theta [\![u]\!] \cdot \{\!\{\nabla_h v\}\!\} - \{\!\{\nabla_h u\}\!\} \cdot [\![v]\!]\right) \,ds, \tag{237}$$

with $\theta = 1$. Then,

$$|B_h(u,v)| \le |||u|||_\delta \, |||v|||_\delta, \qquad \forall u, v \in H^2(\mathcal{T}_h), \tag{238}$$

for any $\delta > 0$, where

$$|||v|||_\delta^2 = \|\nabla_h v\|_{L^2(\Omega)}^2 + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1} \left(\mathbf{n}\cdot\{\!\{\nabla v\}\!\}\right)^2 \, ds + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[v]^2 \, ds. \tag{239}$$

Furthermore, (238) *holds also for $\theta = -1$.*

Proof: We have

$$|\int_\Omega \nabla_h u \cdot \nabla_h v \,\mathrm{d}\mathbf{x}| \le \|\nabla_h u\|_{L^2(\Omega)} \, \|\nabla_h v\|_{L^2(\Omega)}.$$

Furthermore, we use

$$\begin{aligned} |\int_{\Gamma_\mathcal{I}\cup\Gamma_D} [\![u]\!] \cdot \{\!\{\nabla_h v\}\!\} \,\mathrm{d}s| &= |\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{1/2}[u] \, \delta^{-1/2}\mathbf{n}\cdot\{\!\{\nabla_h v\}\!\} \,\mathrm{d}s| \\ &\le \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[u]^2 \,\mathrm{d}s\right)^{1/2} \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1} \left(\mathbf{n}\cdot\{\!\{\nabla_h v\}\!\}\right)^2 \,\mathrm{d}s\right)^{1/2}, \end{aligned} \tag{240}$$

and (240) with u and v exchanged, and obtain

$$\begin{aligned} |B_h(u,v)| &\le |\int_\Omega \nabla_h u \cdot \nabla_h v \,\mathrm{d}\mathbf{x}| + |\int_{\Gamma_\mathcal{I}\cup\Gamma_D} [\![u]\!] \cdot \{\!\{\nabla_h v\}\!\} \,\mathrm{d}s| + |\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \{\!\{\nabla_h u\}\!\} \cdot [\![v]\!] \,\mathrm{d}s| \\ &\le \|\nabla_h u\|_{L^2(\Omega)} \, \|\nabla_h v\|_{L^2(\Omega)} \\ &\quad + \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[u]^2 \,\mathrm{d}s\right)^{1/2} \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1} \left(\mathbf{n}\cdot\{\!\{\nabla_h v\}\!\}\right)^2 \,\mathrm{d}s\right)^{1/2} \\ &\quad + \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1} \left(\mathbf{n}\cdot\{\!\{\nabla_h u\}\!\}\right)^2 \,\mathrm{d}s\right)^{1/2} \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[v]^2 \,\mathrm{d}s\right)^{1/2} \\ &\le \left(\|\nabla_h u\|_{L^2(\Omega)}^2 + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1} \left(\mathbf{n}\cdot\{\!\{\nabla_h u\}\!\}\right)^2 \,\mathrm{d}s + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[u]^2 \,\mathrm{d}s\right)^{1/2} \\ &\quad \left(\|\nabla_h v\|_{L^2(\Omega)}^2 + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1} \left(\mathbf{n}\cdot\{\!\{\nabla_h v\}\!\}\right)^2 \,\mathrm{d}s + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[v]^2 \,\mathrm{d}s\right)^{1/2} \\ &\le |||u|||_\delta \, |||v|||_\delta, \end{aligned}$$

where we used the Cauchy-Schwarz inequality $\sum_i a_i b_i \le \left(\sum_i a_i^2\right)^{1/2} \left(\sum_i b_i^2\right)^{1/2}$. □

Lemma 5.19 (Continuity of NIPG and SIPG) *Let B_h be given as in* (208) *with $\delta > 0$ and $\theta = 1$ for NIPG and $\theta = -1$ for SIPG. Then there is a constant $1 < C \le 2$ such that*

$$|B_h(u,v)| \le C|||u|||_\delta \, |||v|||_\delta, \qquad \forall u, v \in H^2(\mathcal{T}_h), \tag{241}$$

where the norm $|||\cdot|||_\delta$ is as defined (239).

Proof: Using the Cauchy-Schwarz inequality $\int ab \le \left(\int a^2\right)^{1/2} \left(\int b^2\right)^{1/2}$ for $a = \delta^{1/2}[u]$ and $b = \delta^{1/2}[v]$ we obtain

$$|\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[u][v] \,\mathrm{d}s| \le \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[u]^2 \,\mathrm{d}s\right)^{1/2} \left(\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[v]^2 \,\mathrm{d}s\right)^{1/2}.$$

Thereby, using (238) for $\theta = 1$ and $\theta = -1$ we obtain

$$\begin{aligned} |B_h(u,v)| &\le |||u|||_\delta \, |||v|||_\delta + |\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[u][v] \,\mathrm{d}s| \\ &\le \left(|||u|||_\delta^2 + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[u]^2 \,\mathrm{d}s\right)^{1/2} \left(|||v|||_\delta^2 + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[v]^2 \,\mathrm{d}s\right)^{1/2} \\ &\le 2|||u|||_\delta \, |||v|||_\delta. \end{aligned}$$

and hence (241). □

We note that $\delta > 0$ in (238) may be any positive constant. In contrast to that $\delta > 0$ in (241) is the constant of the interior penalty term (198).

Lemma 5.20 (Continuity of BR2, [11]) *Let B_h be given as in* (234)*. Then there is a constant $C > 1$ such that*

$$|B_h(u,v)| \leq C|||u|||_{L_0^e}|||v|||_{L_0^e} \qquad \forall u, v \in H^2(\mathcal{T}_h), \tag{242}$$

where the $|||\cdot|||_{L_0^e}$-norm is given by

$$|||v|||_{L_0^e}^2 = \|\nabla_h v\|_{L^2(\Omega)}^2 + \sum_{e \subset \Gamma_{\mathcal{I}} \cup \Gamma} \|\mathbf{L}_0^e(v)\|_{L^2(\Omega)}^2. \tag{243}$$

Proof: First, we note that $\mathbf{L}_0(v) = \sum_{e\subset\partial\kappa} \mathbf{L}_0^e(v)$ on κ. Furthermore, since the support of each L_0^e is the union of (one or two) elements sharing the edge e, we have

$$\sum_{e \subset \Gamma_{\mathcal{I}} \cup \Gamma} \|\mathbf{L}_0^e(v)\|_{L^2(\Omega)}^2 = \sum_{\kappa \in \mathcal{T}_h} \sum_{e \subset \partial\kappa} \|\mathbf{L}_0^e(v)\|_{L^2(\kappa)}^2. \tag{244}$$

Thereby, using Cauchy-Schwarz inequality, $\left(\sum_{i=1}^N a_i\right)^2 \leq \sum_{i=1}^N 1 \sum_{i=1}^N a_i^2 = N \sum_{i=1}^N a_i^2$,

$$\|\mathbf{L}_0(v)\|_{L^2(\kappa)}^2 = \int_\kappa \left(\sum_{e\subset\partial\kappa} \mathbf{L}_0^e(v)\right)^2 \mathrm{d}\mathbf{x} \leq N \int_\kappa \sum_{e\subset\partial\kappa} (\mathbf{L}_0^e(v))^2 \mathrm{d}\mathbf{x} = N \sum_{e\subset\partial\kappa} \|\mathbf{L}_0^e(v)\|_{L^2(\kappa)}^2, \tag{245}$$

where N is the number of faces $e \subset \partial\kappa$ of an element κ. And, due to (244)

$$\|\mathbf{L}_0(v)\|_{L^2(\Omega)}^2 \leq N \sum_{\kappa\in\mathcal{T}_h} \sum_{e\subset\partial\kappa} \|\mathbf{L}_0^e(v)\|_{L^2(\kappa)}^2 = N \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma} \|\mathbf{L}_0^e(v)\|_{L^2(\Omega)}^2. \tag{246}$$

Given B_h as in (234), using Cauchy-Schwarz, (246) and again Cauchy-Schwarz we obtain

$$\begin{aligned}
B_h(u,v) =& \int_\Omega (\nabla_h u \cdot \nabla_h v + \mathbf{L}_0(u) \cdot \nabla_h v + \nabla_h u \cdot \mathbf{L}_0(v)) \, \mathrm{d}\mathbf{x} + \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma_D} C_{\mathrm{BR2}} \int_\Omega \mathbf{L}_0^e(u) \cdot \mathbf{L}_0^e(v) \, \mathrm{d}\mathbf{x} \\
\leq& \|\nabla_h u\|_{L^2(\Omega)} \|\nabla_h v\|_{L^2(\Omega)} + \|\mathbf{L}_0(u)\|_{L^2(\Omega)} \|\nabla_h v\|_{L^2(\Omega)} + \|\nabla_h u\|_{L^2(\Omega)} \|\mathbf{L}_0(v)\|_{L^2(\Omega)} \\
&+ \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma_D} C_{\mathrm{BR2}} \|\mathbf{L}_0^e(u)\|_{L^2(\Omega)} \|\mathbf{L}_0^e(v)\|_{L^2(\Omega)} \\
\leq& \left(2\|\nabla_h u\|_{L^2(\Omega)}^2 + \|\mathbf{L}_0(u)\|_{L^2(\Omega)}^2 + \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma_D} C_{\mathrm{BR2}} \|\mathbf{L}_0^e(u)\|_{L^2(\Omega)}^2\right) \\
&\times \left(2\|\nabla_h v\|_{L^2(\Omega)}^2 + \|\mathbf{L}_0(v)\|_{L^2(\Omega)}^2 + \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma_D} C_{\mathrm{BR2}} \|\mathbf{L}_0^e(u)\|_{L^2(\Omega)}^2\right) \leq C|||u|||_{L_0^e}|||v|||_{L_0^e}
\end{aligned}$$

with $C = (N + C_{\mathrm{BR2}})^2$. □

Based on the relations shown in the proof of Lemma 5.20 coercivity of BR2 is easily obtained.

Theorem 5.21 (Coercivity of BR2, [11]) *Let B_h be given as in* (234)*. Then there is a constant $C_{BR2}^0 > 0$ ($C_{BR2}^0 = 3$ on triangles, $C_{BR2}^0 = 4$ on quadrilaterals) such that for all $C_{BR2} > C_{BR2}^0$ we have following coercivity property: There is a constant $\gamma > 0$ such that*

$$B_h(v,v) \geq \gamma |||v|||_{L_0^e}^2 \qquad \forall v \in H^2(\mathcal{T}_h), \tag{247}$$

where the $|||\cdot|||_{L_0^e}$-norm is as defined in (243)*.*

Proof: Using $2ab \le \epsilon a^2 + \frac{1}{\epsilon} b^2$ and (245) we obtain

$$\begin{aligned} 2\int_\Omega \nabla_h v \cdot \mathbf{L}_0(v)\,\mathrm{d}\mathbf{x} &\le \epsilon \|\nabla_h v\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|\mathbf{L}_0(v)\|^2_{L^2(\Omega)} \\ &\le \epsilon \|\nabla_h v\|^2_{L^2(\Omega)} + \frac{N}{\epsilon} \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma_D} \|\mathbf{L}_0^e(v)\|^2_{L^2(\Omega)}, \end{aligned} \tag{248}$$

Then, using the definition (234) of B_h we have

$$\begin{aligned} B_h(v,v) &= \|\nabla_h v\|^2_{L^2(\Omega)} + 2\int_\Omega \mathbf{L}_0(v)\cdot\nabla_h v\,\mathrm{d}\mathbf{x} + \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma_D} C_{\mathrm{BR2}}\|\mathbf{L}_0^e(v)\|^2_{L^2(e)} \\ &\ge (1-\epsilon)\|\nabla_h v\|^2_{L^2(\Omega)} + \sum_{e\subset\Gamma_{\mathcal{I}}\cup\Gamma_D}\left(C_{\mathrm{BR2}} - \frac{N}{\epsilon}\right)\|\mathbf{L}_0^e(v)\|^2_{L^2(e)}, \end{aligned} \tag{249}$$

and hence (247) with $\gamma = \min(1-\epsilon, C_{\mathrm{BR2}} - \frac{N}{\epsilon})$ which is positive whenever $\frac{N}{C_{\mathrm{BR2}}} < \epsilon < 1$, i.e. whenever $C_{\mathrm{BR2}} > N$. Thereby, (247) holds provided $C_{\mathrm{BR2}} > C^0_{\mathrm{BR2}} := N$, where N is the number of faces $e \subset \partial\kappa$ of an element κ □

Also the coercivity of the method of Baumann-Oden is easily shown.

Lemma 5.22 (Coercivity of Baumann-Oden) *Let B_h be given as in* (208) *with $\theta = 1$ and $\delta = 0$. Then,*

$$B_h(v,v) = \|\nabla_h v\|^2_{L^2(\Omega)} \qquad \forall v \in H^2(\mathcal{T}_h). \tag{250}$$

Proof:

$$B_h(v,v) = \int_\Omega \nabla_h v\cdot\nabla_h v\,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} (\theta-1)\{\!\{\nabla_h v\}\!\}\cdot[\![v]\!]\,\mathrm{d}s = \|\nabla_h v\|^2_{L^2(\Omega)} \tag{251}$$

for $\theta = 1$. □

We see, that the bilinear form B_h for Baumann-Oden is coercive only with respect to the $H^1(\mathcal{T}_h)$-seminorm. In particular, for any $v_h \in V^d_{h,0}$ we have $B_h(v_h, v_h) = 0$, i.e. the method of Baumann-Oden is unstable. However, considering the discretization of $-\Delta u + cu = f$ with $c \ge c_0 > 0$ we obtain $B_h(v,v) \ge c_0\|v\|^2_{L^2(\Omega)} + \|\nabla_h v\|^2_{L^2(\Omega)}$, hence coercivity in the $H^1(\mathcal{T}_h)$-norm.

Finally, in order to show coercivity of the NIPG and SIPG discretization we first recall the following standard inverse estimate: There is a constant $C > 0$ such that

$$\|\nabla v_h\|_{L^2(\kappa)} \le C h_\kappa^{-1}\|v_h\|_{L^2(\kappa)} \quad \forall v_h \in V_h, \tag{252}$$

In the following we quote from [40], p. 208, a generalization of this estimate to $v_h \in V^d_{h,p}$.

Lemma 5.23 (Inverse estimate on $V^d_{h,p}$) *Let $\mathcal{T}_h$ be a shape regular mesh. Then, there is a constant $C \ge 0$ such that for any $\kappa \in \mathcal{T}_h$ we have*

$$\|\nabla v_h\|_{L^2(\kappa)} \le C\frac{p_\kappa^2}{h_\kappa}\|v_h\|_{L^2(\kappa)}, \qquad \forall v_h \in V^d_{h,p}. \tag{253}$$

Furthermore, let us recall following trace inequality:

Lemma 5.24 (Multiplicative trace inequality, [38]) *Let $\kappa \in \mathcal{T}_h$, with diameter h_κ and radius r_κ of an inscribed circle, with $ch_\kappa < r_\kappa < h_\kappa$, $c > 0$, then*

$$\|v\|^2_{L^2(\partial\kappa)} \le C\left(h_\kappa^{-1}\|v\|^2_{L^2(\kappa)} + \|v\|_{L^2(\kappa)}\|\nabla v\|_{L^2(\kappa)}\right) \qquad \forall v \in H^1(\kappa). \tag{254}$$

Theorem 5.25 (Coercivity of NIPG and SIPG, [38]) *Let B_h be given as in* (208) *with $\delta = C_{IP}\frac{p^2}{h}$. Then there is a constant $C_{IP}^0 \geq 0$ ($C_{IP}^0 = 0$ for NIPG, i.e. $\theta = 1$, and $C_{IP}^0 > 0$ for SIPG, i.e. $\theta = -1$), such that for all $C_{IP} > C_{IP}^0$ we have following coercivity property: There is a constant $\gamma > 0$ such that*

$$B_h(v_h, v_h) \geq \gamma |||v_h|||_\delta^2 \qquad \forall v_h \in V_{h,p}^d, \tag{255}$$

where

$$|||v|||_\delta^2 = \|\nabla_h v\|_{L^2(\Omega)}^2 + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1} \left(\mathbf{n}\cdot\{\!\{\nabla v\}\!\}\right)^2 \, ds + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta [v]^2 \, ds. \tag{256}$$

Proof: This proofs follows the proof given in [38]. We begin by considering the term $\int_e (\mathbf{n}\cdot\nabla v)^2 \, \mathrm{d}s$ on $e \in \Gamma_\mathcal{I}\cup\Gamma_D$. By employing the trace inequality (254) and the inverse estimate (253) we obtain for all $v_h \in V_{h,p}^d$,

$$\begin{aligned}
\int_e (\mathbf{n}\cdot\nabla v_h)^2 \, \mathrm{d}s &\leq C\left(h_\kappa^{-1}\|\nabla v_h\|_{L^2(\kappa)}^2 + \|\nabla v_h\|_{L^2(\kappa)}\|\nabla^2 v_h\|_{L^2(\kappa)}\right) \\
&\leq C\left(\frac{1}{h_\kappa} + \frac{p_\kappa^2}{h_\kappa}\right)\|\nabla v_h\|_{L^2(\kappa)}^2 \\
&\leq C\frac{p_\kappa^2}{h_\kappa}\|\nabla v_h\|_{L^2(\kappa)}^2 \leq \frac{C}{C_{\mathrm{IP}}}\delta\|\nabla v_h\|_{L^2(\kappa)}^2,
\end{aligned} \tag{257}$$

and hence

$$-\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1}\left(\mathbf{n}\cdot\{\!\{\nabla v_h\}\!\}\right)^2 \, \mathrm{d}s \geq -\frac{C}{C_{\mathrm{IP}}}\|\nabla_h v_h\|_{L^2(\Omega)}^2, \tag{258}$$

where we used $(a+b)^2 \leq 2(a^2+b^2)$ for $a = \mathbf{n}\cdot\nabla v_h^+$ and $b = \mathbf{n}\cdot\nabla v_h^-$. From (208) we have

$$B_h(v,v) = \int_\Omega \nabla_h v\cdot\nabla_h v \, \mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} (\theta-1)\{\!\{\nabla_h v\}\!\}\cdot[\![v]\!] \, \mathrm{d}s + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[v]^2 \, \mathrm{d}s. \tag{259}$$

For $\theta = 1$ the second term vanishes and we obtain using (258)

$$B_h(v_h,v_h) - \gamma|||v_h|||_\delta^2 \geq \left(1-\gamma-\gamma\frac{C}{C_{\mathrm{IP}}}\right)\|\nabla v_h\|_{L^2(\Omega)}^2 + (1-\gamma)\int_{\Gamma_\mathcal{I}\cup\Gamma_D}\delta[v]^2 \, \mathrm{d}s.$$

Thereby, for any $C_{\mathrm{IP}} > C_{\mathrm{IP}}^0 = 0$ we find a $0 < \gamma \leq 1/(1+C/C_{\mathrm{IP}})$ such that (255) holds.

For $\theta = -1$ the second term in (259) is bounded using $ab \leq \frac{\epsilon}{4}a^2 + \frac{1}{\epsilon}b^2$,

$$\begin{aligned}
2\int_e \{\!\{\nabla_h v\}\!\}\cdot[\![v]\!] \, \mathrm{d}s &\leq 2\int_e \mathbf{n}\cdot\{\!\{\nabla_h v\}\!\}[v] \, \mathrm{d}s \leq 2\left(\int_e \delta^{-1}\left(\mathbf{n}\cdot\{\!\{\nabla_h v\}\!\}\right)^2 \, \mathrm{d}s\right)^{1/2}\left(\int_e \delta[v]^2 \, \mathrm{d}s\right)^{1/2} \\
&\leq \frac{\epsilon}{4}4\int_e \delta^{-1}\left(\mathbf{n}\cdot\{\!\{\nabla_h v\}\!\}\right)^2 \, \mathrm{d}s + \frac{1}{\epsilon}\int_e \delta[v]^2 \, \mathrm{d}s,
\end{aligned}$$

and hence

$$-2\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \{\!\{\nabla_h v\}\!\}\cdot[\![v]\!] \, \mathrm{d}s \geq -\epsilon\int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta^{-1}\left(\mathbf{n}\cdot\{\!\{\nabla_h v\}\!\}\right)^2 \, \mathrm{d}s - \frac{1}{\epsilon}\int_{\Gamma_\mathcal{I}\cup\Gamma_D}\delta[v]^2 \, \mathrm{d}s. \tag{260}$$

We then obtain

$$B_h(v_h,v_h) - \gamma|||v_h|||_\delta^2 \geq \left(1-\gamma-(\gamma+\epsilon)\frac{C}{C_{\mathrm{IP}}}\right)\|\nabla v_h\|_{L^2(\Omega)}^2 + (1-\gamma-\frac{1}{\epsilon})\int_{\Gamma_\mathcal{I}\cup\Gamma_D}\delta[v]^2 \, \mathrm{d}s.$$

Hence, we require

$$1-\gamma-(\gamma+\epsilon)\frac{C}{C_{\mathrm{IP}}} \geq 0 \quad \text{and} \quad 1-\gamma-\frac{1}{\epsilon} \geq 0.$$

The second inequality is fulfilled if $0 < \gamma \leq 1 - 1/\epsilon$ and $\epsilon > 1$. On the other hand the first inequality requires that $1 - \epsilon C/C_{\text{IP}} \geq \gamma(1 + C/C_{\text{IP}})$ and hence

$$0 < \gamma \leq \frac{1 - \epsilon C/C_{\text{IP}}}{1 + C/C_{\text{IP}}} < \frac{1 - C/C_{\text{IP}}}{1 + C/C_{\text{IP}}} = \frac{C_{\text{IP}} - C}{C_{\text{IP}} + C},$$

for $C_{\text{IP}} > C_{\text{IP}}^0 > 0$, e.g. $C_{\text{IP}}^0 = C$ where C is the constant in (258). □

We emphasize that for the NIPG method any choice of the interior penalty constant $C_{\text{IP}} > 0$ gives a stable scheme. In contrast to that the SIPG method requires $C_{\text{IP}} > C_{\text{IP}}^0 > 0$ for stability with a constant C_{IP}^0 which is in general not known. However, numerical experiments showed that $C_{\text{IP}} = 10 - 20$ is a good choice a for large variety of problems.

We note that whereas continuity of B_h could be shown on $H^2(\mathcal{T}_h)$ coercivity of B_h on $H^2(\mathcal{T}_h)$ does not hold, see Prop. 4.4 in [42]. However, coercivity of B_h on the discrete function space $V_{h,p}^d$ as shown in Theorem 5.25 is sufficient for proving existence and uniqueness of the discrete solution $u_h \in V_{h,p}^d$. This has been discussed in more detail in Section 4.4 for the DG discretization of the linear advection equation.

Remark 5.26 *The estimates* (257) *and* (258) *motivate the particular choice of* $\delta = C_{IP}\frac{p^2}{h}$.

5.5 *A priori* error estimates

In this section we give an *a priori* error estimates for the NIPG and SIPG discretization.

Lemma 5.27 (A priori error estimate for NIPG and SIPG) *Let* $u \in H^{p+1}(\Omega)$ *be the exact solution to Poisson's equation* (162). *Furthermore, let* $u_h \in V_{h,p}^d$ *be the solution to*

$$B_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_{h,p}^d,$$

where B_h *is as given in* (208) *with* $\theta = 1$ *(NIPG) or* $\theta = -1$ *(SIPG) and* $\delta = C_{IP}\frac{p^2}{h}$, $C_{IP} > C_{IP}^0$, *cf. Theorem 5.25. Then*

$$|||u - u_h|||_\delta \leq Ch^p |u|_{H^{p+1}(\Omega)}, \tag{261}$$

where $|||\cdot|||_\delta$ *is the norm as defined in* (239).

Proof: Let the error $e = u - u_h$ be split as follows

$$e = u - u_h = (u - P_h u) - (u_h - P_h u) = \eta - \xi,$$

with $\eta = u - P_h u$ and $\xi = u_h - P_h u$. Here, $P_h := P_{h,p}^d$ is the L^2-projector onto $V_h := V_{h,p}^d$ given in Definition 4.14. Applying coercivity (255) of B_h for $\xi \in V_h$ we obtain

$$\gamma |||\xi|||_\delta^2 \leq B_h(\xi, \xi) = B_h(\eta - e, \xi) = B_h(\eta, \xi),$$

where we used Galerkin orthogonality (236). Using continuity of B_h, (241), we obtain

$$\gamma |||\xi|||_\delta^2 \leq B_h(\eta, \xi) \leq C |||\eta|||_\delta \, |||\xi|||_\delta.$$

In summary, we obtain

$$|||u - u_h|||_\delta \leq |||\eta|||_\delta + |||\xi|||_\delta \leq C |||\eta|||_\delta.$$

Thereby, employing (239) and the approximation estimates in Section 4.5 we obtain

$$\begin{aligned} |||u - u_h|||_\delta^2 \leq & C \left(\|\nabla_h \eta\|_{L^2(\Omega)}^2 + \int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} \delta^{-1} \left(\mathbf{n} \cdot \{\!\{\nabla \eta\}\!\} \right)^2 \, \mathrm{d}s + \int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} \delta [\eta]^2 \, \mathrm{d}s \right) \\ \leq & C \left(Ch^{2p} + \tfrac{1}{C_{\text{IP}}} \tfrac{h}{p^2} Ch^{2(p-1/2)} + Ch^{2(p+1/2)} \right) |u|_{H^{p+1}(\Omega)}^2, \end{aligned}$$

and thus (261). □

Remark 5.28 *We note that estimate* (261) *corresponds to the* H^1*-error estimate*

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^p |u|_{H^{p+1}(\Omega)}. \tag{262}$$

shown in Section 2 for the continuous *Galerkin discretization of Poisson's equation. Estimate* (261) *as well as* (262) *is of optimal order* p *which corresponds to the order of approximation of polynomials of degree* p *in the* H^1*-norm, cf. estimate* (133)*.*

Furthermore, we recall that for the *continuous* Galerkin discretization of Poisson's equation an *a priori* error estimate in the L^2-norm has been obtained via a duality argument (Aubin-Nitsche) which is based on the definition of an appropriate dual (or adjoint) problem. We will use this technique also for the interior penalty discontinuous Galerkin discretization. However, as we will see in the following, application of an duality argument requires an adjoint consistent discretization.

Lemma 5.29 (L^2-error estimates for NIPG and SIPG) *Let* $u \in H^{p+1}(\Omega)$ *be the exact solution to Poisson's equation* (162)*. Furthermore, let* $u_h \in V_{h,p}^d$ *be the solution to*

$$B_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_{h,p}^d,$$

where B_h *is as given in* (208) *with* $\theta = 1$ *(NIPG) or* $\theta = -1$ *(SIPG) and* $\delta = C_{IP}\frac{p^2}{h}$*,* $C_{IP} > C_{IP}^0$*, cf. Theorem 5.25. Then, for NIPG:*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^p |u|_{H^{p+1}(\Omega)}, \tag{263}$$

and for SIPG:

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{p+1} |u|_{H^{p+1}(\Omega)}. \tag{264}$$

Proof: For simplicity we assume homogeneous Dirichlet boundary conditions. We recall from the proof of Theorem 5.12 that given a function $j_\Omega \in L^2(\Omega)$, and the solution $z \in H^2(\Omega)$ to the *adjoint problem*

$$-\Delta z = j_\Omega \quad \text{in } \Omega, \qquad\qquad z = 0 \text{ on } \Gamma, \tag{265}$$

the bilinear form $\hat{B}_h$, see (178) with discretization specific numerical fluxes $\hat{u}_h$ and $\hat{\sigma}_h$, can be written as

$$\hat{B}_h(w, z) = \int_\Omega j_\Omega w \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I} \cup \Gamma} [\![\hat{u}(w)]\!] \cdot \nabla z \,\mathrm{d}s - \int_{\Gamma_\mathcal{I}} [\![\hat{\sigma}(w)]\!] z \,\mathrm{d}s, \tag{266}$$

see Equation (191). In particular, for the SIPG method and Dirichlet boundary conditions we have $[\![\hat{\sigma}_h]\!] = 0$ and $[\![\hat{u}_h]\!] = 0$, cf. proof to Corollary 5.17, and find

$$B_h^s(w, z) = \int_\Omega j_\Omega w \,\mathrm{d}\mathbf{x}, \tag{267}$$

where $B_h^s(\cdot, \cdot)$ denotes the bilinear form of the symmetric interior penalty DG discretizations given in (208) with $\theta = -1$. We define z^s to be the solution to (265) for $j_\Omega := e$. We assume that $z^s \in H^2(\Omega)$ and $\|z^s\|_{H^2(\Omega)} \leq C\|e\|_{L^2(\Omega)}$ which is satisfied if Ω is a convex polygon, for example. In view of (267) we have

$$B_h^s(w, z^s) = \int_\Omega ew \,\mathrm{d}\mathbf{x}. \tag{268}$$

Now choosing $w = e$ we obtain

$$\|e\|_{L^2(\Omega)}^2 = \int_\Omega e^2 \,\mathrm{d}\mathbf{x} = B_h^s(e, z^s) = B_h^s(e, z^s - z_h) \leq |||e|||_\delta \, |||z^s - z_h|||_\delta, \tag{269}$$

where we used Galerkin orthogonality for $z_h = P_{h,p}^d z^s \in V_{h,p}^d$ and continuity (241) of B_h. Thus using (261) and approximation estimates for $z^s - z_h$ we obtain

$$\|e\|_{L^2(\Omega)}^2 \leq |||e|||_\delta \, |||z^s - z_h|||_\delta \leq Ch^p |u|_{H^{p+1}(\Omega)} Ch|z|_{H^2(\Omega)} \leq Ch^{p+1}|u|_{H^{p+1}(\Omega)} \|e\|_{L^2(\Omega)},$$

and hence (264).

For the NIPG method the above argument fails because the method does not satisfy the adjoint consistency condition (188). In fact, for the NIPG method and homogeneous Dirichlet boundary conditions we have $[\![\hat{\sigma}_h]\!] = 0$ and $[\![\hat{u}_h]\!] = 2[\![u_h]\!]$, cf. proof to Corollary 5.17, and

$$B_h^n(w,z) = \int_\Omega j_\Omega w \,\mathrm{d}\mathbf{x} + 2\int_{\Gamma_\mathcal{I}\cup\Gamma} [\![w]\!] \cdot \nabla z \,\mathrm{d}s, \tag{270}$$

where $B_h^n(\cdot,\cdot)$ denotes the bilinear form of the symmetric interior penalty DG discretizations given in (208) with $\theta = 1$. Hence, the analytical solution z^n to

$$B_h^n(w,z^n) = \int_\Omega ew \,\mathrm{d}\mathbf{x}, \tag{271}$$

is mesh-dependent. Furthermore, z^n is not regular which is why we do not obtain an additional order of h from $z^n - z_h$ as we do in the case of the SIPG method. □

5.6 Numerical results

In the following we investigate the experimental order of convergence in the H^1- and the L^2-norm of the SIPG ($\theta = -1$) and the NIPG ($\theta = 1$) discretizations, see Section 5.3.1. According to Theorem 5.25 the penalization parameter is given by $\delta = C_{\mathrm{IP}} \frac{p^2}{h}$. In this example we choose $C_{\mathrm{IP}} = 4$.

Let us consider the following *model problem*: Let $\Omega = (0,1)^2$ and consider Poisson's equation (162) with forcing function f which is chosen so that the analytical solution to (162) is given by

$$u(\mathbf{x}) = \sin(\tfrac{1}{2}\pi x_1)\sin(\tfrac{1}{2}\pi x_2). \tag{272}$$

We impose Dirichlet boundary conditions where the boundary value function g_D on $\Gamma_D = \Gamma = \partial\Omega$ is prescribed based on the analytical solution u.

Figure 2 plots the error in the $H^1(\Omega)$-seminorm, $|u - u_h|_{H^1(\Omega)}$, against the number of elements. We see that for a given polynomial degree p the discretization errors of the SIPG and the NIPG discretization almost coincide. Furthermore, we see that for the discretizations with polynomial degree $p = 1,\ldots,5$, the discretization error in the H^1-seminorm is of order $\mathcal{O}(h^p)$ which is in agreement with the theoretical order of convergence, see Lemma 5.27.

Figure 3a) shows that the error in the $L^2(\Omega)$-norm of the SIPG discretization for the polynomial degrees $p = 1,\ldots,5$, is of order $\mathcal{O}(h^{p+1})$ which again is in perfect agreement with the theoretical result, see Lemma 5.29. In comparison to that, Figure 3b) shows the $L^2(\Omega)$-error of the NIPG discretization. Here, we see that the discretization behaves like $\mathcal{O}(h^{p+1})$ for odd p and like $\mathcal{O}(h^p)$ for even p. This sub-optimal convergence of the NIPG method is attributed to the lack of adjoint consistency and the resulting lack of smoothness of the adjoint solution, see Lemma 5.29. We note that similar results for a different test case have been obtained in [23].

Figures 2 and 3 show that there is a significant advantage of using higher order discretizations over using low order discretization methods. In fact, in Figure 2 we see that the discretization error in the $H^1(\Omega)$-seminorm for $p = 3$ on the coarsest mesh is of similar size as the error for $p = 1$ on the finest mesh. Similarly, in Figure 3a) the discretization error in the $L^2(\Omega)$-norm for $p = 4$ on the coarsest mesh is comparable to the error for $p = 1$ on the finest mesh. We emphasize that here the solutions are of similar accuracy although the finest mesh has by a factor of 16384 more elements

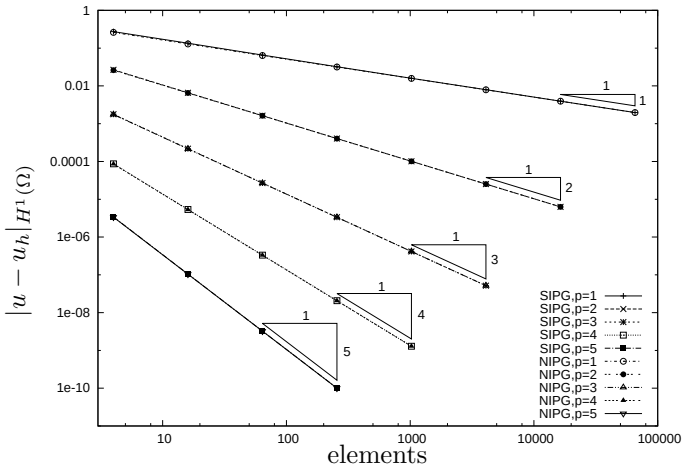
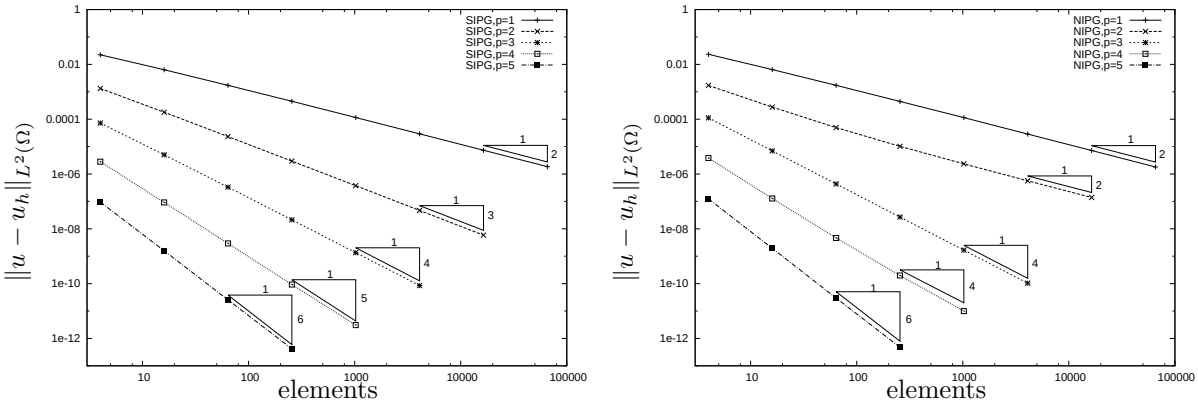
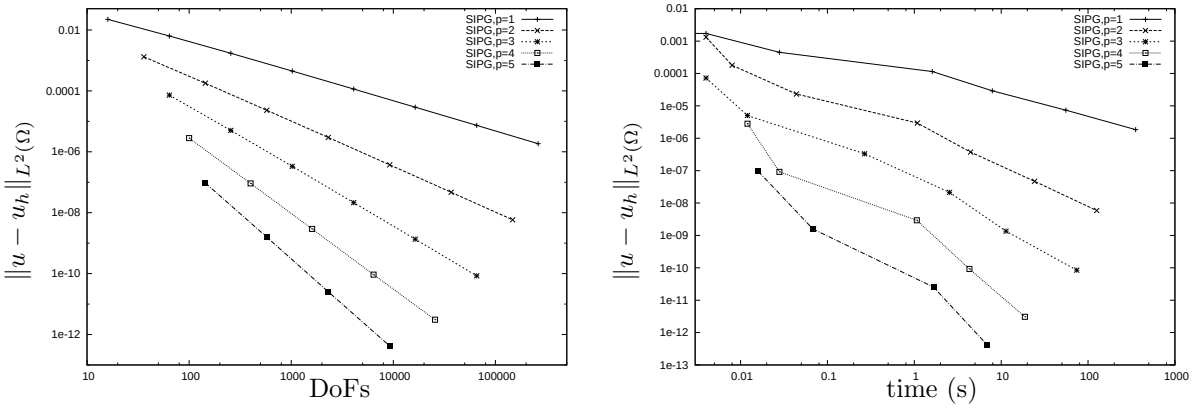


Figure 2: Model problem: The discretization error $|u - u_h|_{H^1(\Omega)}$ of the SIPG and NIPG methods with $p = 1, \ldots, 5$ is of order $\mathcal{O}(h^p)$.



(a) (b)

Figure 3: Model problem: Convergence of the discretization error $\|u - u_h\|_{L^2(\Omega)}$ for a) the SIPG and b) the NIPG methods with global mesh refinement.



(a) (b)

Figure 4: Model problem: The discretization error $\|u - u_h\|_{L^2(\Omega)}$ of the SIPG method plotted a) against the number of degrees of freedom (DoFs) and b) against the computing time in seconds.

than the coarsest mesh. Clearly, a discretization method of higher order requires more degrees of freedom (DoFs) per element, 25 DoFs/element for $p = 4$ in comparison to 4 DoFs/element for $p = 1$ in this case, but still the $p = 1$ discretization requires a factor of more than 2600 as many DoFs for the same accuracy as the $p = 4$ discretization. In more detail this is seen in Figure 4a) which plots the L^2-error against the number of DoFs. The large factor in the number of DoFs for the specific accuracy translates into a large factor in the computing time required. In fact, in Figure 4b) we see that the discretization with $p = 1$ on the finest mesh requires a by a factor of several thousands larger computing time for the same accuracy as the discretization with $p = 4$ on the coarsest mesh.

Admittedly, the model problem considered here is ideal in the sense that the geometry (unit square) and the governing equations (Poisson's equation) are particularly simple, also the solution is perfectly smooth. However, also for more complicated problems like aerodynamic flows, see e.g. Section 9.4, a significant gain of higher order methods over low order methods can be expected.

6 Consistency and adjoint consistency for linear problems

We recall that one of the most important properties of a discretization is its consistency with the differential equations to be discretized. In fact, consistency ensures that the "right" equations are solved. In finite element methods consistency directly implies the well-known Galerkin orthogonality. Provided the discretization is stable and using standard interpolation/approximation estimates this gives optimal *a priori* order estimates in the so-called energy norm, like e.g. the $\|\cdot\|_{H^1}$-norm for Poisson's equation.

Furthermore, we recall that in continuous as well as in discontinuous Galerkin finite element methods a duality argument has been applied for deriving error estimates in the L^2-norm. This approach introduces an appropriate adjoint (dual) problem, which is then used to represent the L^2-norm of the discretization error $e = u - u_h$ in terms of the discretization and the adjoint solution z. Again by Galerkin orthogonality and by using smoothness properties of the adjoint solution the L^2-error estimates are derived, see Theorem 2.17 for the continuous Galerkin discretization and Lemma 5.29 for the discontinuous Galerkin discretization of Poisson's equation.

Optimal order L^2-error estimates depend on the applicability of the duality argument as well as on the smoothness of the adjoint solution. Both, however, are connected to the so-called *adjoint consistency* of the discretization. As we have seen in Lemma 5.29 adjoint consistency of the SIPG results in optimal error estimates in the L^2-norm whereas the lack of adjoint consistency in the case of the NIPG discretization results in a suboptimal order of convergence in the L^2-norm.

In the following we introduce the adjoint consistency analysis following [27]. In particular, we revisit the consistency property and extend the definition of adjoint consistency for the homogeneous Dirichlet problem of Poisson's equation, see Definition 5.10, to the general case of linear problems with inhomogeneous boundary conditions. We then give a general framework for analyzing consistency and adjoint consistency and apply it to the interior penalty DG discretization of Poisson's equation and to the upwind DG discretization of the linear advection equation. The generalization of this analysis to nonlinear problems will be given in Section 8.5.

6.1 Definition of consistency and adjoint consistency

Let Ω be a bounded open domain in $\mathbb{R}^d$ with boundary Γ. Given the linear problem

$$Lu = f \quad \text{in } \Omega, \qquad\qquad Bu = g \quad \text{on } \Gamma, \tag{273}$$

where $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$, L denotes a linear differential operators on Ω, and B denotes a linear boundary operator on Γ. Let J be a linear target functional given by

$$J(u) = (j_\Omega, u)_\Omega + (j_\Gamma, Cu)_\Gamma \equiv \int_\Omega j_\Omega \, u \, \mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma \, Cu \, \mathrm{d}s, \tag{274}$$

where $j_\Omega \in L^2(\Omega)$, $j_\Gamma \in L^2(\Gamma)$, C is an operator on Γ which may be differential, and $(\cdot,\cdot)_\Omega$ and $(\cdot,\cdot)_\Gamma$ denote the $L^2(\Omega)$ and $L^2(\Gamma)$ scalar products, respectively. We assume that the target functional (274) is *compatible* with the primal problem (273), i.e. we assume that there are linear operators L^*, B^* and C^* such that following *compatibility condition* holds:

$$(Lu, z)_\Omega + (Bu, C^*z)_\Gamma = (u, L^*z)_\Omega + (Cu, B^*z)_\Gamma. \tag{275}$$

Then, L^*, B^* and C^* are the so-called *adjoint operators* to L, B and C, respectively. We note that for given operators L and B associated with the primal problem (273) only some target functionals (274) with operators C are compatible whereas others are not. However, *assuming* that (275) holds the adjoint problem associated to (273) and (274) is given by

$$L^*z = j_\Omega \quad \text{in } \Omega, \qquad\qquad B^*z = j_\Gamma \quad \text{on } \Gamma. \tag{276}$$

In an adjoint-based optimization framework, see e.g. [20], this ensures that

$$\begin{aligned} J(u) &= (u, j_\Omega)_\Omega + (Cu, j_\Gamma)_\Gamma = (u, L^*z)_\Omega + (Cu, B^*z)_\Gamma \\ &= (Lu, z)_\Omega + (Bu, C^*z)_\Gamma = (f, z)_\Omega + (g, C^*z)_\Gamma. \end{aligned} \tag{277}$$

Let Ω be subdivided into shape-regular meshes $\mathcal{T}_h = \{\kappa\}$ consisting of elements κ and let V_h be a discrete function space on $\mathcal{T}_h$. Furthermore, let problem (273) be discretized as follows: find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_h, \tag{278}$$

where $B_h(\cdot,\cdot)$ is a bilinear form and $F_h(\cdot)$ a linear form including the prescribed primal force and boundary data functions f and g. Then the discretization (278) is said to be *consistent* if the exact solution $u \in V$ to the primal problem (273) satisfies:

$$B_h(u, v) = F_h(v) \quad \forall v \in V, \tag{279}$$

where V is a suitably chosen function space such that $u \in V$ and $V_h \subset V$. Furthermore, the discretization (278) is said to be *adjoint consistent* if the exact solution $z \in V$ to the adjoint problem (276) satisfies:

$$B_h(w, z) = J(w) \quad \forall w \in V. \tag{280}$$

In other words, a discretization is adjoint consistent if the discrete adjoint problem is a consistent discretization of the continuous adjoint problem.

6.2 The consistency and adjoint consistency analysis

Based on the definition of consistency and adjoint consistency in the previous subsection we now follow [27] and outline a framework for analyzing consistency and adjoint consistency of discontinuous Galerkin discretizations. This framework can also be used to find specific terms due to which some DG discretizations may not be adjoint consistent. In these cases the analysis gives some insight into how an adjoint inconsistent DG discretization together with a specific target functional could be modified to recover an adjoint consistent discretization.

Given a primal problem, a discontinuous Galerkin discretization of the problem and a target functional, the adjoint consistency analysis consists of the following steps:

- **Derivation of the continuous adjoint problem:** Let the primal problem be given by (273). Furthermore, assume that $J(\cdot)$ is a linear (or linearized) target functional as in (274) which is compatible with (273). Then, we derive the continuous adjoint problem with continuous adjoint boundary conditions as given in (276).

- **Consistency analysis of the discrete primal problem:** We rewrite the discontinuous Galerkin discretization (278) of problem (273) in following element-based primal residual form: find $u_h \in V_h$ such that

$$\int_\Omega R(u_h)v_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} r(u_h)v_h + \boldsymbol{\rho}(u_h)\cdot\nabla_h v_h \,\mathrm{d}s + \int_\Gamma r_\Gamma(u_h)v_h + \boldsymbol{\rho}_\Gamma(u_h)\cdot\nabla_h v_h \,\mathrm{d}s = 0 \quad \forall v_h \in V_h, \tag{281}$$

  where $R(u_h)$ denotes the element residual, $r(u_h)$ and $\boldsymbol{\rho}(u_h)$ denote the interior face residuals, and $r_\Gamma(u_h)$ and $\boldsymbol{\rho}_\Gamma(u_h)$ denote the boundary residuals. We note, that this is a generalization of (159) to include also $\nabla_h v_h$ terms. According to (279) the discretization (278) is consistent if the exact solution u to (273) satisfies

$$\int_\Omega R(u)v \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} r(u)v + \boldsymbol{\rho}(u)\cdot\nabla_h v \,\mathrm{d}s + \int_\Gamma r_\Gamma(u)v + \boldsymbol{\rho}_\Gamma(u)\cdot\nabla_h v \,\mathrm{d}s = 0 \quad \forall v \in V, \tag{282}$$

  which holds provided u satisfies

$$\begin{aligned} R(u) &= 0 && && \text{in } \kappa, \kappa\in\mathcal{T}_h, \\ r(u) &= 0, && \boldsymbol{\rho}(u) = 0 && \text{on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h, \\ r_\Gamma(u) &= 0, && \boldsymbol{\rho}_\Gamma(u) = 0 && \text{on } \Gamma. \end{aligned} \tag{283}$$

- **Adjoint consistency of element, interior face and boundary terms:** Given the discretization (278) and the target functional (274), we rewrite the discrete adjoint problem: find $z_h \in V_h$ such that

$$B_h(w_h, z_h) = J(w_h) \quad \forall w_h \in V_h, \tag{284}$$

  in following element-based adjoint residual form: find $z_h \in V_h$ such that

$$\int_\Omega w_h\, R^*(z_h) \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} w_h\, r^*(z_h) + \nabla w_h\cdot\boldsymbol{\rho}^*(z_h) \,\mathrm{d}s + \int_\Gamma w_h\, r_\Gamma^*(z_h) + \nabla w_h\cdot\boldsymbol{\rho}_\Gamma^*(z_h) \,\mathrm{d}s = 0 \quad \forall w_h \in V_h, \tag{285}$$

  where $R^*(z_h)$, $r^*(z_h)$, $\boldsymbol{\rho}^*(z_h)$, $r_\Gamma^*(z_h)$ and $\boldsymbol{\rho}_\Gamma^*(z_h)$ denote the element, interior face and boundary adjoint residuals, respectively. According to (280) the discretization (278) is adjoint consistent if the exact solution $z \in V$ to (276) satisfies

$$\int_\Omega w\, R^*(z) \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} w\, r^*(z) + \nabla w\cdot\boldsymbol{\rho}^*(z) \,\mathrm{d}s + \int_\Gamma w\, r_\Gamma^*(z) + \nabla w\cdot\boldsymbol{\rho}_\Gamma^*(z) \,\mathrm{d}s = 0 \quad \forall w \in V, \tag{286}$$

which holds provided z satisfies

$$\begin{aligned} R^*(z) &= 0 && && \text{in } \kappa, \kappa \in \mathcal{T}_h, \\ r^*(z) &= 0, && \boldsymbol{\rho}^*(z) = 0 && \text{on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \\ r^*_\Gamma(z) &= 0, && \boldsymbol{\rho}^*_\Gamma(z) = 0 && \text{on } \Gamma. \end{aligned} \tag{287}$$

We note that the adjoint problem and consequently the adjoint consistency of a discretization depends on the specific target functional $J(\cdot)$ under consideration. Given a target functional of the form (274), we see that $R^*(z)$ depends on $j_\Omega(\cdot)$, and $r^*_\Gamma(z)$ depends on $j_\Gamma(\cdot)$. In order to obtain an adjoint consistent discretization it might be necessary to modify the target functional as follows

$$\tilde{J}(u_h) = J(i(u_h)) + \int_\Gamma r_J(u_h)\,\mathrm{d}s, \tag{288}$$

where $i(\cdot)$ and $r_J(\cdot)$ are functions to be specified. A modification of a target functional is called *consistent* if $\tilde{J}(u) = J(u)$ holds for the exact solution u. Thereby, the modification in (288) is consistent if the exact solution u satisfies $i(u) = u$ and $r_J(u) = 0$. Although the true value of the target functional is unchanged, $\tilde{J}(u) = J(u)$, the computed value $J(u_h)$ of the target functional is modified, and more importantly for nonlinear $\tilde{J}$ functionals, $\tilde{J}'[u_h]$ differs from $J'[u_h]$. This modification can be used to recover an adjoint consistent discretization. We note, that (288) is not a unique choice of a consistent modification of $J(\cdot)$; other examples are $\tilde{J}(u_h) = J(u_h) + \int_\Omega R_J(u_h)\,\mathrm{d}\mathbf{x}$, with $R_J(u) = 0$; or $\tilde{J}(u_h) = m(J(u_h), J(i(u_h)))$ with $i(u) = u$ and $m(j,j) = j$. However the consistent modification as given in (288) will be sufficient for our purpose.

In the following subsections we will perform the complete consistency and adjoint consistency analysis as outlined above for the interior penalty discontinuous Galerkin discretization of the Dirichlet-Neumann problem (162) of Poisson's equation and for the upwind discontinuous Galerkin discretization of the linear advection equation.

6.3 Adjoint consistency analysis of the IP discretization

We consider the Dirichlet-Neumann boundary value problem (162) of Poisson's equation,

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D, \qquad \mathbf{n}\cdot\nabla u = g_N \quad \text{on } \Gamma_N, \tag{289}$$

where $f \in L^2(\Omega)$, $g_D \in L^2(\Gamma_D)$ and $g_N \in L^2(\Gamma_N)$ are given functions. We assume that Γ_D and Γ_N are disjoint subsets with union Γ. We also assume that Γ_D is nonempty.

6.3.1 The continuous adjoint problem to Poisson's equation

In order to derive the continuous adjoint problem, we multiply the left hand side of (289) by z and integrate twice by parts over the domain Ω. Thereby, we obtain

$$(-\Delta u, z)_\Omega = (\nabla u, \nabla z)_\Omega - (\mathbf{n}\cdot\nabla u, z)_\Gamma = (u, -\Delta z)_\Omega + (u, \mathbf{n}\cdot\nabla z)_\Gamma - (\mathbf{n}\cdot\nabla u, z)_\Gamma.$$

Splitting the boundary terms according to $\Gamma = \Gamma_D \cup \Gamma_N$ and shuffling terms we arrive at

$$(-\Delta u, z)_\Omega + (u, -\mathbf{n}\cdot\nabla z)_{\Gamma_D} + (\mathbf{n}\cdot\nabla u, z)_{\Gamma_N} = (u, -\Delta z)_\Omega + (\mathbf{n}\cdot\nabla u, -z)_{\Gamma_D} + (u, \mathbf{n}\cdot\nabla z)_{\Gamma_N}.$$

Comparing with the compatibility condition (275), we see that for $Lu = -\Delta u$ in Ω and

$$\begin{aligned} Bu &= u, && Cu = \mathbf{n}\cdot\nabla u && \text{on } \Gamma_D, \\ Bu &= \mathbf{n}\cdot\nabla u, && Cu = u && \text{on } \Gamma_N, \end{aligned}$$

the adjoint operators are given by $L^*z = -\Delta z$ on Ω and

$$\begin{aligned} B^*z &= -z, & C^*z &= -\mathbf{n}\cdot\nabla z & &\text{on } \Gamma_D,\\ B^*z &= \mathbf{n}\cdot\nabla z, & C^*z &= z & &\text{on } \Gamma_N. \end{aligned}$$

In particular, for

$$\begin{aligned} J(u) &= \int_\Omega j_\Omega\, u \,\mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma\, Cu \,\mathrm{d}s \\ &= \int_\Omega j_\Omega\, u \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} j_D\, \mathbf{n}\cdot\nabla u \,\mathrm{d}s + \int_{\Gamma_N} j_N\, u \,\mathrm{d}s, \end{aligned} \tag{290}$$

the continuous adjoint problem is given by

$$-\Delta z = j_\Omega \quad \text{in } \Omega, \tag{291}$$

subject to the boundary conditions

$$-z = j_D \quad \text{on } \Gamma_D, \qquad\qquad \mathbf{n}\cdot\nabla z = j_N \quad \text{on } \Gamma_N. \tag{292}$$

6.3.2 Primal residual form of the interior penalty DG discretization

We begin by recalling the unified form of the method by Baumann-Oden, of the symmetric and the non-symmetric interior penalty discontinuous Galerkin discretization of (289) as given in (207) and (208): find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_h, \tag{293}$$

where

$$\begin{aligned} B_h(u,v) &= \int_\Omega \nabla_h u \cdot \nabla_h v \,\mathrm{d}\mathbf{x} \\ &\quad + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \left(\theta[\![u]\!]\cdot\{\!\{\nabla_h v\}\!\} - \{\!\{\nabla_h u\}\!\}\cdot[\![v]\!]\right) \mathrm{d}s + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[\![u]\!]\cdot[\![v]\!]\,\mathrm{d}s, \\ F_h(v) &= \int_\Omega fv \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} \theta g_D\, \mathbf{n}\cdot\nabla v \,\mathrm{d}s + \int_{\Gamma_D} \delta g_D v \,\mathrm{d}s + \int_{\Gamma_N} g_N v \,\mathrm{d}s. \end{aligned} \tag{294}$$

For $\delta \neq 0$ and $\theta = -1$ this represents the symmetric, whereas for $\delta \neq 0$ and $\theta = 1$ the non-symmetric version of the interior penalty DG method. Furthermore, for $\delta = 0$ and $\theta = 1$ this scheme reduces to the method of Baumann-Oden. The discretization is given in *face-based* form, i.e. in terms of $\int_{\Gamma_\mathcal{I}}$. In order to rewrite this in the element-based primal residual form as given in (281) we first must rewrite $B_h(\cdot,\cdot)$ in (294) in *element-based* formulation, i.e. in terms of $\sum_\kappa \int_{\partial\kappa}$. However, before doing so we introduce some more notation: In addition to the jump operator $[\![\cdot]\!]$ defined in Definition (5.2) we define the jump operator $[\cdot]$.

Definition 6.1 *Let $e \in \Gamma_\mathcal{I}$ be an interior edge between two adjacent elements κ^+ and κ^-. Let $q \in T(\mathcal{T}_h)$ be the traces of a scalar. Then, we define the jump operator $[\cdot]$ by*

$$\begin{aligned} [q] &= q^+ - q^- \; on\; e \in \Gamma_\mathcal{I}, \\ [q] &= q^+ \qquad\; on\; e \in \Gamma, \end{aligned}$$

Remark 6.2 *The jump $[\cdot]$ has already been used in the definition of the generic flux in* (117).

Remark 6.3 *Note that $[\![q]\!] = [q]\mathbf{n}$ and $[\![q]\!]\cdot\mathbf{n} = [q]$ for all $q \in T(\mathcal{T}_h)$.*

Then we show following two lemmas which will later be used to transfer from face-based to element-based formulations.

Lemma 6.4 *Let $q, \phi \in T(\mathcal{T}_h)$ and $\psi \in H^1(\mathcal{T}_h)$, then*

$$\int_{\Gamma_{\mathcal{I}} \cup \Gamma} [\![q]\!] \cdot [\![\phi]\!] \, ds = \sum_{\kappa} \int_{\partial\kappa} [q]\phi \, ds, \tag{295}$$

$$\int_{\Gamma_{\mathcal{I}} \cup \Gamma} [\![q]\!] \cdot \{\!\{\nabla_h \psi\}\!\} \, ds = \sum_{\kappa} \int_{\partial\kappa} q\mathbf{n} \cdot \{\!\{\nabla_h \psi\}\!\} \, ds, \tag{296}$$

$$\int_{\Gamma_{\mathcal{I}} \cup \Gamma} [\![q]\!] \cdot \{\!\{\nabla_h \psi\}\!\} \, ds = \frac{1}{2} \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma} [\![q]\!] \cdot \nabla_h \psi \, ds + \int_{\Gamma} q\mathbf{n} \cdot \nabla_h \psi \, ds. \tag{297}$$

Proof: We have

$$\sum_{\kappa} \int_{\partial\kappa\setminus\Gamma} [q]\phi \, \mathrm{d}s = \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma} [\![q]\!] \cdot \mathbf{n}\phi \, \mathrm{d}s = \int_{\Gamma_{\mathcal{I}}} [\![q]\!] \cdot (\mathbf{n}^+\phi^+ + \mathbf{n}^-\phi^-) \, \mathrm{d}s = \int_{\Gamma_{\mathcal{I}}} [\![q]\!] \cdot [\![\phi]\!] \, \mathrm{d}s,$$

$$\sum_{\kappa} \int_{\partial\kappa\setminus\Gamma} q\mathbf{n} \cdot \{\!\{\nabla_h \psi\}\!\} \, \mathrm{d}s = \int_{\Gamma_{\mathcal{I}}} \left(q^+\mathbf{n}^+ + q^-\mathbf{n}^-\right) \cdot \{\!\{\nabla_h \psi\}\!\} \, \mathrm{d}s = \int_{\Gamma_{\mathcal{I}}} [\![q]\!] \cdot \{\!\{\nabla_h \psi\}\!\} \, \mathrm{d}s,$$

$$\frac{1}{2} \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma} [\![q]\!] \cdot \nabla_h \psi \, \mathrm{d}s = \int_{\Gamma_{\mathcal{I}}} [\![q]\!] \cdot \{\!\{\nabla_h \psi\}\!\} \, \mathrm{d}s.$$

and use the definitions of $[\![\cdot]\!]$ and $\{\!\{\cdot\}\!\}$ on Γ. □

Lemma 6.5 *Let $q \in T(\mathcal{T}_h)$ and $\psi \in H^1(\mathcal{T}_h)$, then*

$$q^+ \, \mathbf{n}^+ \cdot \nabla_h \psi^+ = q^+ \mathbf{n}^+ \cdot \{\!\{\nabla_h \psi\}\!\} + \frac{1}{2} q^+ \, [\![\nabla_h \psi]\!]. \tag{298}$$

Proof:

$$\begin{aligned} q^+ \, \mathbf{n}^+ \cdot \nabla_h \psi^+ &= \frac{1}{2}(q^+ \, \mathbf{n}^+ \cdot \nabla_h \psi^+ + q^+ \, \mathbf{n}^+ \cdot \nabla_h \psi^-) + \frac{1}{2}(q^+ \, \mathbf{n}^+ \cdot \nabla_h \psi^+ - q^+ \, \mathbf{n}^+ \cdot \nabla_h \psi^-) \\ &= q^+ \, \mathbf{n}^+ \cdot \{\!\{\nabla_h \psi\}\!\} + \frac{1}{2} q^+ \, [\![\nabla_h \psi]\!]. \end{aligned}$$

□

Using (297) for $q := u_h$ and $\psi := v_h$, using (296) for $\psi := u_h$ and $q := v_h$, and using (295) for $q := u_h$ and $\phi := v_h$, the bilinear form B_h in (294) can be rewritten as follows

$$\begin{aligned} B_h(u_h, v_h) \equiv &\int_{\Omega} \nabla_h u_h \cdot \nabla_h v_h \, \mathrm{d}\mathbf{x} + \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma} \frac{1}{2} \theta [\![u_h]\!] \cdot \nabla_h v_h \, \mathrm{d}s \\ &- \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma_N} \{\!\{\nabla_h u_h\}\!\} \cdot \mathbf{n} \, v_h \, \mathrm{d}s + \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma_N} \delta [u_h] v_h \, \mathrm{d}s + \int_{\Gamma_D} \theta u_h \mathbf{n} \cdot \nabla_h v_h \, \mathrm{d}s. \end{aligned} \tag{299}$$

Using integration by parts on each element κ and summing over all elements yields

$$\int_{\Omega} \nabla_h u_h \cdot \nabla_h v_h \, \mathrm{d}\mathbf{x} = -\int_{\Omega} \Delta_h u_h \, v_h \, \mathrm{d}\mathbf{x} + \sum_{\kappa} \int_{\partial\kappa} \nabla_h u_h \cdot \mathbf{n} \, v_h \, \mathrm{d}s. \tag{300}$$

Substituting this into (299) we obtain

$$\begin{aligned} B_h(u_h, v_h) \equiv &-\int_{\Omega} \Delta_h u_h v_h \, \mathrm{d}\mathbf{x} + \sum_{\kappa} \int_{\partial\kappa} \nabla_h u_h \cdot \mathbf{n} \, v_h \, \mathrm{d}s + \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma} \frac{1}{2} \theta [\![u_h]\!] \cdot \nabla_h v_h \, \mathrm{d}s \\ &- \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma_N} \{\!\{\nabla_h u_h\}\!\} \cdot \mathbf{n} \, v_h \, \mathrm{d}s + \sum_{\kappa} \int_{\partial\kappa\setminus\Gamma_N} \delta [u_h] v_h \, \mathrm{d}s + \int_{\Gamma_D} \theta u_h \mathbf{n} \cdot \nabla_h v_h \, \mathrm{d}s. \end{aligned}$$

This is simplified by using (298) for $q := v_h$ and $\psi := u_h$ on $\partial\kappa \setminus \Gamma$ and $\nabla_h u_h - \{\!\{\nabla_h u_h\}\!\} = 0$ on Γ_D. In summary, the discretization (293) and (294) can be rewritten as

$$\begin{aligned} B_h(u_h, v_h) \equiv & -\int_\Omega \Delta_h u_h v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} \mathbf{n}\cdot\nabla_h u_h v_h \,\mathrm{d}s + \sum_\kappa \int_{\partial\kappa\setminus\Gamma} \frac{1}{2}\theta[\![u_h]\!]\cdot\nabla_h v_h \,\mathrm{d}s \\ & + \sum_\kappa \int_{\partial\kappa\setminus\Gamma} \left(\frac{1}{2}[\![\nabla_h u_h]\!] + \delta[u_h]\right) v_h \,\mathrm{d}s + \int_{\Gamma_D} \theta u_h \mathbf{n}\cdot\nabla_h v_h \,\mathrm{d}s + \int_{\Gamma_D} \delta u_h v_h \,\mathrm{d}s \\ = & \int_\Omega f v_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} \theta g_D \,\mathbf{n}\cdot\nabla v_h \,\mathrm{d}s + \int_{\Gamma_D} \delta g_D v_h \,\mathrm{d}s + \int_{\Gamma_N} g_N v_h \,\mathrm{d}s. \end{aligned} \tag{301}$$

This can be expressed in element-based residual form as follows: find $u_h \in V_h$ such that

$$\begin{aligned} \int_\Omega R(u_h) v_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} r(u_h) v_h + \boldsymbol{\rho}(u_h)\cdot\nabla_h v_h \,\mathrm{d}s \\ + \int_\Gamma r_\Gamma(u_h) v_h + \boldsymbol{\rho}_\Gamma(u_h)\cdot\nabla_h v_h \,\mathrm{d}s = 0 \quad \forall v_h \in V_h, \end{aligned}$$

where the residuals are given by $R(u_h) = f + \Delta_h u_h$ on Ω, and

$$\begin{aligned} r(u_h) &= -\tfrac{1}{2}[\![\nabla_h u_h]\!] - \delta[u_h], & \boldsymbol{\rho}(u_h) &= -\tfrac{1}{2}\theta[\![u_h]\!] && \text{on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h, \\ r_\Gamma(u_h) &= \delta(g_D - u_h), & \boldsymbol{\rho}_\Gamma(u_h) &= \theta(g_D - u_h)\mathbf{n} && \text{on } \Gamma_D, \\ r_\Gamma(u_h) &= g_N - \mathbf{n}\cdot\nabla_h u_h, & \boldsymbol{\rho}_\Gamma(u_h) &= 0 && \text{on } \Gamma_N. \end{aligned} \tag{302}$$

In particular, we see that the exact solution $u \in H^2(\Omega)$ to Poisson's equation (162) satisfies $R(u) = 0$, $r(u) = 0$, $\boldsymbol{\rho}(u) = 0$, $r_\Gamma(u) = 0$ and $\boldsymbol{\rho}_\Gamma(u) = 0$. Thereby u satisfies the equation,

$$\int_\Omega R(u) v \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} r(u) v + \boldsymbol{\rho}(u)\cdot\nabla_h v \,\mathrm{d}s + \int_\Gamma r_\Gamma(u) v + \boldsymbol{\rho}_\Gamma(u)\cdot\nabla_h v \,\mathrm{d}s = 0,$$

for all $v \in V$ which is equivalent to u satisfying

$$B_h(u, v) = F_h(v) \qquad \forall v \in V,$$

i.e. the DG discretizations as given in (293) or (301) are consistent. We note that we have shown this property already in Corollary 5.15 which is based on the consistency of the numerical fluxes $\hat{u}(u)$, $\hat{\sigma}(u, \nabla u)$ in Theorem 5.7. In this subsection we now analyzed consistency using an alternative way which is based on the primal residual form of the discretization. Here, consistency can easily be checked based on primal residuals which do (or do not) vanish for the exact solution u of the underlying equations.

We end this section by noting that from (301) we see that the discrete solution u_h satisfies following problem in a weak sense

$$-\Delta u = f \quad \text{in } \Omega, \tag{303}$$

subject to inter-element conditions

$$\begin{aligned} \frac{1}{2}[\![\nabla_h u]\!] + \delta[u] = 0 \quad \text{on } \partial\kappa\setminus\Gamma, \\ [u] = 0 \quad \text{on } \partial\kappa\setminus\Gamma, \end{aligned} \tag{304}$$

and boundary conditions

$$\begin{aligned} u = g_D \quad \text{on } \partial\kappa\cap\Gamma_D, \\ \mathbf{n}\cdot\nabla_h u = g_N \quad \text{on } \partial\kappa\cap\Gamma_N. \end{aligned} \tag{305}$$

In fact, this is the mesh-dependent counterpart of the original equations (162) to be solved.

6.3.3 Adjoint residual form of the interior penalty DG discretization

Given the target functional defined in (290), the discrete adjoint problem (284) to the discretization (293) and (294) is given by: find $z_h \in V_h$ such that

$$\int_\Omega \nabla_h w_h \cdot \nabla_h z_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I} \cup \Gamma_D} (\theta [\![w_h]\!] \cdot \{\!\{ \nabla_h z_h \}\!\} - \{\!\{ \nabla_h w_h \}\!\} \cdot [\![z_h]\!] + \delta [\![w_h]\!] \cdot [\![z_h]\!]) \,\mathrm{d}s = J(w_h),$$

for all $w_h \in V_h$. Then, in element-based form, we have: find $z_h \in V_h$ such that

$$\int_\Omega \nabla_h w_h \cdot \nabla_h z_h \,\mathrm{d}\mathbf{x} + \sum_\kappa \int_{\partial\kappa \setminus \Gamma_N} w_h \left(\theta \mathbf{n} \cdot \{\!\{ \nabla_h z_h \}\!\} + \delta [z_h]\right) \mathrm{d}s \\ - \frac{1}{2} \sum_\kappa \int_{\partial\kappa \setminus \Gamma} \nabla_h w_h \cdot [\![z_h]\!] \,\mathrm{d}s - \int_{\Gamma_D} \nabla_h w_h \cdot \mathbf{n}\, z_h \,\mathrm{d}s = J(w_h), \quad (306)$$

for all $w_h \in V_h$. Using (300) with u_h and v_h replaced by z_h and w_h, and using (298) with q and ψ replaced by w_h and z_h yields

$$\begin{aligned} \int_\Omega \nabla_h z_h \cdot \nabla_h w_h \,\mathrm{d}\mathbf{x} &= -\int_\Omega \Delta_h z_h \, w_h \,\mathrm{d}\mathbf{x} + \sum_\kappa \int_{\partial\kappa} w_h \, \mathbf{n} \cdot \nabla_h z_h \,\mathrm{d}s \\ &= -\int_\Omega w_h \, \Delta_h z_h \,\mathrm{d}\mathbf{x} + \sum_\kappa \int_{\partial\kappa} \left(w_h \, \mathbf{n} \cdot \{\!\{ \nabla_h z_h \}\!\} + \frac{1}{2} w_h \, [\![\nabla_h z_h]\!] \right) \mathrm{d}s. \end{aligned}$$

Substituting this into (306) we obtain

$$-\int_\Omega w_h \Delta_h z_h \,\mathrm{d}x + \sum_\kappa \int_{\partial\kappa \setminus \Gamma} w_h \left(\frac{1}{2} [\![\nabla_h z_h]\!] + (1+\theta) \mathbf{n} \cdot \{\!\{ \nabla_h z_h \}\!\} + \delta [z_h] \right) \mathrm{d}s \\ - \frac{1}{2} \sum_\kappa \int_{\partial\kappa \setminus \Gamma} \nabla_h w_h \cdot [\![z_h]\!] \,\mathrm{d}s + \int_{\Gamma_N} w_h \mathbf{n} \cdot \nabla_h z_h \,\mathrm{d}s \\ + \int_{\Gamma_D} w_h \left((1+\theta) \mathbf{n} \cdot \nabla_h z_h + \delta z_h \right) \mathrm{d}s - \int_{\Gamma_D} \nabla_h w_h \cdot \mathbf{n}\, z_h \,\mathrm{d}s \\ = \int_\Omega w_h \, j_\Omega \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} \nabla w_h \cdot \mathbf{n} \, j_D \,\mathrm{d}s + \int_{\Gamma_N} w_h \, j_N \,\mathrm{d}s, \quad (307)$$

and we arrive at the element-based adjoint residual form: find $z_h \in V_h$ such that

$$\int_\Omega w_h \, R^*(z_h) \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} w_h \, r^*(z_h) + \nabla w_h \cdot \boldsymbol{\rho}^*(z_h) \,\mathrm{d}s \\ + \int_\Gamma w_h \, r^*_\Gamma(z_h) + \nabla w_h \cdot \boldsymbol{\rho}^*_\Gamma(z_h) \,\mathrm{d}s = 0 \quad \forall w_h \in V_h, \quad (308)$$

where the adjoint residuals are given by $R^*(z_h) = j_\Omega + \Delta_h z_h$ on Ω, by

$$r^*(z_h) = -\tfrac{1}{2} [\![\nabla_h z_h]\!] - (1+\theta) \mathbf{n} \cdot \{\!\{ \nabla_h z \}\!\} - \delta [z_h], \qquad \boldsymbol{\rho}^*(z_h) = \tfrac{1}{2} [\![z_h]\!], \quad (309)$$

on interior faces $\partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h$, and by

$$\begin{aligned} &r^*_\Gamma(z_h) = -(1+\theta) \mathbf{n} \cdot \nabla_h z_h - \delta z_h, && \boldsymbol{\rho}^*_\Gamma(z_h) = (j_D + z_h) \mathbf{n} && \text{on } \Gamma_D, \\ &r^*_\Gamma(z_h) = j_N - \mathbf{n} \cdot \nabla_h z_h, && \boldsymbol{\rho}^*_\Gamma(z_h) = 0 && \text{on } \Gamma_N. \end{aligned} \quad (310)$$

From (309) we see that the exact solution z to the adjoint problem (291) satisfies $r^*(z) = 0$ provided $\theta = -1$. Furthermore, we have $\boldsymbol{\rho}^*(z) = 0$ and $R^*(z) = 0$. This shows that NIPG and

the method by Baumann-Oden are adjoint inconsistent whereas the interior face terms of SIPG are adjoint consistent. This has already been shown in Corollary 5.17 for the Dirichlet problem with homogeneous boundary conditions (163). Furthermore, in [23] it has been demonstrated that the lack of adjoint consistency of the NIPG method leads to non-smooth adjoint solutions and a sub-optimal convergence of the method. In contrast to that the adjoint consistent SIPG method shows an optimal order of convergence.

As $r_\Gamma^*(z) = 0$ and $\boldsymbol{\rho}_\Gamma^*(z) = 0$ on Γ_N, the SIPG method is also adjoint consistent on Γ_N. However, on Γ_D the requirements $r_\Gamma^*(z) = 0$ and $\boldsymbol{\rho}_\Gamma^*(z) = 0$ reduce to the conditions $z = 0$ (note that $\theta = -1$) and $z = -j_D$, which are compatible for $j_D = 0$, but conflict for $j_D \neq 0$. This incompatibility can be resolved by modifying the target functional according to (288), with $i(u_h) = u_h$ and

$$r_J(u_h) = -\delta(u_h - g_D)j_D. \tag{311}$$

which, in the following, will be denoted by the *IP modification* of the target functional. This modification is consistent, as $i(u) = u$ and $r_J(u) = 0$ holds for the exact solution u to (289). As the modified functional is not linear in u_h (it is affine), the discrete adjoint problem includes its linearization as follows: find $z_h \in V_h$ such that

$$B_h(w_h, z_h) = \tilde{J}'[u_h](w_h) \quad \forall w_h \in V_h, \tag{312}$$

where

$$\tilde{J}'[u_h](w_h) = J'[u_h](w_h) + \int_{\Gamma_D} r_J'[u_h](w_h)\,\mathrm{d}s = J(w_h) - \int_{\Gamma_D} w_h\,\delta j_D\,\mathrm{d}s. \tag{313}$$

Then, the adjoint residuals on Γ_D are given by

$$r_\Gamma^*(z_h) = -\delta j_D - (1+\theta)\mathbf{n}\cdot\nabla_h z_h - \delta z_h, \qquad \boldsymbol{\rho}_\Gamma^*(z_h) = (j_D + z_h)\mathbf{n}, \qquad \text{on } \Gamma_D, \tag{314}$$

which vanish for $z = -j_D$. Thereby, the SIPG method is adjoint consistent also on Γ_D. Finally, we see that the discrete adjoint solution z_h must satisfy following problem in a weak sense

$$-\Delta z = j_\Omega \quad \text{in } \kappa, \kappa \in \mathcal{T}_h, \tag{315}$$

subject to inter-element conditions

$$\begin{aligned} \frac{1}{2}[\![\nabla_h z]\!] + (1+\theta)\mathbf{n}\cdot\{\!\{\nabla_h z\}\!\} + \delta[z] = 0 \quad &\text{on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \\ [z] = 0 \quad &\text{on } \partial\kappa \setminus \Gamma, \end{aligned} \tag{316}$$

and boundary conditions

$$\begin{aligned} z &= -j_D && \text{on } \partial\kappa \cap \Gamma_D, \\ (1+\theta)\mathbf{n}\cdot\nabla_h z + \delta z &= -\delta j_D && \text{on } \partial\kappa \cap \Gamma_D, \\ \mathbf{n}\cdot\nabla_h z &= j_N && \text{on } \partial\kappa \cap \Gamma_N. \end{aligned} \tag{317}$$

Note, that for $\theta = -1$ there is a correspondence to the primal equations (303)-(305). In fact, the discrete adjoint equations correspond to the discrete primal equations, with u, f, g_D and g_N replaced by z, j_Ω, $-j_D$ and j_N respectively; I.e. the discrete adjoint equation to the SIPG discretization is equivalent to the SIPG discretization of the continuous adjoint equation.

Finally, we note that [22] considers the target functional $J(u) = \int_{\Gamma_0} \mathbf{n}\cdot\nabla u j_D\,\mathrm{d}s$, $\Gamma_0 \subset \Gamma_D$, which is a special case of (290) with $j_\Omega \equiv 0$ in Ω, $j_N \equiv 0$ on Γ_N and $j_D \equiv 0$ on $\Gamma_D \setminus \Gamma_0$. Numerical experiments in [22] have shown, that the (discrete) adjoint solution associated with this target functional is non-smooth near Γ_0. Furthermore, it has been demonstrated that either by setting $\delta = 0$ on Γ_0, or by modifying the target functional appropriately, this effect vanishes, and the adjoint solution becomes smooth. We note, that the modification of the target functional supposed in [22] is connected to (313). However, here, we followed [27] and derived (313) in the more general framework of consistent modifications of target functionals, see (288).

6.4 Adjoint consistency analysis of the upwind DG discretization

In this section we apply the consistency and adjoint consistency analysis outlined in Section 6.2 to the upwind discontinuous Galerkin discretization of the linear advection equation.

We begin by recalling the linear advection equation:

$$\nabla \cdot (\mathbf{b}u) + cu = f \quad \text{in } \Omega, \qquad\qquad u = g \quad \text{on } \Gamma_-, \tag{318}$$

where $f \in L^2(\Omega)$, $\mathbf{b} \in [C^1(\Omega)]^d$, $c \in L^\infty(\Omega)$ and $g \in L^2(\Gamma_-)$, where

$$\Gamma_- = \{\mathbf{x} \in \Gamma, \mathbf{b}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\} \tag{319}$$

denotes the inflow part of the boundary $\Gamma = \partial\Omega$. Furthermore, we adopt the hypothesis (101).

6.4.1 The continuous adjoint problem to the linear advection equation

In order to derive the continuous adjoint problem, we multiply the left hand side of (318) by z, integrate over the domain Ω and integrate by parts. Thereby, we obtain

$$(\nabla \cdot (\mathbf{b}u) + cu, z)_\Omega + (u, -\mathbf{b} \cdot \mathbf{n}\, z)_{\Gamma_-} = (u, -\mathbf{b} \cdot \nabla z + cz)_\Omega + (u, \mathbf{b} \cdot \mathbf{n}\, z)_{\Gamma_+}. \tag{320}$$

Comparing with (275), we see that for $Lu = \nabla \cdot (\mathbf{b}u) + cu$ in Ω and

$$\begin{aligned} &Bu = u, && Cu = 0 && \text{on } \Gamma_-, \\ &Bu = 0, && Cu = u && \text{on } \Gamma_+, \end{aligned}$$

the adjoint operators are given by $L^*z = -\mathbf{b} \cdot \nabla z + cz$ in Ω and

$$\begin{aligned} &B^*z = 0, && C^*z = -\mathbf{b} \cdot \mathbf{n}\, z && \text{on } \Gamma_-, \\ &B^*z = \mathbf{b} \cdot \mathbf{n}\, z, && C^*z = 0 && \text{on } \Gamma_+. \end{aligned}$$

In particular, for

$$J(u) = \int_\Omega j_\Omega\, u \,\mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma\, Cu \,\mathrm{d}s = \int_\Omega j_\Omega\, u \,\mathrm{d}\mathbf{x} + \int_{\Gamma_+} j_\Gamma\, u \,\mathrm{d}s, \tag{321}$$

the continuous adjoint problem is given by

$$-\mathbf{b} \cdot \nabla z + cz = j_\Omega \quad \text{in } \Omega, \tag{322}$$

subject to the boundary condition

$$\mathbf{b} \cdot \mathbf{n}\, z = j_\Gamma \quad \text{on } \Gamma_+. \tag{323}$$

6.4.2 Primal residual form of the DG discretization based on upwind

We recall the discontinuous Galerkin discretization of the linear advection equation based on the upwind flux: find $u_h \in V_{h,p}^d$ such that

$$\begin{aligned} -\int_\Omega (\mathbf{b}u_h) \cdot \nabla_h v_h \,\mathrm{d}\mathbf{x} + \int_\Omega c u_h v_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa_- \setminus \Gamma} \mathbf{b} \cdot \mathbf{n}\, u_h^- v_h^+ \,\mathrm{d}s + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa_+} \mathbf{b} \cdot \mathbf{n}\, u_h^+ v_h^+ \,\mathrm{d}s \\ = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n}\, g v_h \,\mathrm{d}s \quad \forall v_h \in V_{h,p}^d. \end{aligned} \tag{324}$$

Furthermore, we recall the primal residual form as given in Section 4.7.3: find $u_h \in V_{h,p}^d$ such that

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa R(u_h)v\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} r(u_h)v\,\mathrm{d}s + \int_\Gamma r_\Gamma(u_h)v\,\mathrm{d}s = 0 \quad \forall v \in V_{h,p}^d, \tag{325}$$

where $R(u_h)$, $r(u_h)$ and $r_\Gamma(u_h)$ denote the element, interior face and boundary residuals, respectively, given by

$$\begin{aligned} R(u_h) &= f - \nabla_h\cdot(\mathbf{b}u_h) - cu_h && \text{in } \kappa, \kappa\in\mathcal{T}_h,\\ r(u_h) &= \mathbf{b}\cdot\mathbf{n}\,(u_h^+ - u_h^-) && \text{on } \partial\kappa_- \setminus \Gamma, \kappa\in\mathcal{T}_h,\\ r_\Gamma(u_h) &= \mathbf{b}\cdot\mathbf{n}(u_h - g) && \text{on } \Gamma_-, \end{aligned}$$

and $r_\Gamma(u_h) \equiv 0$ on Γ_+. I.e. we have (282) with $\boldsymbol{\rho}_\Gamma(u_h) \equiv 0$ and $\boldsymbol{\rho}(u_h) \equiv 0$. Furthermore, we see that the exact solution $u \in H^{1,\mathbf{b}}(\Omega)$ to (318) satisfies (282) with $R(u) = 0$, $r(u) = 0$ and $r_\Gamma(u) = 0$. Thereby, (324) is a consistent discretization of (318). Finally, we see that the discrete solution u_h to (318) must satisfy following problem in a weak sense

$$\nabla\cdot(\mathbf{b}u) + cu = f \quad \text{in } \kappa, \kappa\in\mathcal{T}_h, \tag{326}$$

subject to inter-element conditions

$$-\mathbf{b}\cdot\mathbf{n}\,[u] = 0 \quad \text{on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h, \tag{327}$$

and boundary conditions

$$\mathbf{b}\cdot\mathbf{n}\,u = \mathbf{b}\cdot\mathbf{n}\,g \quad \text{on } \Gamma_-. \tag{328}$$

This is the mesh-dependent counterpart of the original equations (318) to be solved.

6.4.3 Adjoint residual form of the DG discretization based on upwind

Substituting

$$\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_-\setminus\Gamma} \mathbf{b}\cdot\mathbf{n}\,u_h^- v^+\,\mathrm{d}s = -\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_+\setminus\Gamma} \mathbf{b}\cdot\mathbf{n}\,u_h^+ v^-\,\mathrm{d}s$$

in (324) we find that the discrete adjoint problem (284) to the discretization (324) is given by: find $z_h \in V_h$ such that

$$B_h(w_h, z_h) \equiv \int_\Omega w_h\,(-\mathbf{b}\cdot\nabla_h z_h + cz_h)\,\mathrm{d}\mathbf{x} + \sum_\kappa\int_{\partial\kappa_+} w_h^+\,\mathbf{b}\cdot\mathbf{n}[z_h]\,\mathrm{d}s = J(w_h),$$

for all $w_h \in V_h$. Hence, for a target functional $J(\cdot)$ as in (321), we have (285) with

$$\begin{aligned} R^*(z_h) &= j_\Omega + \mathbf{b}\cdot\nabla_h z_h - cz_h && \text{in } \kappa, \kappa\in\mathcal{T}_h,\\ r^*(z_h) &= -\mathbf{b}\cdot\mathbf{n}\,[z_h] && \text{on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h,\\ r^*(z_h) &= j_\Gamma - \mathbf{b}\cdot\mathbf{n}\,z_h && \text{on } \Gamma_+, \end{aligned}$$

$r^*(z_h) \equiv 0$ on Γ_-, $\boldsymbol{\rho}^*(z_h) \equiv 0$ on $\partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h$, and $\boldsymbol{\rho}_\Gamma^*(z_h) \equiv 0$ on Γ. As (286) with (287) holds for the exact solution z to (322) and (323), we conclude, that (324) is an adjoint consistent discretization of (318). Furthermore, we see that the discrete adjoint solution z_h must satisfy following problem in a weak sense

$$-\mathbf{b}\cdot\nabla z + cz = j_\Omega \quad \text{in } \kappa, \kappa\in\mathcal{T}_h, \tag{329}$$

subject to inter-element conditions

$$\mathbf{b}\cdot\mathbf{n}\,[z] = 0 \quad \text{on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h, \tag{330}$$

and boundary conditions

$$\mathbf{b}\cdot\mathbf{n}\,z = j_\Gamma \quad \text{on } \Gamma_+. \tag{331}$$

This is the mesh-dependent counterpart of the adjoint equations given in Section 6.4.1.

7 *A priori* error estimates for target functionals $J(\cdot)$

In this section we derive *a priori* error estimates with respect to target functionals $J(\cdot)$ for adjoint consistent and adjoint inconsistent discontinuous Galerkin discretizations. In particular, we will see that analogous to the (sub-)optimal order of convergence in the L^2-norm for adjoint (in-)consistent discretizations also the order of convergence in $J(\cdot)$ is (sub-)optimal.

We consider following general linear problem:

$$Lu = f \quad \text{in } \Omega, \qquad\qquad Bu = g \quad \text{on } \Gamma, \tag{332}$$

where $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$, L denotes a linear differential operator on Ω and B a linear differential boundary operator on Γ. Let (332) be discretized as follows: find $u_h \in V_{h,p}^d$ such that

$$B_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_{h,p}^d, \tag{333}$$

where the bilinear form $B_h(\cdot,\cdot)$ is continuous on V with respect to a specific $|||\cdot|||$-norm, i.e.

$$B_h(w, v) \leq C_B |||w||| \, |||v||| \quad \forall w, v \in V.$$

Here, V is suitably chosen function space such that $u \in \tilde{V} \subset V$ and $V_h \subset V$ but possibly $V_h \not\subset \tilde{V}$.

Remark 7.1 *We recall that for the DG discretization of Poisson's equation we have $\tilde{V} = H^2(\Omega)$ and $V = H^2(\mathcal{T}_h)$, and for the DG discretization of the linear advection equation we have $\tilde{V} = H^{1,\mathbf{b}}(\Omega)$ and $V = H^{1,\mathbf{b}}(\mathcal{T}_h)$.*

We assume that the discretization (333) is consistent, i.e. the exact solution $u \in \tilde{V} \subset V$ to (332) satisfies

$$B_h(u, v) = F_h(v) \quad \forall v \in V, \tag{334}$$

which implies the Galerkin orthogonality

$$B_h(u - u_h, v_h) = 0 \quad \forall v_h \in V_{h,p}^d. \tag{335}$$

Furthermore, we assume that following *a priori* error estimate in the $|||\cdot|||$-norm holds: There are constants $C > 0$ and $r = r(p) > 0$ such that

$$|||u - u_h||| \leq C h^r |u|_{H^{p+1}(\Omega)} \quad \forall u \in H^{p+1}(\Omega). \tag{336}$$

Finally, we assume that the local projection operator $P_{h,p}^d$ as defined in Section 4.5 satisfies following approximation estimate in the $|||\cdot|||$ norm: There are constants $C > 0$ and $\tilde{r} = \tilde{r}(p) > 0$ such that

$$|||v - P_{h,p}^d v||| \leq C h^{\tilde{r}} |v|_{H^{p+1}(\Omega)} \qquad \forall v \in H^{p+1}(\Omega). \tag{337}$$

In the following we want to measure the discretization error $e = u - u_h$ not in some global norm like the $|||\cdot|||$-norm but with respect to target functionals $J(\cdot)$ of the form

$$J(u) = (j_\Omega, u)_\Omega + (j_\Gamma, Cu)_\Gamma \equiv \int_\Omega j_\Omega \, u \, \mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma \, Cu \, \mathrm{d}s, \tag{338}$$

where $j_\Omega \in L^2(\Omega)$ and $j_\Gamma \in L^2(\Gamma)$, and C is an differential boundary operator on Γ. We assume that $J(\cdot)$ is compatible with (332) as defined in Section 6.1. Then, there are differential operators L^*, B^* and C^* which are the adjoint operators to L, B and C, respectively, and the continuous adjoint problem is given by:

$$L^* z = j_\Omega \quad \text{in } \Omega, \qquad\qquad B^* z = j_\Gamma \quad \text{on } \Gamma. \tag{339}$$

Finally, we recall that a discretization together with a target functional is called adjoint consistent, if the exact solution $z \in \tilde{V} \subset V$ to the adjoint problem (339) satisfies:

$$B_h(w, z) = J(w) \quad \forall w \in V. \tag{340}$$

Theorem 7.2 (*A priori* error estimates in $J(\cdot)$) *Let the situation be as described above. Furthermore, assume that j_Ω and j_Γ are smooth functions on Ω and Γ, respectively. Finally, assume that the adjoint solution z in* (339) *is smooth, $z \in H^{p+1}(\Omega)$. Then, we have following estimates:*

a) If the discretization (333) *together with the target functional $J(\cdot)$ is* adjoint consistent*, then there is a constant $C > 0$ such that*

$$|J(u) - J(u_h)| \leq Ch^{r+\tilde{r}}|u|_{H^{p+1}(\Omega)}|z|_{H^{p+1}(\Omega)} \qquad \forall u \in H^{p+1}(\Omega). \tag{341}$$

b) If, however, the discretization is adjoint inconsistent *we only have:*

$$|J(u) - J(u_h)| \leq Ch^{r}|u|_{H^{p+1}(\Omega)} \qquad \forall u \in H^{p+1}(\Omega). \tag{342}$$

c) If the discretization is adjoint inconsistent*, but $z \in \tilde{V}$ in* (339) *satisfies*

$$B_h(w, z) = J(w) \quad \forall w \in \tilde{V}, \tag{343}$$

i.e. the discrete adjoint problem reduces to a weak formulation of the continuous adjoint problem if tested with smooth functions $w \in \tilde{V}$ instead of $w \in V_{h,p}^d$ or $w \in V$, then

$$|J(u) - J(u_h)| \leq Ch^{r+\tilde{r}}|u|_{H^{p+1}(\Omega)}|z|_{H^{p+1}(\Omega)} + B_h(u_h, z - z_h) \qquad \forall u \in H^{p+1}(\Omega), \tag{344}$$

where $z_h \in V_{h,p}^d$ is the solution to the discrete adjoint problem

$$B_h(w_h, z_h) = J(w_h) \quad \forall w_h \in V_{h,p}^d. \tag{345}$$

Proof: a) For an adjoint consistent discretization we set $w := e = u - u_h \in V$ in (340),

$$\begin{aligned} |J(u) - J(u_h)| = |J(e)| = |B_h(e, z)| = |B_h(u - u_h, z - P_h z)| &\leq C|||u - u_h||| \, |||z - P_h z||| \\ &\leq Ch^{r}|u|_{H^{p+1}(\Omega)} Ch^{\tilde{r}}|z|_{H^{p+1}(\Omega)}, \end{aligned} \tag{346}$$

where here and in the following we use P_h as a short notation for $P_{h,p}^d$. Hence we have (341).

b) For an adjoint inconsistent discretization we do not have (340). Thereby, in order to represent the error $J(u) - J(u_h)$ we define following mesh-dependent adjoint problem: find $\psi \in V$ such that

$$B_h(w, \psi) = J(w) \quad \forall w \in V. \tag{347}$$

We note, that for an adjoint consistent discretization the solution ψ to (347) coincides with the solution z to the continuous adjoint solution (339) and thus is smooth. For an adjoint inconsistent discretization, however, we cannot expect ψ to be smooth. In that case ψ is mesh-dependent and in general discontinuous across interior faces. We then proceed as follows

$$\begin{aligned} |J(u) - J(u_h)| = |J(e)| = |B_h(e, \psi)| = |B_h(u - u_h, \psi - P_h\psi)| &\leq C|||u - u_h||| \, |||\psi - P_h\psi||| \\ &\leq Ch^{r}|u|_{H^{p+1}(\Omega)}, \end{aligned}$$

where due to the lack of smoothness of ψ here we do not gain additional orders of h from $|||\psi - P_h\psi|||$.

c) For an adjoint inconsistent discretization we do not have (340). But if (343) holds we still can represent the error $J(u) - J(u_h)$ as follows:

$$J(u) - J(u_h) = B_h(u, z) - B_h(u_h, z_h), \tag{348}$$

where $z_h \in V_{h,p}^d$ is the solution to the discrete adjoint problem

$$B_h(w_h, z_h) = J(w_h) \quad \forall w_h \in V_{h,p}^d. \tag{349}$$

Using Galerkin orthogonality (335) for $v_h := P_h z \in V_{h,p}^d$ we obtain

$$\begin{aligned} J(u) - J(u_h) &= B_h(u, z) - B_h(u_h, z_h) \\ &= B_h(u - u_h, z) + B_h(u_h, z) - B_h(u_h, z_h) \\ &= B_h(u - u_h, z - P_h z) + B_h(u_h, z - z_h). \end{aligned} \tag{350}$$

Here, the first term is the standard error which can be bounded

$$B_h(u - u_h, z - P_h z) \le C h^{r+s} |u|_{H^{p+1}(\Omega)} |z|_{H^{p+1}(\Omega)} \tag{351}$$

like in part a) of this proof which together with (350) results in (344). □

Remark 7.3 *The second term in* (344) *represents the adjoint consistency error. For an adjoint consistent discretization this term vanishes due to the* adjoint Galerkin orthogonality

$$B_h(w_h, z - z_h) = B_h(w_h, z) - B_h(w_h, z_h) = J(w_h) - J(w_h) = 0, \qquad \forall w_h \in V_{h,p}^d, \tag{352}$$

which we obtain by subtracting (345) *from* (340)*. However, for adjoint inconsistent discretizations the adjoint consistency error* $B_h(u_h, z - z_h)$ *does not vanish. Instead, by continuity of* B_h *we obtain*

$$B_h(u_h, z - z_h) \le C |||u_h||| \, |||z - z_h|||. \tag{353}$$

7.1 Upwind DG of the linear advection equation: Estimates in $J(\cdot)$

We consider the linear advection equation

$$\nabla \cdot (\mathbf{b}u) + cu = f \quad \text{in } \Omega, \qquad\qquad u = g \quad \text{on } \Gamma_-, \tag{354}$$

and recall its DG discretization based on the upwind flux: find $u_h \in V_{h,p}^d$ such that

$$\begin{aligned} B_h(u_h, v_h) \equiv &- \int_\Omega (\mathbf{b}u_h) \cdot \nabla_h v_h \, \mathrm{d}\mathbf{x} + \int_\Omega c u_h v_h \, \mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa_- \setminus \Gamma} \mathbf{b} \cdot \mathbf{n} \, u_h^- v_h^+ \, \mathrm{d}s \\ &+ \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa_+} \mathbf{b} \cdot \mathbf{n} \, u_h^+ v_h^+ \, \mathrm{d}s = \int_\Omega f v_h \, \mathrm{d}\mathbf{x} - \int_{\Gamma_-} \mathbf{b} \cdot \mathbf{n} \, g v_h \, \mathrm{d}s \quad \forall v_h \in V_{h,p}^d. \end{aligned} \tag{355}$$

This discretization is consistent, see Section 6.4.2. Furthermore, the bilinear form B_h is continuous with respect to the $|||\cdot|||_{b_0}$-norm given by

$$|||v|||_{b_0}^2 = \|c_0 v\|^2 + \sum_{e \subset \Gamma_{\mathcal{I}}} \int_e b_0 \, [v]^2 \, \mathrm{d}s + \frac{1}{2} \int_\Gamma |\mathbf{b} \cdot \mathbf{n}| \, v^2 \, \mathrm{d}s. \tag{356}$$

where $b_0 = \frac{1}{2}|\mathbf{b} \cdot \mathbf{n}|$.

Furthermore, we have an approximation estimate for the local projection operator $P_{h,p}^d$:

$$|||v - P_{h,p}^d v|||_{b_0} \le C h^{p+1/2} |v|_{H^{p+1}(\Omega)} \qquad \forall v \in H^{p+1}(\Omega), \tag{357}$$

see [proof of Theorem 8.20]. Finally, we recall the *a priori* error estimate in Theorem 4.17: There is a constant $C > 0$ such that

$$|||u - u_h|||_{b_0} \le C h^{p+1/2} |u|_{H^{p+1}(\Omega)} \qquad \forall u \in H^{p+1}(\Omega), \tag{358}$$

which in view of (357) is optimal. In the following we give *a priori* error estimates for the discretization error $e = u - u_h$ measured in terms of target quantities $J(\cdot)$ of the form

$$J(u) = \int_\Omega j_\Omega \, u \, \mathrm{d}\mathbf{x} + \int_{\Gamma_+} j_\Gamma \, u \, \mathrm{d}s, \tag{359}$$

which are compatible with the linear advection, see Section 6.4.1. The corresponding continuous adjoint problem is

$$-\mathbf{b} \cdot \nabla z + cz = j_\Omega \quad \text{in } \Omega, \qquad\qquad \mathbf{b} \cdot \mathbf{n} \, z = j_\Gamma \quad \text{on } \Gamma_+. \tag{360}$$

Finally, we recall from Section 6.4.3 that the discretization (355) together with the target functional (359) is adjoint consistent.

If we now had continuity of the bilinear form $B_h(\cdot,\cdot)$ in (355) all necessary conditions would be fulfilled in order to employ Theorem 7.2a) for obtaining an *a priori* error estimate in $J(\cdot)$. However, continuity of $B_h(\cdot,\cdot)$, i.e. $|B_h(u,v)| \leq C|||u|||_{b_0}|||v|||_{b_0}$ for all $u, v \in H^{1,\mathbf{b}}(\mathcal{T}_h)$, is not available. Nevertheless, there are alternative proofs which result in following estimate:

Corollary 7.4 (*A priori* error estimate in $J(\cdot)$) *Let $B_h(\cdot,\cdot)$ and $J(\cdot)$ be given by* (355) *and* (359)*, respectively. Let j_Ω and j_Γ in* (359) *be smooth functions such that solution z to the continuous adjoint problem* (360) *is smooth, $z \in H^{p+1}(\Omega)$. Then, there is a constant $C > 0$ such that*

$$|J(u) - J(u_h)| \leq Ch^{2p+1}|u|_{H^{p+1}(\Omega)}|z|_{H^{p+1}(\Omega)} \qquad \forall u \in H^{p+1}(\Omega). \tag{361}$$

Proof: See [35, 23]. □

We see, that the order of convergence in $J(\cdot)$ is $\mathcal{O}(h^{2p+1})$ provided both primal and adjoint solutions are smooth, $u \in H^{p+1}(\Omega)$ and $z \in H^{p+1}(\Omega)$. If, however, u or z are less regular we obtain an estimate with a correspondingly reduced order of convergence in the target quantity $J(\cdot)$:

Corollary 7.5 (*A priori* error estimate in $J(\cdot)$ with reduced regularity) *Let $B_h(\cdot,\cdot)$ and $J(\cdot)$ be given by* (355) *and* (359)*, respectively. Furthermore, assume that $u \in H^{s+1}(\Omega)$ and $z \in H^{\tilde{s}+1}(\Omega)$ hold for the exact solutions u and z to the primal and adjoint problems* (354) *and* (360)*, respectively. Then, there is a constant $C > 0$ such that*

$$|J(u) - J(u_h)| \leq Ch^{t+\tilde{t}+1}|u|_{H^{t+1}(\Omega)}|z|_{H^{\tilde{t}+1}(\Omega)} \qquad \forall u \in H^{s+1}(\Omega), \tag{362}$$

where $t = \min(s,p)$ and $\tilde{t} = \min(\tilde{s},p)$.

7.2 IP DG discretization for Poisson's equation: Estimates in $J(\cdot)$

We consider the Dirichlet-Neumann boundary value problem of Poisson's equation,

$$-\Delta u = f \quad \text{in } \Omega, \qquad\qquad u = g_D \quad \text{on } \Gamma_D, \qquad\qquad \mathbf{n} \cdot \nabla u = g_N \quad \text{on } \Gamma_N, \tag{363}$$

and its symmetric ($\theta = -1$) and unsymmetric ($\theta = 1$) interior penalty DG discretization given by: find $u_h \in V_{h,p}^d$ such that

$$\begin{aligned} \int_\Omega \nabla_h u_h \cdot \nabla_h v_h \, \mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I} \cup \Gamma_D} \left(\theta [\![u_h]\!] \cdot \{\!\{\nabla_h v_h\}\!\} - \{\!\{\nabla_h u_h\}\!\} \cdot [\![v_h]\!]\right) \mathrm{d}s + \int_{\Gamma_\mathcal{I} \cup \Gamma_D} \delta [\![u_h]\!] \cdot [\![v_h]\!] \, \mathrm{d}s, \\ = \int_\Omega f v_h \, \mathrm{d}\mathbf{x} + \int_{\Gamma_D} \theta g_D \, \mathbf{n} \cdot \nabla v_h \, \mathrm{d}s + \int_{\Gamma_D} \delta g_D v_h \, \mathrm{d}s + \int_{\Gamma_N} g_N v_h \, \mathrm{d}s \qquad \forall v_h \in V_{h,p}^d. \end{aligned} \tag{364}$$

This discretization is consistent, see Section (6.3.2). We recall, see Lemma 5.19, that B_h given by the left hand side in (364) is continuous,

$$B_h(w,v) \le C_B |||w|||_\delta |\, |||v|||_\delta \qquad \forall w, v \in V,$$

with respect to the $|||\cdot|||_\delta^2$-norm defined by

$$|||v|||_\delta^2 = \|\nabla_h v\|_{L^2(\Omega)}^2 + \int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} \delta^{-1} \left(\mathbf{n} \cdot \{\!\{\nabla v\}\!\}\right)^2 \,\mathrm{d}s + \int_{\Gamma_{\mathcal{I}} \cup \Gamma_D} \delta [v]^2 \,\mathrm{d}s.$$

Furthermore, we have an approximation estimate for the local projection operator $P_{h,p}^d$:

$$|||v - P_{h,p}^d v|||_\delta \le C h^p |v|_{H^{p+1}(\Omega)} \qquad \forall v \in H^{p+1}(\Omega), \tag{365}$$

see proof of Lemma 5.27. Finally, we recall the *a priori* error estimate in Lemma 5.27: There is a constant $C > 0$ such that

$$|||u - u_h|||_\delta \le C h^p |u|_{H^{p+1}(\Omega)} \qquad \forall u \in H^{p+1}(\Omega), \tag{366}$$

which in view of (365) is optimal. In the following we give *a priori* error estimates for the discretization error $e = u - u_h$ measured in terms of target quantities $J(\cdot)$ of the form

$$J(u) = \int_\Omega j_\Omega \, u \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} j_D \, \mathbf{n} \cdot \nabla u \,\mathrm{d}s + \int_{\Gamma_N} j_N \, u \,\mathrm{d}s, \tag{367}$$

which are compatible with Poisson's equation, see Section 6.3.1. The corresponding continuous adjoint problem is

$$-\Delta z = j_\Omega \quad \text{in } \Omega, \qquad\qquad -z = j_D \quad \text{on } \Gamma_D, \qquad\qquad \mathbf{n} \cdot \nabla z = j_N \quad \text{on } \Gamma_N. \tag{368}$$

Finally, we recall from Section 6.4.3 that the symmetric version of the IP discretization (364) together with following consistent modification

$$\begin{aligned} \tilde{J}(u_h) &= J(u_h) - \int_{\Gamma_D} \delta (u_h - g_D) j_D \,\mathrm{d}s, \\ \tilde{J}'[u_h](w_h) &= J(w_h) - \int_{\Gamma_D} w_h \, \delta j_D \,\mathrm{d}s \end{aligned} \tag{369}$$

of the target functional (367) is adjoint consistent whereas the unsymmetric version is adjoint inconsistent. Having collected these results we now can use Theorem 7.2 to obtain following estimates.

Corollary 7.6 (*A priori* error estimate in $J(\cdot)$) *Let $B_h(\cdot,\cdot)$ be given by the left hand side of (364). Furthermore, let $\tilde{J}(\cdot)$ be given by (369) and (367). Let j_Ω, j_D and j_N in (367) be smooth functions such that the adjoint solution z to (368) is smooth, $z \in H^{p+1}(\Omega)$. Then, we have following estimate for an adjoint consistent SIPG discretization:*

$$|J(u) - \tilde{J}(u_h)| \le C h^{2p} |u|_{H^{p+1}(\Omega)} |z|_{H^{p+1}(\Omega)} \qquad \forall u \in H^{p+1}(\Omega). \tag{370}$$

and for an adjoint inconsistent IP discretization, e.g. NIPG:

$$|J(u) - J(u_h)| \le C h^p |u|_{H^{p+1}(\Omega)} \qquad \forall u \in H^{p+1}(\Omega). \tag{371}$$

Proof: Use (341) in Theorem 7.2 with $r(p) = \tilde{r}(p) = p$ and (342) with $r(p) = p$. □

In this proof we applied statement a) of Theorem 7.2. In the following we give an alternative proof, see also [23], which is connected to part c) of Theorem 7.2 as well as to the proof of Lemma 5.29.
Proof: (Alternative proof of Corollary 7.6, [23]): Let $B_h^s(\cdot,\cdot)$ and $B_h^n(\cdot,\cdot)$, and $F_h^s(\cdot)$ and $F_h^n(\cdot)$ denote the bilinear and linear forms in (364) for, respectively, the **s**ymmetric ($\theta = -1$) and **n**onsymmetric ($\theta = 1$) interior penalty DG discretizations. Then the discrete solutions u_h^s and u_h^n satisfy the following problems: find $u_h^s \in V_{h,p}^d$ such that

$$B_h^s(u_h^s, v_h) = F_h^s(v_h) \qquad \forall v_h \in V_{h,p}^d; \tag{372}$$

and find $u_h^n \in V_{h,p}^d$ such that

$$B_h^n(u_h^n, v_h) = F_h^n(v_h) \qquad \forall v_h \in V_{h,p}^d, \tag{373}$$

respectively. Furthermore, let z^s and z^n be the analytical solutions to following adjoint problems: find $z^s \in H^2(\mathcal{T}_h)$ such that

$$B_h^s(w, z^s) = J(w) \qquad \forall w \in H^2(\mathcal{T}_h); \tag{374}$$

and find $z^n \in H^2(\mathcal{T}_h)$ such that

$$B_h^n(w, z^n) = J(w) \qquad \forall w \in H^2(\mathcal{T}_h), \tag{375}$$

respectively. Then we obtain

$$J(u) - J(u_h^s) = B_h^s(u - u_h^s, z^s) = B_h^s(u - u_h^s, z^s - z_h), \tag{376}$$

due to Galerkin orthogonality, where $z_h \in V_{h,p}^d$ is any discrete function. We note that due to the adjoint consistency of B_h^s the solution z^s to the adjoint problem (374) coincides with the solution z to the continuous adjoint problem (368) and thus is smooth. By using coercivity of B_h^s, smoothness of the adjoint solution $z^s \in H^{p+1}(\Omega)$, and estimates (366) and (365) in (376) we obtain (370) for the SIPG discretization. Similarly, for the NIPG discretization we obtain

$$J(u) - J(u_h^n) = B_h^n(u - u_h^n, z^n) = B_h^n(u - u_h^n, z^n - z_h). \tag{377}$$

In the following we rewrite the error representation (377) in terms of the (smooth) adjoint solution z^s instead of the solution z^n (374) which is mesh-dependent and in general non-smooth. Before we proceed, we note that

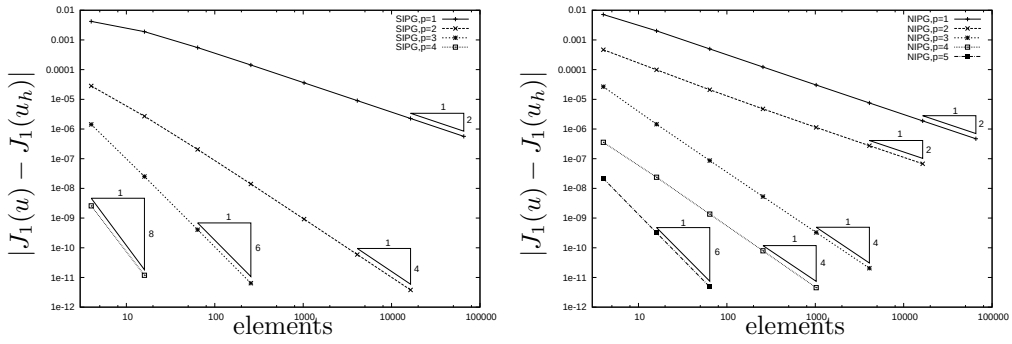
$$B_h^n(w_h, v_h) = B_h^s(w_h, v_h) + 2\int_{\Gamma_\mathcal{I}\cup\Gamma_D} [\![w_h]\!] \cdot \{\!\{ \nabla_h v_h \}\!\} \,\mathrm{d}\mathbf{x}.$$

Thereby and by using the Galerkin orthogonality of the NIPG discretization we obtain:

$$\begin{aligned}
J(u) - J(u_h^n) &= B_h^n(u - u_h^n, z^n) = B_h^s(u - u_h^n, z^s) \\
&= B_h^n(u - u_h^n, z^s) - 2\int_{\Gamma_\mathcal{I}\cup\Gamma_D} [\![u - u_h^n]\!] \cdot \{\!\{ \nabla_h z^s \}\!\} \,\mathrm{d}\mathbf{x} \\
&= B_h^n(u - u_h^n, z^s - z_h) - 2\int_{\Gamma_\mathcal{I}\cup\Gamma_D} [\![u - u_h^n]\!] \cdot \{\!\{ \nabla_h z^s \}\!\} \,\mathrm{d}\mathbf{x}, \\
&\leq Ch^{2p}|u|_{H^{p+1}(\Omega)}|z|_{H^{p+1}(\Omega)} + C|\!|\!| u - u_h |\!|\!|_\delta \, \|z^s\|_{H^2(\mathcal{T}_h)} \\
&\leq Ch^{2p}|u|_{H^{p+1}(\Omega)}|z|_{H^{p+1}(\Omega)} + Ch^p|u|_{H^{p+1}(\Omega)}\|z^s\|_{H^2(\mathcal{T}_h)} \leq Ch^p|u|_{H^{p+1}(\Omega)},
\end{aligned}$$

where we used $\|z^s\|_{H^2(\mathcal{T}_h)} \leq C$. Hence we have (371). □

We see, that the order of convergence in $J(\cdot)$ is $\mathcal{O}(h^{2p})$ for SIPG and $\mathcal{O}(h^p)$ for NIPG provided both primal and adjoint solutions are smooth, $u \in H^{p+1}(\Omega)$ and $z \in H^{p+1}(\Omega)$. If, however, u or z are less regular we obtain an estimate with a correspondingly reduced order of convergence in $J(\cdot)$:



(a) (b)

Figure 5: Example 1: Convergence of the error $|J_1(u) - J_1(u_h)|$ for a) the SIPG and b) the NIPG discretizations with global mesh refinement.

Corollary 7.7 (*A priori* error estimate in $J(\cdot)$ with reduced regularity) *Let $B_h(\cdot,\cdot)$ be given by the left hand side of* (364) *and let $\tilde{J}(\cdot)$ be given by* (369) *and* (367)*. Assume that $u \in H^{s+1}(\Omega)$ and $z \in H^{\tilde{s}+1}(\Omega)$ hold for the exact solutions u and z to the primal and adjoint problems* (363) *and* (368)*, respectively. Then, we have following estimates for an adjoint consistent SIPG discretization:*

$$|J(u) - J(u_h)| \leq Ch^{t+\tilde{t}}|u|_{H^{t+1}(\Omega)}|z|_{H^{\tilde{t}+1}(\Omega)} \qquad \forall u \in H^{s+1}(\Omega), \tag{378}$$

and for an adjoint inconsistent discretization, e.g. NIPG:

$$|J(u) - J(u_h)| \leq Ch^{t}|u|_{H^{t+1}(\Omega)} \qquad \forall u \in H^{s+1}(\Omega), \tag{379}$$

where $t = \min(s,p)$ and $\tilde{t} = \min(\tilde{s},p)$.

7.3 Numerical results

Example 1 We begin by investigating the experimental order of convergence of the SIPG and NIPG discretizations when measuring the error in terms of specific target quantities $J(\cdot)$.

To this end we revisit the experimental *model problem* introduced in Section 5.6. This problem is based on Poisson's equations with an inhomogeneous Dirichlet boundary value function g_D and a forcing function f chosen so that the analytical solution u is given by (272).

First, we choose the target quantity to represent the (weighted) mean value of u over Ω, i.e.

$$J_1(u_h) = \int_\Omega j_\Omega \, u_h \, \mathrm{d}\mathbf{x}; \tag{380}$$

here, we define the weight function j_Ω by

$$j_\Omega(\mathbf{x}) = \sin(\pi x_1)\sin(\pi x_2).$$

Thereby the true value of the target quantity is given by $J(u) = 0.1801265486975$. We note that the target quantity (380) is compatible with Poisson's equation (289). In fact, it is a special case of the target quantity given in (290) with $\Gamma_N = \emptyset$ and $j_D = 0$ on $\Gamma_D = \Gamma$.

Figure 5a) shows the error of the SIPG discretization measured in terms of the target quantity $J_1(\cdot)$ given by (380). We see that under global mesh refinement the error $|J_1(u) - J_1(u_h)|$ behaves like $O(h^{2p})$ which is in perfect agreement with the theoretical order of convergence, see estimate (370). Figure 5b) shows the respective plot for the NIPG discretization. Here, we see that the error $|J_1(u) - J_1(u_h)|$ behaves like $\mathcal{O}(h^{p+1})$ for odd p and like $\mathcal{O}(h^p)$ for even p. This convergence

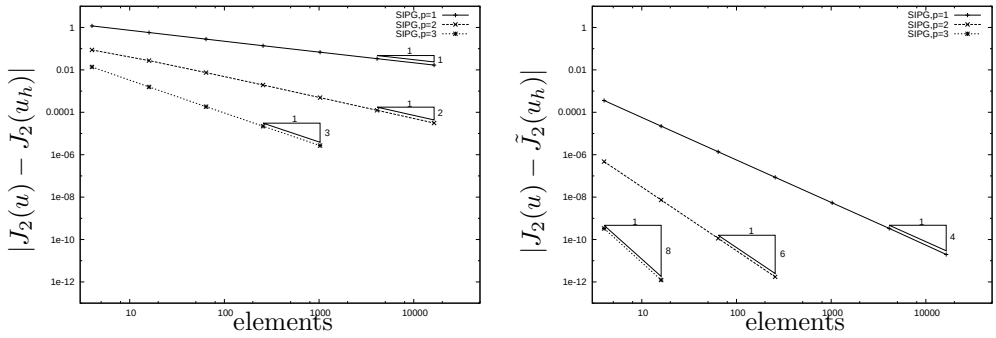


Figure 6: Example 2: Convergence of a) the error $|J_2(u) - J_2(u_h)|$ and b) the error $|J_2(u) - \tilde{J}_2(u_h)|$ for the SIPG discretization with global mesh refinement.

behavior is similar to the convergence behavior in the $L^2(\Omega)$-norm encountered for the NIPG scheme in Section 5.6. Again, due to the lack of adjoint consistency the order of convergence in $J_1(\cdot)$ for the NIPG scheme is lower than in the case of the adjoint consistent SIPG scheme. We note that similar results for a different test case have been obtained in [23].

Example 2 We consider the same *model problem* as in the previous example. However, instead of the mean value quantity (380) here we now choose the target quantity to represent the mean value of the normal derivative of u over the boundary $\Gamma = \partial\Omega$, i.e.

$$J_2(u_h) = \int_\Gamma j_D \, \mathbf{n} \cdot \nabla_h u_h \, \mathrm{d}s, \tag{381}$$

with $j_D \equiv 1$ on $\Gamma_D = \Gamma$. Thereby the true value of the target quantity is given by $J(u) = -2$. We note that this target quantity is compatible with Poisson's equation (289). In fact, it is a special case of the target quantity given in (290) with $\Gamma_N = \emptyset$, and $j_\Omega = 0$ on Ω. Furthermore, we note that the solution z to the corresponding continuous adjoint problem

$$-\Delta z = 0 \quad \text{in } \Omega, \qquad\qquad -z = j_D \quad \text{on } \Gamma_D \tag{382}$$

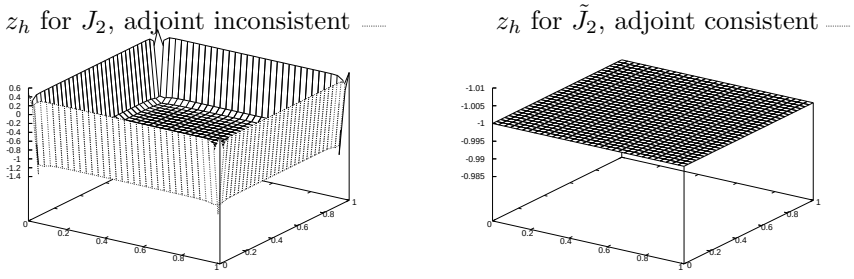
is given by $z \equiv -1$ on Ω. Figure 6a) shows that the error $|J_2(u) - J_2(u_h)|$ behaves like $O(h^p)$ for the SIPG discretizations with $p = 1, 2, 3$. Following the discussion in Section 6.3.3 we recognize that the SIPG discretization in combination with the target functional $J_2(\cdot)$ in (381) is adjoint *in*consistent. Thus the order of convergence $O(h^p)$ encountered in Figure 6a) is, in fact, the expected order of convergence for this adjoint inconsistent discretization, see estimate (371).

However, we recall from Sections 6.3.3 and 7.2 that the following modification

$$\tilde{J}_2(u_h) := J_2(u_h) - \int_\Gamma \delta(u_h - g_D) j_D \, \mathrm{d}s \tag{383}$$

of the target functional $J_2(u_h)$ leads to an adjoint consistent discretization. Here, δ is the penalization parameter of the IP discretization and g_D is the boundary value function of the *model problem* considered. Note, that $\tilde{J}_2(u_h)$ in (383) is a consistent modification of $J_2(u_h)$ as the true value of the target quantity is unchanged: $\tilde{J}_2(u) = J_2(u) = -2$ holds for the exact solution u.

Figure 6b) shows the error of the SIPG discretization measured in terms of the modified target quantity $\tilde{J}_2(\cdot)$ given in (383). We see that the consistent modification of the target functional leads to a significant increase in the accuracy and the order of convergence of the discretization. In



(a) (b)

Figure 7: Example 2: Discrete adjoint solution z_h corresponding a) to the adjoint inconsistent SIPG discretization with $J_2(u_h)$ and b) to the adjoint consistent SIPG discretization with $\tilde{J}_2(u_h)$.

fact, we see that under global mesh refinement the error $|J_2(u) - \tilde{J}_2(u_h)|$ behaves like $O(h^{2(p+1)})$ which is even larger than the expected order $O(h^{2p})$, see estimate (370), of an adjoint consistent discretization.

In the following we want to highlight the connection between adjoint consistency and the smoothness of the adjoint solution. To this end, Figure 7a) shows the discrete adjoint solution z_h connected to the (original) target quantity $J_2(\cdot)$; i.e. z_h is the solution to the discrete adjoint problem given at the top of Section 6.3.3 with right hand side $J_2(\cdot)$. In Figure 7a) we see that z_h is irregular in the neighborhood of the boundary. We note that this irregularity does not vanish under mesh refinement. Thereby, the discrete adjoint solution does not converge to the exact solution, $z \equiv -1$, of the continuous adjoint problem (382). This behavior corresponds to the fact that the SIPG discretization in combination with the target quantity $J_2(\cdot)$ is adjoint *in*consistent.

In comparison to that, Figure 7b) shows the discrete adjoint solution z_h connected to the modified target quantity $\tilde{J}_2(\cdot)$. Here, we see that z_h is perfectly smooth. Furthermore, we note that z_h converges to the exact adjoint solution $z \equiv -1$. In fact, we have $z_h \equiv -1$ from the second coarsest mesh onwards. That is, the discrete adjoint solution is a consistent discretization of the continuous adjoint solution. In other words: the SIPG discretization in combination with the modified target quantity $\tilde{J}_2(\cdot)$ is adjoint consistent.

Finally, we recall that the experimental order of convergence of the error $|J_2(u) - \tilde{J}_2(u_h)|$ of the adjoint consistent SIPG discretization behaves like $O(h^{2(p+1)})$ which is two powers of h larger than the theoretically expected order $O(h^{2p})$, see estimate (370). A possible reason for this might be a too simple *model problem* in combination with a particularly simple target quantity which results in the constant continuous adjoint solution $z \equiv -1$.

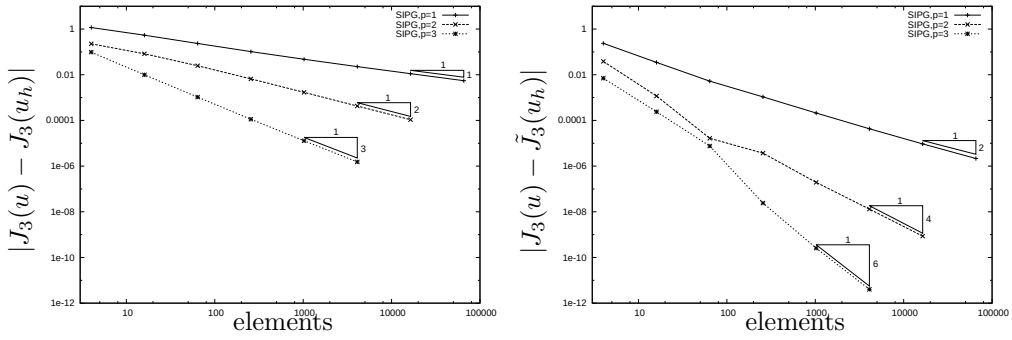
Example 3 In order to demonstrate that the estimate (370) is sharp we consider the following problem: Let $\Omega = (0,1) \times (0.1,1)$ and consider Poisson's equation (162) with forcing function f which is chosen so that the analytical solution to (162) is given by

$$u(\mathbf{x}) = \tfrac{1}{4}(1 + x_1)^2 \sin(2\pi x_1 x_2). \tag{384}$$

We note that this is a modification of the problem considered in [23]. Again, we impose Dirichlet boundary conditions where the boundary value function g_D on $\Gamma_D = \Gamma$ is prescribed based on the solution u. We consider the target quantity $J_3(u_h)$ and its modification $\tilde{J}_3(u_h)$ given as follows

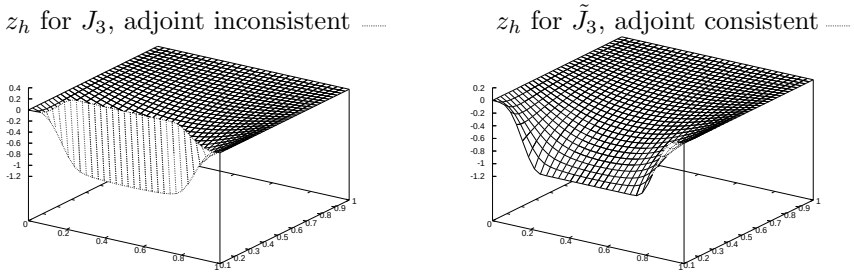
$$J_3(u_h) = \int_\Gamma j_D \, \mathbf{n} \cdot \nabla_h u_h \, \mathrm{d}s, \tag{385}$$

$$\tilde{J}_3(u_h) = J_3(u_h) - \int_\Gamma \delta(u_h - g_D) j_D \, \mathrm{d}s. \tag{386}$$



(a) (b)

Figure 8: Example 3: Convergence of a) the error $|J_3(u) - J_3(u_h)|$ and b) the error $|J_3(u) - \tilde{J}_3(u_h)|$ for the SIPG discretization with global mesh refinement.



(a) (b)

Figure 9: Example 3: Discrete adjoint solution z_h corresponding a) to the adjoint inconsistent SIPG discretization with $J_3(u_h)$ and b) to the adjoint consistent SIPG discretization with $\tilde{J}_3(u_h)$.

and choose $j_D \in L^2(\Gamma)$ to be given by

$$j_D(\mathbf{x}) = \begin{cases} \exp\left(4 - \frac{1}{16}((x_1 - \frac{1}{4})^2 - \frac{1}{8})^{-2}\right) & \text{for } \mathbf{x} \in (0, \frac{1}{4}) \times (0.1, 1), \\ \exp\left(4 - \frac{1}{16}((x_1 - \frac{3}{4})^2 - \frac{1}{8})^{-2}\right) & \text{for } \mathbf{x} \in (\frac{3}{4}, 1) \times (0.1, 1), \\ 1 & \text{for } \mathbf{x} \in (\frac{1}{4}, \frac{3}{4}) \times (0.1, 1), \\ 0 & \text{elsewhere on } \Gamma. \end{cases}$$

Thereby, the true value of the target quantity is $J_3(u) = -1.2825165799606$.

Figure 8a) shows that the convergence behavior of the error $|J_3(u) - J_3(u_h)|$ behaves like $O(h^p)$ which is in perfect agreement with the estimate (371) for an adjoint *in*consistent discretization. Furthermore, Figure 8b) shows that the convergence behavior of the error $|J_3(u) - \tilde{J}_3(u_h)|$ behaves like $O(h^{2p})$ which is as expected, see the estimate (370), for an adjoint consistent discretization.

Finally, Figure 9a) shows the discrete adjoint solution z_h connected to the (original) target quantity $J_3(\cdot)$. We see that in the neighborhood of the bottom boundary $[0, 1] \times \{0.1\} \subset \Gamma$ the discrete adjoint solution is irregular which corresponds to the fact that the SIPG discretization in combination with the target quantity $J_3(\cdot)$ is adjoint inconsistent. In contrast to that the corresponding Figure 9b) shows that the discrete adjoint solution z_h connected to the modified target quantity $\tilde{J}_3(\cdot)$ is entirely smooth which corresponds to the fact that the SIPG discretization in combination with the modified target quantity $\tilde{J}_3(\cdot)$ is adjoint consistent.

8 Discontinuous Galerkin discretizations of the compressible Euler equations

The compressible Euler equations are a nonlinear system of conservation equations (conservation of mass, momentum and energy) describing inviscid compressible flows frequently used as a simple model for gas flows. In order to introduce some new notation we first consider a system of linear hyperbolic equations.

8.1 Hyperbolic conservation equations

Given a final time $T > 0$, we consider the following system of conservation equations,

$$\begin{aligned} \partial_t \mathbf{u} + \sum_{i=1}^{d} \partial_{x_i} \mathbf{f}_i^c(\mathbf{u}) = 0 \qquad & \text{in } (0,T] \times \Omega, \\ \mathbf{u}(0,\cdot) = \mathbf{u}_0(\cdot) \quad & \text{in } \Omega, \end{aligned} \tag{387}$$

where Ω is a bounded connected domain in $\mathbb{R}^d$, $d \geq 1$, $\mathbf{u} = (u_1, \ldots, u_m)^\top$, $\mathcal{F}^c(\mathbf{u}) = (\mathbf{f}_1^c(\mathbf{u}), \ldots, \mathbf{f}_d^c(\mathbf{u}))^\top$ and $\mathbf{f}_i^c : \mathbb{R}^m \to \mathbb{R}^m$, $i = 1, \ldots, d$, are continuously differentiable. In particular, we will be concerned with the solution of the *stationary* system of conservation equations,

$$\nabla \cdot \mathcal{F}^c(\mathbf{u}) = 0 \quad \text{in } \Omega, \tag{388}$$

subject to appropriate boundary conditions described below. We say that (387) is hyperbolic, if the matrix

$$B(\mathbf{u}, \boldsymbol{\nu}) := \sum_{i=1}^{d} \nu_i A_i(\mathbf{u}) \tag{389}$$

has m real eigenvalues and a complete set of linearly independent eigenvectors for all vectors $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d) \in \mathbb{R}^d$. Here, $A_i(\mathbf{u})$ denotes the Jacobi matrix of the flux $\mathbf{f}_i^c(u)$, i.e.

$$A_i(\mathbf{u}) := \partial_{\mathbf{u}} \mathbf{f}_i^c(\mathbf{u}), \qquad i = 1, \ldots, d. \tag{390}$$

The system of conservation equations (387) must be supplemented by appropriate boundary conditions; for example at inflow/outflow boundaries, we require that

$$B^-(\mathbf{u}, \mathbf{n})\,(\mathbf{u} - \mathbf{g}) = 0, \quad \text{on } \Gamma \tag{391}$$

where $\mathbf{n}$ denotes the unit outward normal vector to the boundary $\Gamma = \partial\Omega$ and $\mathbf{g}$ is a (given) vector function. Here, $B^\pm(\mathbf{u}, \mathbf{n})$ denotes the negative/positive part of $B(\mathbf{u}, \mathbf{n})$,

$$B^\pm(\mathbf{u}, \mathbf{n}) = P \Lambda^\pm P^{-1}, \tag{392}$$

where $P = [\mathbf{r}_1, \ldots, \mathbf{r}_m]$ denotes the $m \times m$ matrix of eigenvectors of $B(\mathbf{u}, \mathbf{n})$ and $\Lambda^- = \operatorname{diag}(\min(\lambda_i, 0))$ and $\Lambda^+ = \operatorname{diag}(\max(\lambda_i, 0))$ the $m \times m$ diagonal matrix of the negative/positive eigenvalues of $B(\mathbf{u}, \mathbf{n})$ with $B\mathbf{r}_i = \lambda_i \mathbf{r}_i$, $i = 1, \ldots, d$.

Example 8.1 *The simplest hyperbolic problem is given by the linear advection equation*

$$\nabla \cdot (\mathbf{b} u) = f \quad \text{in } \Omega, \qquad\qquad u = g \quad \text{on } \Gamma_-, \tag{393}$$

i.e. the model problem previously considered in (99) *with vanishing reaction term $c = 0$. In fact, setting $\mathcal{F}^c(u) = (b_1 u, \ldots, b_d u)^\top = \mathbf{b}u$, with a scalar function u, i.e. $m = 1$, we have $B(u, \mathbf{n}) = \mathbf{b} \cdot \mathbf{n} = \lambda \in \mathbb{R}$. The boundary condition* (391) *is given by*

$$B^-(u, \mathbf{n})(u - g) = \lambda^-(u - g) = 0 \quad \text{on } \Gamma,$$

where $\lambda^- = \min(\lambda, 0)$ and reduces to the boundary condition given in (393).

8.2 The compressible Euler equations

The Euler equations of compressible gas dynamics represent an important example of the hyperbolic system (387). We consider the stationary equations in two dimensions given by

$$\nabla \cdot \mathcal{F}^c(\mathbf{u}) = 0 \quad \text{in } \Omega, \tag{394}$$

subject to various boundary conditions; In particular, slip-wall boundary conditions are imposed at solid wall boundaries $\Gamma_W \subset \Gamma$, with vanishing normal velocity, $\mathbf{n} \cdot \mathbf{v} = n_1 v_1 + n_2 v_2 = 0$, i.e.

$$B\mathbf{u} = n_1 u_2 + n_2 u_3 = 0 \quad \text{on } \Gamma_W.$$

Here, the vector of conservative variables $\mathbf{u}$ and the convective flux $\mathcal{F}^c(\mathbf{u}) = (\mathbf{f}_1^c(\mathbf{u}), \mathbf{f}_2^c(\mathbf{u}))^\top$ are defined by

$$\mathbf{u} = \begin{bmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho E \end{bmatrix}, \quad \mathbf{f}_1^c(\mathbf{u}) = \begin{bmatrix} \rho v_1 \\ \rho v_1^2 + p \\ \rho v_1 v_2 \\ \rho H v_1 \end{bmatrix} \text{ and } \mathbf{f}_2^c(\mathbf{u}) = \begin{bmatrix} \rho v_2 \\ \rho v_1 v_2 \\ \rho v_2^2 + p \\ \rho H v_2 \end{bmatrix}, \tag{395}$$

where ρ, $\mathbf{v} = (v_1, v_2)^\top$, p and E denote the density, velocity vector, pressure and specific total energy, respectively. Additionally, H is the total enthalpy given by

$$H = E + \frac{p}{\rho} = e + \tfrac{1}{2}\mathbf{v}^2 + \frac{p}{\rho}, \tag{396}$$

where e is the specific static internal energy, and the pressure is determined by the equation of state of an ideal gas

$$p = (\gamma - 1)\rho e, \tag{397}$$

where $\gamma = c_p/c_v$ is the ratio of specific heat capacities at constant pressure, c_p, and constant volume, c_v; for dry air, $\gamma = 1.4$. The flux Jacobians $A_i(\mathbf{u})$ defined in (390) are given by

$$A_1(\mathbf{u}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -v_1^2 + \frac{1}{2}(\gamma-1)\mathbf{v}^2 & (3-\gamma)v_1 & -(\gamma-1)v_2 & \gamma - 1 \\ -v_1 v_2 & v_2 & v_1 & 0 \\ v_1\left(\frac{1}{2}(\gamma-1)\mathbf{v}^2 - H\right) & H - (\gamma-1)v_1^2 & -(\gamma-1)v_1 v_2 & \gamma v_1 \end{pmatrix},$$

$$A_2(\mathbf{u}) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ -v_1 v_2 & v_2 & v_1 & 0 \\ -v_2^2 + \frac{1}{2}(\gamma-1)\mathbf{v}^2 & -(\gamma-1)v_1 & (3-\gamma)v_2 & \gamma - 1 \\ v_2\left(\frac{1}{2}(\gamma-1)\mathbf{v}^2 - H\right) & -(\gamma-1)v_1 v_2 & H - (\gamma-1)v_2^2 & \gamma v_2 \end{pmatrix}.$$

Finally, the eigenvalues of the matrix $B(\mathbf{u}, \mathbf{n}) = \sum_{i=1}^2 n_i A_i(\mathbf{u})$ are

$$\lambda_1 = \mathbf{v} \cdot \mathbf{n} - c, \quad \lambda_2 = \lambda_3 = \mathbf{v} \cdot \mathbf{n}, \quad \lambda_4 = \mathbf{v} \cdot \mathbf{n} + c \tag{398}$$

where $c = \sqrt{\gamma p/\rho}$ denotes the speed of sound. Considering the signs of λ_i, $i = 1, \ldots, 4$, we distinguish four cases of boundary conditions (391):

- supersonic inflow: $\lambda_i < 0$, $i = 1, \ldots, 4$,
- subsonic inflow: $\lambda_i < 0$, $i = 1, 2, 3$, $\lambda_4 > 0$,
- subsonic outflow: $\lambda_1 < 0$, $\lambda_i > 0$, $i = 2, 3, 4$, and
- supersonic outflow: $\lambda_i > 0$, $i = 1, \ldots, 4$.

Each eigenvalue smaller than zero corresponds to an inflow characteristic. The number of variables to be prescribed on the boundary depend on the number of inflow characteristics.

8.3 The DG discretization of the compressible Euler equations

We begin by introducing the vector-valued counterpart of the discrete function space $V_{h,p}^d$ defined in (123). Let $\mathbf{V}_{h,p}^d$ be the finite element space consisting of discontinuous vector–valued polynomial functions of degree $p \geq 0$, defined by

$$\begin{aligned}\mathbf{V}_{h,p}^d = \{\mathbf{v}_h \in \left[L^2(\Omega)\right]^m : \mathbf{v}_h|_\kappa \circ \sigma_\kappa \in [Q_p(\hat{\kappa})]^m \text{ if } \hat{\kappa} \text{ is the unit hypercube, and} \\ \mathbf{v}_h|_\kappa \circ \sigma_\kappa \in [P_p(\hat{\kappa})]^m \text{ if } \hat{\kappa} \text{ is the unit simplex}, \kappa \in \mathcal{T}_h\},\end{aligned} \tag{399}$$

For deriving discontinuous Galerkin discretizations of the compressible Euler equations we proceed similarly as we did for the case of the linear advection equation in Section 4.7. In order to introduce a weak formulation of (388) we multiply it by an arbitrary smooth (vector-)function $\mathbf{v}$ and integrate by parts over an element κ in the mesh $\mathcal{T}_h$; thereby, we obtain

$$-\int_\kappa \mathcal{F}^c(\mathbf{u}) : \nabla \mathbf{v} \, \mathrm{d}\mathbf{x} + \int_{\partial\kappa} (\mathbf{n} \cdot \mathcal{F}^c(\mathbf{u})) \cdot \mathbf{v} \, \mathrm{d}s = 0. \tag{400}$$

To discretize (400), we replace the analytical solution $\mathbf{u}$ by the Galerkin finite element approximation $\mathbf{u}_h$ and the test function $\mathbf{v}$ by $\mathbf{v}_h$, where $\mathbf{u}_h$ and $\mathbf{v}_h$ both belong to the finite element space $\mathbf{V}_{h,p}^d$. In addition, since the numerical solution $\mathbf{u}_h$ is discontinuous between element interfaces, we must replace the flux $\mathcal{F}^c(\mathbf{u}) \cdot \mathbf{n}$ by a *numerical flux* function $\mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n})$, which depends on both the interior– and outer–trace of $\mathbf{u}_h$ on $\partial\kappa$, $\kappa \in \mathcal{T}_h$, and the unit outward normal $\mathbf{n}$ to $\partial\kappa$. Thereby, summing over the elements κ in the mesh $\mathcal{T}_h$, yields the discontinuous Galerkin discretization of (388) as follows: find $\mathbf{u}_h \in \mathbf{V}_{h,p}^d$ such that

$$-\int_\Omega \mathcal{F}^c(\mathbf{u}_h) : \nabla_h \mathbf{v}_h \, \mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} \mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}) \cdot \mathbf{v}_h^+ \, \mathrm{d}s = 0 \qquad \forall \mathbf{v}_h \in \mathbf{V}_{h,p}^d. \tag{401}$$

We remark that the replacement of the flux $\mathcal{F}^c(\mathbf{u}) \cdot \mathbf{n}$ by the numerical flux function $\mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n})$ on the boundary of element κ, κ in $\mathcal{T}_h$, corresponds to the weak imposition of the boundary data. Like in the case of the linear advection equation in Section 4 the numerical flux $\mathcal{H}(\cdot, \cdot, \cdot)$ must be consistent and conservative. We recall that

(i) $\mathcal{H}(\cdot, \cdot, \cdot)|_{\partial\kappa}$ is consistent with the flux $\mathcal{F}^c(\cdot) \cdot \mathbf{n}$ for each κ in $\mathcal{T}_h$; i.e.

$$\mathcal{H}(\mathbf{v}, \mathbf{v}, \mathbf{n})|_{\partial\kappa} = \mathcal{F}^c(\mathbf{v}) \cdot \mathbf{n} \quad \forall \kappa \in \mathcal{T}_h;$$

(ii) $\mathcal{H}(\cdot, \cdot, \cdot)$ is conservative, i.e. given any two neighboring elements κ and κ' from the finite element partition $\mathcal{T}_h$, at each point $x \in \partial\kappa \cap \partial\kappa' \neq \emptyset$, noting that $\mathbf{n}_{\kappa'} = -\mathbf{n}$, we have that

$$\mathcal{H}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = -\mathcal{H}(\mathbf{w}, \mathbf{v}, -\mathbf{n}).$$

There are several numerical flux functions satisfying these conditions, such as the Godunov, Engquist–Osher, Lax–Friedrichs, Roe or the Vijayasundaram flux. As examples, here we consider three different numerical fluxes:

- The **(local) Lax–Friedrichs flux** $\mathcal{H}_{LF}(\cdot, \cdot, \cdot)$ is defined by

$$\mathcal{H}_{LF}(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n})|_{\partial\kappa} = \tfrac{1}{2}\left(\mathcal{F}^c(\mathbf{u}^+) \cdot \mathbf{n} + \mathcal{F}^c(\mathbf{u}^-) \cdot \mathbf{n} + \alpha\left(\mathbf{u}^+ - \mathbf{u}^-\right)\right),$$

  for $\kappa \in \mathcal{T}_h$, where α is the maximum over $\mathbf{u}^+$ and $\mathbf{u}^-$,

$$\alpha = \max_{\mathbf{v}=\mathbf{u}^+, \mathbf{u}^-} \{|\lambda(B(\mathbf{v}, \mathbf{n}))|\},$$

  of the largest eigenvalue (in absolute value) $|\lambda(B)|$ of the matrix $B(\mathbf{v}, \mathbf{n}) = \sum_{i=0}^d n_i A_i(\mathbf{u})$ defined in (389).

- The **Vijayasundaram flux** $\mathcal{H}_V(\cdot,\cdot,\cdot)$ is defined by

$$\mathcal{H}_V(\mathbf{u}^+,\mathbf{u}^-,\mathbf{n})|_{\partial\kappa} = B^+(\bar{\mathbf{u}},\mathbf{n})\mathbf{u}^+ + B^-(\bar{\mathbf{u}},\mathbf{n})\mathbf{u}^- \quad \text{for } \kappa \in \mathcal{T}_h,$$

  where $B^+(\bar{\mathbf{u}},\mathbf{n})$ and $B^-(\bar{\mathbf{u}},\mathbf{n})$ denote the positive and negative parts, cf. (392), of the matrix $B(\bar{\mathbf{u}},\mathbf{n})$, respectively, evaluated at an average state $\bar{\mathbf{u}}$ between $\mathbf{u}^+$ and $\mathbf{u}^-$.

- The **HLLE flux** $\mathcal{H}_{HLLE}(\cdot,\cdot,\cdot)$ is given by

$$\mathcal{H}_{HLLE}(\mathbf{u}^+,\mathbf{u}^-,\mathbf{n})|_{\partial\kappa} = \tfrac{1}{\lambda^+-\lambda^-}\left(\lambda^+\mathcal{F}^c(\mathbf{u}^+)\cdot\mathbf{n} - \lambda^-\mathcal{F}^c(\mathbf{u}^-)\cdot\mathbf{n} - \lambda^+\lambda^-\left(\mathbf{u}^+-\mathbf{u}^-\right)\right),$$

  where $\lambda^+ = \max(\lambda_{\max},0)$ and $\lambda^- = \min(\lambda_{\min},0)$.

Remark 8.2 *We note that when applied to the linear advection equation* (393)*, most numerical fluxes, in particular the numerical fluxes introduced above, reduce to the upwind flux given in* (115)*:*

$$\mathcal{H}_{uw}(u^+,u^-,\mathbf{n}) = \begin{cases} \mathbf{b}\cdot\mathbf{n}\,u^-, & \text{for } (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) < 0, \text{ i.e. } \mathbf{x}\in\partial\kappa_-, \\ \mathbf{b}\cdot\mathbf{n}\,u^+, & \text{for } (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) \geq 0, \text{ i.e. } \mathbf{x}\in\partial\kappa_+. \end{cases}$$

8.4 Boundary conditions

For boundary faces $\partial\kappa\cap\Gamma\neq\emptyset$ we replace $\mathbf{u}_h^-$ by an appropriate boundary function $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$ which realizes the boundary conditions to be imposed.

First we define several *farfield boundary conditions*:

- Supersonic inflow corresponds to Dirichlet boundary conditions where

$$\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{g}_D = \mathbf{u}_\infty \quad \text{on} \quad \Gamma_{\text{D,sup}}.$$

- Supersonic outflow corresponds to Neumann boundary conditions where

$$\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u} \quad \text{on} \quad \Gamma_{\text{N}}.$$

- The subsonic inflow boundary condition takes the pressure from the flow field and imposes all other variables based on freestream conditions $\mathbf{u}_\infty$, i.e.

$$\mathbf{u}_\Gamma(\mathbf{u}) = \left(\rho_\infty, \rho_\infty v_{1,\infty}, \rho_\infty v_{2,\infty}, \frac{p(\mathbf{u})}{\gamma-1} + \rho_\infty\left(v_{1,\infty}^2 + v_{2,\infty}^2\right)\right)^\top \quad \text{on} \quad \Gamma_{\text{D,sub-in}}.$$

  Here, $p\equiv p(\mathbf{u})$ denotes the pressure evaluated using the equation of state (397).

- The subsonic outflow boundary condition imposes an outflow pressure p_{out} and takes all other variables from the flow field, i.e.

$$\mathbf{u}_\Gamma(\mathbf{u}) = \left(u_1, u_2, u_3, \frac{p_{\text{out}}}{\gamma-1} + \frac{u_2^2+u_3^2}{2u_1}\right)^\top \quad \text{on} \quad \Gamma_{\text{D,sub-out}}.$$

- The characteristic farfield boundary condition imposes Dirichlet boundary conditions based on free-stream conditions on characteristic inflow variables. No boundary conditions are imposed on characteristic outflow variables. This corresponds to using the Vijayasundaram flux on the farfield boundary.

Finally, we define following *wall boundary condition*:

- For slip wall boundary conditions used at reflective walls we set

$$\mathbf{u}_\Gamma(\mathbf{u}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2n_1^2 & -2n_1n_2 & 0 \\ 0 & -2n_1n_2 & 1-2n_2^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \mathbf{u} \quad \text{on} \quad \Gamma_{\text{refl}}, \tag{402}$$

  which originates from $\mathbf{u}$ by inverting the sign of the normal velocity component of $\mathbf{u}$, i.e. $\mathbf{v} = (v_1, v_2)$ is replaced by $\mathbf{v}^- = \mathbf{v} - 2(\mathbf{v}\cdot\mathbf{n})\mathbf{n}$. This choice ensures a vanishing average normal velocity, $\bar{\mathbf{v}}\cdot\mathbf{n} = \frac{1}{2}(\mathbf{v}+\mathbf{v}^-)\cdot\mathbf{n} = 0$.

Given the boundary value function $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$ as defined above the DG discretization of (388) including boundary conditions is given as follows: find $\mathbf{u}_h \in \mathbf{V}_{h,p}^d$ such that

$$N_h(\mathbf{u}_h, \mathbf{v}_h) \equiv -\int_\Omega \mathcal{F}^c(\mathbf{u}_h) : \nabla_h \mathbf{v}_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} \mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n})\cdot\mathbf{v}_h^+ \,\mathrm{d}s + \int_\Gamma \mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n})\cdot\mathbf{v}_h^+ \,\mathrm{d}s = 0 \quad \mathbf{v}_h \in \mathbf{V}_{h,p}^d, \tag{403}$$

where $\mathcal{H}_\Gamma$ is usually the same numerical flux $\mathcal{H}$ as used on interior faces $\partial\kappa\setminus\Gamma$, $\kappa\in\mathcal{T}_h$.

8.5 Consistency and adjoint consistency for nonlinear problems

In Section 6.1 we introduced the consistency and adjoint consistency analysis for linear problems. In this section we now give the generalization of this analysis to nonlinear problems of the form:

$$Nu = 0 \quad \text{in } \Omega, \qquad\qquad Bu = 0 \quad \text{on } \Gamma, \tag{404}$$

where N is a nonlinear differential (and Fréchet-differentiable) operator and B is a (possibly nonlinear) boundary operator. Let $J(\cdot)$ be a nonlinear target functional

$$J(u) = \int_\Omega j_\Omega(u)\,\mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma(Cu)\,\mathrm{d}s, \tag{405}$$

with Fréchet derivative

$$J'[u](w) = \int_\Omega j_\Omega'[u]\,w\,\mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma'[Cu]\,C'[u]w\,\mathrm{d}s, \tag{406}$$

where $j_\Omega(\cdot)$ and $j_\Gamma(\cdot)$ may be nonlinear with derivatives j_Ω' and j_Γ', respectively, and C is a differential boundary operator on Γ and may be nonlinear with derivative C'. Here, $'$ denotes the (total) Fréchet derivative and the square bracket $[\cdot]$ denotes the state about which linearization is performed. Again, we say that the target functional (405) is *compatible* with (404) provided the following compatibility condition holds

$$(N'[u]w, z)_\Omega + (B'[u]w, (C'[u])^* z)_\Gamma = (w, (N'[u])^* z)_\Omega + (C'[u]w, (B'[u])^* z)_\Gamma, \tag{407}$$

where $(N'[u])^*$, $(B'[u])^*$ and $(C'[u])^*$ denote the adjoint operators to $N'[u]$, $B'[u]$ and $C'[u]$. This condition is analogous to (275), with L, B and C replaced by $N'[u]$, $B'[u]$ and $C'[u]$, respectively. *Assuming* that (407) holds the continuous adjoint problem associated to (404) and (406) is:

$$(N'[u])^* z = j_\Omega'[u] \quad \text{in } \Omega, \qquad\qquad (B'[u])^* z = j_\Gamma'[Cu] \quad \text{on } \Gamma. \tag{408}$$

We note that in an optimization framework [20] this ensures, analogous to (277), that

$$\begin{aligned} J'[u](w) &= (w, j'_\Omega[u])_\Omega + (C'[u]w, j'_\Gamma[Cu])_\Gamma = (w, (N'[u])^* z)_\Omega + (C'[u]w, (B'[u])^* z)_\Gamma \\ &= (N'[u]w, z)_\Omega + (B'[u]w, (C'[u])^* z)_\Gamma. \end{aligned} \tag{409}$$

Let $N_h : V \times V \to \mathbb{R}$ be a semi-linear form, nonlinear in its first and linear in its second argument, such that the nonlinear problem (404) is discretized as follows: find $u_h \in V_h$ such that

$$N_h(u_h, v_h) = 0 \quad \forall v_h \in V_h. \tag{410}$$

Then, the discretization (410) is said to be *consistent* if the exact solution $u \in V$ to the primal problem (404) satisfies the following equation:

$$N_h(u, v) = 0 \quad \forall v \in V. \tag{411}$$

Furthermore, the discretization (410) is said to be *adjoint consistent* if the exact solutions $u, z \in V$ to the primal and adjoint problems (404) and (408), respectively, satisfy the following equation:

$$N'_h[u](w, z) = J'[u](w) \quad \forall w \in V, \tag{412}$$

where $N'_h[u]$ denotes the Fréchet derivatives of $N_h(u, v)$ with respect to u.

In other words, a discretization is adjoint consistent if the discrete adjoint problem is a consistent discretization of the continuous adjoint problem. Finally, we note that in case of a linear problem and target functional the definition of adjoint consistency in (412) reduces to the definition of linear adjoint consistency given in Section 6. The definition of adjoint consistency for nonlinear problems as given in (412) was introduced by Lu [36]. Furthermore, we note that [36] also gives a definition of asymptotically adjoint consistent methods.

8.5.1 The consistency and adjoint consistency analysis

Based on the definition of consistency and adjoint consistency in the previous subsection we now follow [27] and generalize the framework for analyzing consistency and adjoint consistency of discontinuous Galerkin discretizations for linear problems as given in Section 6 to the case of nonlinear problems. We recall that this framework can also be used to find specific terms due to which some DG discretizations may not be adjoint consistent. In these cases the analysis gives some insight into how an adjoint inconsistent DG discretization together with a specific target functional could be modified to recover an adjoint consistent discretization.

Given a primal problem, a discontinuous Galerkin discretization of the problem and a target functional, the adjoint consistency analysis consists of the following steps:

- **Derivation of the continuous adjoint problem:** Let the primal problem be given by (404). Furthermore, assume that $J(\cdot)$ is a nonlinear functional (405) which is compatible with the primal problem (404). Then we derive the continuous adjoint problem (408) including adjoint boundary conditions.

  We note that the derivation of the adjoint operator $(N'[u])^*$ for nonlinear systems is a considerably more complicated task than deriving L^* for scalar linear problems. Still more involved is the derivation of the adjoint boundary operators $(B'[u])^*$. In the framework of optimal design, [20] gives a general approach of deriving $(B'[u])^*$ and $(C'[u])^*$ assumed to be connect to B, C, N and $(N'[u])^*$ through (407). This approach is based on a matrix representation of boundary operators which for systems of equations leads to lengthy and error prone derivations. In contrast to optimization where both $(B'[u])^*$ and C^* are required, in the following analysis we require only the adjoint operator $(B'[u])^*$. Due to this we can circumvent the approach described in [20] and use a simpler way of deriving the adjoint operators $(B'[u])^*$.

- **Consistency analysis of the discrete primal problem:** We rewrite the discontinuous Galerkin discretization (410) of problem (404) in the following element-based primal residual form: find $u_h \in V_h$ such that

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa R(u_h)v_h\,\mathrm{d}\mathbf{x}+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} r(u_h)v_h\,\mathrm{d}s+\int_\Gamma r_\Gamma(u_h)v_h\,\mathrm{d}s=0\quad\forall v_h\in V_h, \tag{413}$$

where $R(u_h)$, $r(u_h)$ and $r_\Gamma(u_h)$ denote the element, interior face and boundary residuals, respectively. According to (411), the discretization (410) is consistent if the exact solution u to (404) satisfies

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa R(u)v\,\mathrm{d}\mathbf{x}+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} r(u)v\,\mathrm{d}s+\int_\Gamma r_\Gamma(u)v\,\mathrm{d}s=0\quad\forall v\in V, \tag{414}$$

which holds provided u satisfies

$$R(u)=0\ \text{ in }\kappa,\kappa\in\mathcal{T}_h,\qquad r(u)=0\ \text{ on }\partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h,\qquad r_\Gamma(u)=0\ \text{ on }\Gamma. \tag{415}$$

- **Derivation of the discrete adjoint problem** Given the discretization (410), the target functional (405) and its linearization (406), we derive the discrete adjoint problem: find $z_h \in V_h$ such that

$$\mathcal{N}'[u_h](w_h,z_h)=J'[u_h](w_h)\quad\forall w_h\in V_h. \tag{416}$$

$\mathcal{N}'[u_h]$ is called the Jacobian of the numerical scheme and is required also for implicit and adjoint methods, e.g. Newton iteration, a posteriori error estimation, adjoint-based adaptation, see [25], and for optimization.

- **Adjoint consistency of element, interior face and boundary terms** We rewrite the discrete adjoint problem (416) in element-based adjoint residual form: find $z_h \in V_h$ such that

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa w_h\,R^*[u_h](z_h)\,\mathrm{d}\mathbf{x}+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} w_h\,r^*[u_h](z_h)\,\mathrm{d}s+\int_\Gamma w_h\,r^*_\Gamma[u_h](z_h)\,\mathrm{d}s=0, \tag{417}$$

for all $w_h \in V_h$, where $R^*[u_h](z_h)$, $r^*[u_h](z_h)$ and $r^*_\Gamma[u_h](z_h)$ denote the element, interior face and boundary adjoint residuals, respectively. According to (412), the discretization (410) is adjoint consistent if the exact solutions u and z satisfy

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa w\,R^*[u](z)\,\mathrm{d}\mathbf{x}+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} w\,r^*[u](z)\,\mathrm{d}s+\int_\Gamma w\,r^*_\Gamma[u](z)\,\mathrm{d}s=0\quad\forall w\in V, \tag{418}$$

which holds provided u and z satisfy

$$R^*[u](z)=0\ \text{ in }\kappa,\qquad r^*[u](z)=0\ \text{ on }\partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h,\qquad r^*_\Gamma[u](z)=0\ \text{ on }\Gamma. \tag{419}$$

We note that the adjoint problem and consequently the adjoint consistency of a discretization depends on the specific target functional $J(\cdot)$ under consideration. Given a target functional of the form (405), we see that $R^*[u](z)$ depends on $j_\Omega(\cdot)$, and $r^*_\Gamma[u](z)$ depends on $j_\Gamma(\cdot)$. For obtaining an adjoint consistent discretization it is in some cases, see following Sections, necessary to modify the target functional as follows

$$\tilde{J}(u_h)=J(i(u_h))+\int_\Gamma r_J(u_h)\,\mathrm{d}s, \tag{420}$$

where $i(\cdot)$ and $r_J(\cdot)$ are functions to be specified. We recall from Section 6.2 that a modification of a target functional is called *consistent* if $\tilde{J}(u)=J(u)$ holds for the exact solution u.

8.6 Adjoint consistency analysis of DG for the compressible Euler equations

8.6.1 The continuous adjoint problem to the compr. Euler equations

The most important target quantities in inviscid compressible flows are the pressure induced drag and lift coefficients, c_{dp} and c_{lp}, defined by

$$J(\mathbf{u}) = \int_\Gamma j(\mathbf{u})\,\mathrm{d}s = \frac{1}{C_\infty}\int_{\Gamma_W} p\,\mathbf{n}\cdot\boldsymbol{\psi}\,\mathrm{d}s, \tag{421}$$

where $j(\mathbf{u}) = \frac{1}{C_\infty} p\,\mathbf{n}\cdot\boldsymbol{\psi}$ on Γ_W and $j(\mathbf{u}) \equiv 0$ elsewhere. Here, $C_\infty = \frac{1}{2}\gamma p_\infty M_\infty^2 \bar{l} = \frac{1}{2}\gamma \frac{|\mathbf{v}_\infty|^2}{c_\infty^2} p_\infty \bar{l} = \frac{1}{2}\rho_\infty |\mathbf{v}_\infty|^2 \bar{l}$, where M denotes the Mach number, c the sound speed defined by $c^2 = \gamma p/\rho$, $\bar{l}$ denotes a reference length, and $\boldsymbol{\psi}$ is given by $\boldsymbol{\psi}_d = (\cos(\alpha), \sin(\alpha))^\top$ or $\boldsymbol{\psi}_l = (-\sin(\alpha), \cos(\alpha))^\top$ for the drag and lift coefficient, respectively. Subscripts $_\infty$ indicate free-stream quantities.

In order to derive the continuous adjoint problem, we multiply the left hand side of (394) by $\mathbf{z}$, integrate by parts and linearize about $\mathbf{u}$ to obtain

$$\left(\nabla\cdot\left(\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}](\mathbf{w})\right), \mathbf{z}\right)_\Omega = -\left(\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}](\mathbf{w}), \nabla\mathbf{z}\right)_\Omega + \left(\mathbf{n}\cdot\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}](\mathbf{w}), \mathbf{z}\right)_\Gamma, \tag{422}$$

where $\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}] := (\mathcal{F}^c)'[\mathbf{u}]$ denotes the Fréchet derivative of $\mathcal{F}^c$ with respect to $\mathbf{u}$. Here, we already use the subscript $_{\mathbf{u}}$ notation, which we require in Section 9 to distinguish from subscript $_{\nabla\mathbf{u}}$ denoting the derivative with respect to $\nabla\mathbf{u}$. Thereby, the variational formulation of the continuous adjoint problem is given by: find $\mathbf{z}$ such that

$$-\left(\mathbf{w}, \left(\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}]\right)^\top \nabla\mathbf{z}\right)_\Omega + \left(\mathbf{w}, \left(\mathbf{n}\cdot\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}]\right)^\top \mathbf{z}\right)_\Gamma = J'[\mathbf{u}](\mathbf{w}) \quad \forall\mathbf{w}\in V, \tag{423}$$

and the continuous adjoint problem is given by

$$-\left(\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}]\right)^\top \nabla\mathbf{z} = 0 \quad \text{in } \Omega, \qquad\qquad \left(\mathbf{n}\cdot\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}]\right)^\top \mathbf{z} = j'[\mathbf{u}] \quad \text{on } \Gamma. \tag{424}$$

Using $\mathcal{F}^c(\mathbf{u})\cdot\mathbf{n} = p(0, n_1, n_2, 0)^\top$ on Γ_W, and the definition of j in (421) we obtain

$$p'[\mathbf{u}](0, n_1, n_2, 0)\cdot\mathbf{z} = \frac{1}{C_\infty} p'[\mathbf{u}]\mathbf{n}\cdot\boldsymbol{\psi} \quad \text{on } \Gamma_W,$$

which reduces to the boundary condition of the adjoint compressible Euler equations,

$$(B'[\mathbf{u}])^*\mathbf{z} = n_1 z_2 + n_2 z_3 = \frac{1}{C_\infty}\mathbf{n}\cdot\boldsymbol{\psi} \quad \text{on } \Gamma_W. \tag{425}$$

8.6.2 Primal residual form of DG for the compr. Euler equations

Using integration by parts on (403) we obtain the residual form: find $\mathbf{u}_h \in \mathbf{V}^d_{h,p}$ such that

$$\int_\Omega \mathbf{R}(\mathbf{u}_h)\cdot\mathbf{v}_h\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathbf{r}(\mathbf{u}_h)\cdot\mathbf{v}_h^+\,\mathrm{d}s + \int_\Gamma \mathbf{r}_\Gamma(\mathbf{u}_h)\cdot\mathbf{v}_h^+\,\mathrm{d}s = 0 \quad \forall\mathbf{v}_h\in\mathbf{V}^d_{h,p}, \tag{426}$$

where the primal residuals are given by

$$\begin{aligned} \mathbf{R}(\mathbf{u}_h) &= -\nabla\cdot\mathcal{F}^c(\mathbf{u}_h) && \text{in } \kappa, \kappa\in\mathcal{T}_h,\\ \mathbf{r}(\mathbf{u}_h) &= \mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+) && \text{on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h,\\ \mathbf{r}_\Gamma(\mathbf{u}_h) &= \mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+) && \text{on } \Gamma. \end{aligned} \tag{427}$$

Given the consistency of the numerical flux, $\mathcal{H}(\mathbf{w}, \mathbf{w}, \mathbf{n}) = \mathbf{n}\cdot\mathcal{F}^c(\mathbf{w})$, and the consistency of the boundary function, i.e. $\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ for the exact solution $\mathbf{u}$ to (394), we find that $\mathbf{u}$ satisfies following equations

$$\mathbf{R}(\mathbf{u}) = 0 \ \text{ in } \kappa, \kappa\in\mathcal{T}_h, \qquad \mathbf{r}(\mathbf{u}) = 0 \ \text{ on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h, \qquad \mathbf{r}_\Gamma(\mathbf{u}) = 0 \ \text{ on } \Gamma. \tag{428}$$

We conclude that (403) is a consistent discretization of (394).

8.6.3 Adjoint residual form of DG for the compr. Euler equations

For the target functional $J(\cdot)$ defined in (421) with Fréchet derivative, $J'[\mathbf{u}](\cdot)$, the discrete adjoint problem is given by: find $\mathbf{z}_h \in \mathbf{V}_{h,p}^d$ such that

$$N_h'[\mathbf{u}_h](\mathbf{w}_h, \mathbf{z}_h) = J'[\mathbf{u}_h](\mathbf{w}_h) \quad \forall \mathbf{w}_h \in \mathbf{V}_{h,p}^d, \tag{429}$$

where

$$\begin{aligned} N_h'[\mathbf{u}_h](\mathbf{w}, \mathbf{z}_h) \equiv & -\int_\Omega (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_h]\mathbf{w}) : \nabla_h \mathbf{z}_h \,\mathrm{d}\mathbf{x} \\ & + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} \left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\mathbf{w}^+ + \mathcal{H}'_{\mathbf{u}^-}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\mathbf{w}^- \right) \mathbf{z}_h^+ \,\mathrm{d}s \\ & + \int_\Gamma \left(\mathcal{H}'_{\Gamma,\mathbf{u}^+}\left(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+\right) + \mathcal{H}'_{\Gamma,\mathbf{u}^-}\left(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+\right) \mathbf{u}_\Gamma'[\mathbf{u}_h^+] \right) \mathbf{w}^+ \mathbf{z}_h^+ \,\mathrm{d}s. \end{aligned} \tag{430}$$

Here $\mathbf{v} \to \mathcal{H}'_{\mathbf{u}^+}(\mathbf{v}^+, \mathbf{v}^-, \mathbf{n})$ and $\mathbf{v} \to \mathcal{H}'_{\mathbf{u}^-}(\mathbf{v}^+, \mathbf{v}^-, \mathbf{n})$ denote the derivatives of the flux function $\mathcal{H}(\cdot,\cdot,\cdot)$ with respect to its first and second arguments, respectively. As the numerical flux is conservative, $\mathcal{H}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = -\mathcal{H}(\mathbf{w}, \mathbf{v}, -\mathbf{n})$, we obtain $\mathcal{H}'_{\mathbf{u}^-}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = \partial_{\mathbf{w}}\mathcal{H}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = -\partial_{\mathbf{w}}\mathcal{H}(\mathbf{w}, \mathbf{v}, -\mathbf{n}) = -\mathcal{H}'_{\mathbf{u}^+}(\mathbf{w}, \mathbf{v}, -\mathbf{n})$, and

$$\begin{aligned} \int_{\Gamma_\mathcal{I}} \mathcal{H}'_{\mathbf{u}^-}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\mathbf{w}^-\mathbf{z}^+ \,\mathrm{d}s &= -\int_{\Gamma_\mathcal{I}} \mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^-, \mathbf{u}_h^+, \mathbf{n}^-)\mathbf{w}^-\mathbf{z}^+ \,\mathrm{d}s \\ &= -\int_{\Gamma_\mathcal{I}} \mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\mathbf{w}^+\mathbf{z}^- \,\mathrm{d}s, \end{aligned} \tag{431}$$

where we exchanged notations $^+$ and $^-$ on $\Gamma_\mathcal{I}$. Then, the discrete adjoint problem (429) with (430) is given in adjoint residual form as follows: find $\mathbf{z}_h \in \mathbf{V}_{h,p}^d$ such that

$$\int_\Omega \mathbf{w}_h \cdot \mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h) \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} \mathbf{w}_h^+ \cdot \mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) \,\mathrm{d}s + \int_\Gamma \mathbf{w}_h^+ \cdot \mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h) \,\mathrm{d}s = 0, \tag{432}$$

for all $\mathbf{w}_h \in \mathbf{V}_{h,p}^d$, where the adjoint residuals are given by

$$\begin{aligned} \mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h) =& (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_h])^\top \nabla \mathbf{z}_h && \text{in } \kappa, \kappa \in \mathcal{T}_h, \\ \mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) =& -\left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\right)^\top [\![\mathbf{z}_h]\!] \cdot \mathbf{n} && \text{on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \\ \mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h) =& j'[\mathbf{u}_h] - \left(\mathcal{H}'_{\Gamma,\mathbf{u}^+} + \mathcal{H}'_{\Gamma,\mathbf{u}^-}\mathbf{u}_\Gamma'[\mathbf{u}_h]\right)^\top \mathbf{z}_h^+ && \text{on } \Gamma, \end{aligned} \tag{433}$$

where $\mathcal{H}'_{\Gamma,\mathbf{u}^+} := \mathcal{H}'_{\Gamma,\mathbf{u}^+}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+)$ and $\mathcal{H}'_{\Gamma,\mathbf{u}^-} := \mathcal{H}'_{\Gamma,\mathbf{u}^-}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+)$.

Comparing the discrete adjoint boundary condition

$$\left(\mathcal{H}'_{\Gamma,\mathbf{u}^+} + \mathcal{H}'_{\Gamma,\mathbf{u}^-}\mathbf{u}_\Gamma'[\mathbf{u}_h]\right)^\top \mathbf{z}_h^+ = j'[\mathbf{u}_h] \quad \text{on } \Gamma, \tag{434}$$

and the continuous adjoint boundary condition in (424), we notice that not all choices of $\mathcal{H}_\Gamma$ give rise to an adjoint consistent discretization. In fact, we require $\mathcal{H}_\Gamma$ to have following properties: In order to incorporate boundary conditions in the primal discretization (403), $\mathcal{H}_\Gamma$ must depend on $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$, hence $\mathcal{H}'_{\Gamma,\mathbf{u}^-} \neq 0$. Furthermore, we require $\mathcal{H}'_{\Gamma,\mathbf{u}^+} = 0$, as otherwise the left hand side in (434) involves two summands which is in contrast to the continuous adjoint boundary condition in (424). Finally, we recall that $\mathcal{H}_\Gamma$ is consistent, $\mathcal{H}_\Gamma(\mathbf{v}, \mathbf{v}, \mathbf{n}) = \mathbf{n} \cdot \mathcal{F}^c(\mathbf{v})$, and conclude that

$\mathcal{H}_\Gamma$ is given by $\mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}) = \mathbf{n} \cdot \mathcal{F}^c(\mathbf{u}_\Gamma(\mathbf{u}_h^+))$. Employing a modified target functional $\tilde{J}(\mathbf{u}_h) = J(\mathbf{i}(\mathbf{u}_h))$, i.e. (288) with $r_j(\mathbf{u}_h) \equiv 0$, (434) yields

$$\left(\mathbf{n} \cdot \left(\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_\Gamma(\mathbf{u}_h^+)]\right) \mathbf{u}_\Gamma'[\mathbf{u}_h^+]\right)^\top \mathbf{z} = j'[\mathbf{i}(\mathbf{u}_h^+)]\mathbf{i}'[\mathbf{u}_h^+]. \tag{435}$$

We find the modification $\mathbf{i}(\mathbf{u}_h) = \mathbf{u}_\Gamma(\mathbf{u}_h)$ which is consistent as $\mathbf{i}(\mathbf{u}) = \mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ holds for the exact solution $\mathbf{u}$. Thereby (435) reduces to

$$\left(\mathbf{n} \cdot \mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_\Gamma(\mathbf{u}_h^+)]\right)^\top \mathbf{z} = j'[\mathbf{u}_\Gamma(\mathbf{u}_h^+)], \tag{436}$$

which represents a discretization of the continuous adjoint boundary condition in (424). In order to obtain a discretization of the adjoint boundary condition at solid wall boundaries (425), we require $B\mathbf{u}_\Gamma(\mathbf{u}_h^+) = 0$ on Γ_W. This condition is satisfied by

$$\mathbf{u}_\Gamma(\mathbf{u}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-n_1^2 & -n_1 n_2 & 0 \\ 0 & -n_1 n_2 & 1-n_2^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \mathbf{u} \quad \text{on} \quad \Gamma_W, \tag{437}$$

which originates from $\mathbf{u}$ by subtracting the normal velocity component of $\mathbf{u}$, i.e. $\mathbf{v} = (v_1, v_2)$ is replaced by $\mathbf{v}_\Gamma = \mathbf{v} - (\mathbf{v}\cdot\mathbf{n})\mathbf{n}$ which ensures that the normal velocity component vanishes, $\mathbf{v}_\Gamma \cdot \mathbf{n} = 0$.

In summary, let $\mathbf{u}_\Gamma$ be given by (437) and $\mathcal{H}_\Gamma$ and $\tilde{J}$ be defined by

$$\mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}) = \mathbf{n} \cdot \mathcal{F}_\Gamma^c(\mathbf{u}_h^+), \qquad \tilde{J}(\mathbf{u}_h) = J_\Gamma(\mathbf{u}_h), \tag{438}$$

where $\mathcal{F}_\Gamma^c(\mathbf{u}_h^+) := \mathcal{F}^c(\mathbf{u}_\Gamma(\mathbf{u}_h^+))$, $J_\Gamma(\mathbf{u}_h) := J(\mathbf{u}_\Gamma(\mathbf{u}_h))$ and $j_\Gamma(\mathbf{u}_h) := j(\mathbf{u}_\Gamma(\mathbf{u}_h))$, then the adjoint residuals (433) are given by:

$$\begin{aligned}
\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h) =& (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_h])^\top \nabla \mathbf{z}_h && \text{in } \kappa, \kappa \in \mathcal{T}_h, \\
\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) =& -\left(\mathcal{H}_{\mathbf{u}^+}'(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\right)^\top [\![\mathbf{z}_h]\!] \cdot \mathbf{n} && \text{on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \\
\mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h) =& j_\Gamma'[\mathbf{u}_h^+] - \left(\mathbf{n} \cdot \mathcal{F}_{\Gamma,\mathbf{u}}^c[\mathbf{u}_h^+]\right)^\top \mathbf{z}_h^+ && \text{on } \Gamma.
\end{aligned} \tag{439}$$

In particular, the discretization (403) together with (438) is adjoint consistent as the exact solutions $\mathbf{u}$ and $\mathbf{z}$ to (394) and (424), respectively, satisfy

$$\mathbf{R}^*[\mathbf{u}](\mathbf{z}) = 0 \ \text{ in } \kappa, \kappa \in \mathcal{T}_h, \qquad \mathbf{r}^*[\mathbf{u}](\mathbf{z}) = 0 \ \text{ on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \qquad \mathbf{r}_\Gamma^*[\mathbf{u}](\mathbf{z}) = 0 \ \text{ on } \Gamma.$$

Note, that the adjoint residuals in (439) reduce to the adjoint residuals of the linear advection equation with $b = 0$ in Section 6.4.3, when setting $\mathcal{F}^c(u) = \mathbf{b}u$ and $\mathcal{H}_{u^+}' = \mathbf{b} \cdot \mathbf{n}$.

Also note, that the standard discontinuous Galerkin discretizations for the compressible Euler equations take the same numerical flux function on the boundary Γ as in the interior of the domain, and simply replace $\mathbf{u}_h^-$ in $\mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n})$ by the boundary function $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$ resulting in $\mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n})$. Furthermore, the definition of $\mathbf{u}_\Gamma$ in (402) based on $\mathbf{v}_\Gamma = \mathbf{v} - 2(\mathbf{v}\cdot\mathbf{n})\mathbf{n}$ ensures a vanishing average normal velocity, $\bar{\mathbf{v}}\cdot\mathbf{n} = \frac{1}{2}(\mathbf{v} + \mathbf{v}_\Gamma)\cdot\mathbf{n} = 0$. However, $\mathbf{v}_\Gamma \cdot \mathbf{n} = 0$ and $B\mathbf{u}_\Gamma(\mathbf{u}_h^+) = 0$, as required in (436), is not satisfied. Thereby, the discontinuous Galerkin discretization based on the standard choice of $\mathcal{H}_\Gamma$ and $\mathbf{u}_\Gamma$ is not adjoint consistent.

In fact, numerical experiments indicated large gradients i.e. an irregular adjoint solution near solid wall boundaries. The lack of adjoint consistency of this standard approach was first analyzed by [36] who also proposed the adjoint consistent approach (438) and demonstrated that this approach gives rise to smooth adjoint solutions for an inviscid compressible flow over a Gaussian bump. The smoothness of the discrete adjoint has been confirmed in [26] for an inviscid compressible flow around a NACA0012 airfoil, see also Section 8.7. Furthermore, [26] studies the effect of adjoint consistency on the accuracy of the flow solution and on error cancellation in an *a posteriori* error estimation approach.

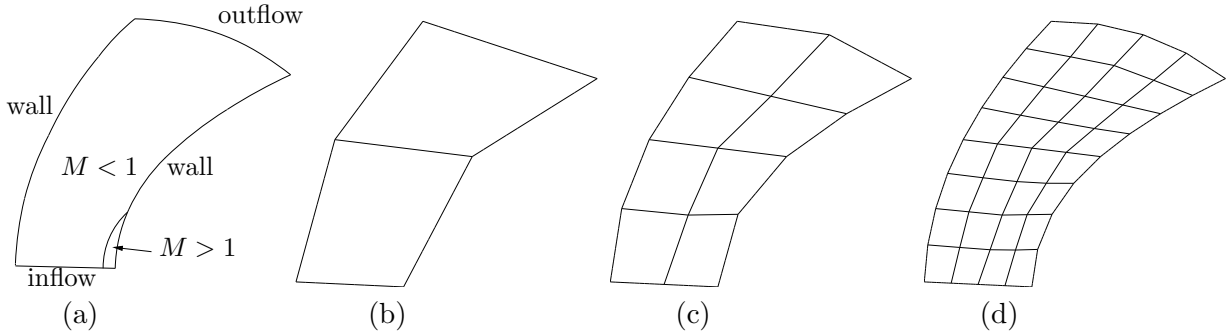


Figure 10: Ringleb flow problem: a) Regions of sub- and supersonic flow denoted by the Mach number $M < 1$ and $M > 1$; b)-d) Coarse meshes with 2, 8 and 32 elements, respectively.

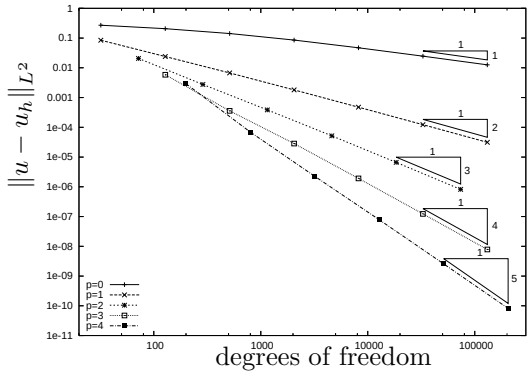


Figure 11: Ringleb flow problem: The L^2-error of the DG(p), $p = 0, \ldots, 4$, discretizations of the compressible Euler equations is of order $\mathcal{O}(h^{p+1})$, [24].

8.7 Numerical results

Ringleb flow problem For discretizations of the 2d stationary compressible Euler equations there are virtually no *a priori* error estimates available. Therefore, in the following we examine the order of convergence of the DG discretization experimentally. In particular, we consider the solution to the 2d compressible Euler equations for the Ringleb flow problem. This is one of the few non-trivial problems of the 2d Euler equations for which a smooth analytical solution is known. For this problem the analytical solution may be obtained be employing the hodograph transformation, see [15] or the appendix of [24]. This problem represents a transonic flow in a channel, see Figure 10a), with inflow and outflow boundaries given by the lower and upper boundaries of the domain, and reflective (slip wall) boundaries with vanishing normal velocity, $\mathbf{v} \cdot \mathbf{n} = 0$, on the left and right boundary. The solution to this flow problem is smooth but it is transonic with a small supersonic region near the lower right corner. The computational domain is subdivided into quadrilateral elements. Figure 10 shows the coarsest three meshes in a sequence of globally refined meshes. In order to suppress the discretization effects of slip wall boundaries here we impose the boundary condition, $B^-(\mathbf{u}, \mathbf{n})(\mathbf{u} - \mathbf{g}) = \mathbf{0}$ on the whole boundary Γ of the domain, where $\mathbf{g}$ is the boundary value function taken from the exact solution to the Ringleb flow problem. This boundary condition represents an inflow boundary condition for characteristic variables on inflow parts (with respect to the corresponding characteristics) of the boundary. Figure 11, by [24], plots the $L^2(\Omega)$-error of the DG(p), $0 \leq p \leq 4$, solutions against the number of degrees of freedom (DoFs) on the sequence of globally refined meshes. We observe an experimental order $O(h^{p+1})$ of convergence which is optimal for polynomial trial and test functions of degree p.

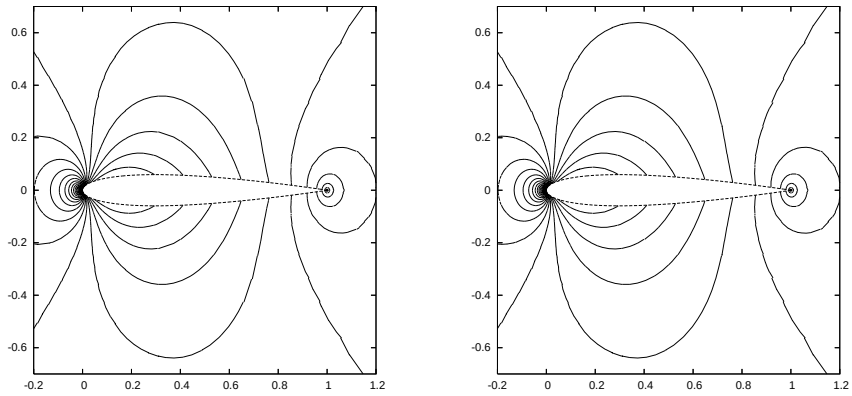


Figure 12: $M = 0.5, \alpha = 0^\circ$ inviscid flow around the NACA0012 airfoil: Mach isolines of the flow solution $\mathbf{u}_h$ to (left) the standard and (right) the adjoint consistent DG discretization, [28].

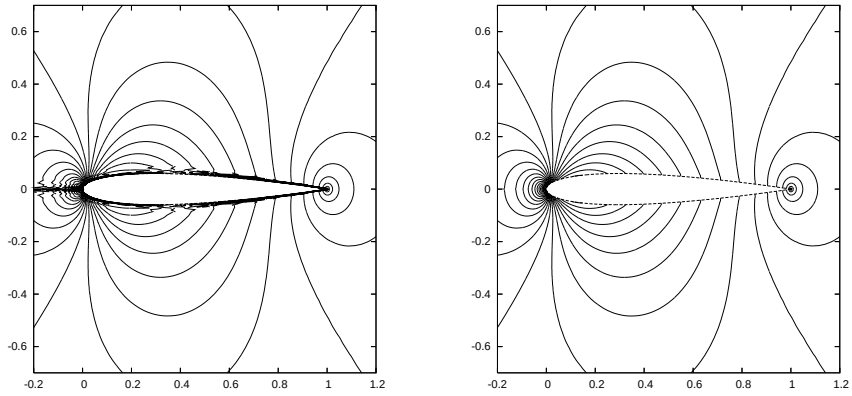


Figure 13: $M = 0.5, \alpha = 0^\circ$ inviscid flow around the NACA0012 airfoil: z_1 isolines of the discrete adjoint solution $\mathbf{z}_h$ to (left) the standard and (right) the adjoint consistent DG discretization, [28].

Inviscid flow around the NACA0012 airfoil In the following, we investigate the smoothness of the discrete adjoint solution when employing the adjoint consistent discretization based on (438) in comparison to the standard (classical) approach of choosing $\mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}) = \mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n})$ and an unmodified target functional $J(\mathbf{u}_h)$. To this end, we consider an inviscid Mach $M = 0.5$ flow at a zero angle of attack, i.e. $\alpha = 0^\circ$, around the NACA0012 airfoil. Here, the upper and lower surfaces of the airfoil geometry are specified by the function $g^\pm$, respectively, where

$$g^\pm(s) = \pm 5 \times 0.12 \times (0.2969 s^{1/2} - 0.126 s - 0.3516 s^2 + 0.2843 s^3 - 0.1015 s^4).$$

As the chord length l of the airfoil is $l \approx 1.00893$ we use a rescaling of g in order to yield an airfoil of unit (chord) length. The computational domain Ω is subdivided into quadrilateral elements. Curved boundaries are approximated by piecewise quadratic polynomials. In Figure 12, by [28], we compare the (primal) flow solutions $\mathbf{u}_h \in \mathbf{V}_h^1$ for the standard and the adjoint consistent DG discretizations and find no visible difference. However, when comparing the adjoint solutions corresponding to the pressure induced drag coefficient c_{dp}, see Figure 13, we notice that the discrete adjoint solution to the standard DG discretization is irregular near and upstream the airfoil. In contrast to that, the adjoint solution to the adjoint consistent discretization is entirely smooth. Furthermore, in [28] it has been shown that for this test case on a sequence of locally refined meshes the error in the c_{dp} value for the adjoint consistent discretization is by a factor 1.3-2.4 smaller than for the adjoint inconsistent discretization.

9 DG discretizations of the compressible Navier-Stokes equations

The compressible Euler equations as considered in the last section serve as a simple model for gas flows. In fact, while ignoring all viscous effects they describe an inviscid compressible flow. In the following, we will enrich the physical model by including also viscous terms. The resulting compressible Navier-Stokes equations serve as a model for laminar viscous compressible flows.

9.1 The compressible Navier-Stokes equations

In the following we give a detailed description of the two–dimensional steady state compressible Navier-Stokes equations. Like in Section 8.2, ρ, $\mathbf{v} = (v_1, v_2)^\top$, p and E denote the density, velocity vector, pressure and specific total energy, respectively. Furthermore, T denotes the temperature. The equations of motion are given by

$$\nabla \cdot (\mathcal{F}^c(\mathbf{u}) - \mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u})) \equiv \frac{\partial}{\partial x_k}\mathbf{f}_k^c(\mathbf{u}) - \frac{\partial}{\partial x_k}\mathbf{f}_k^v(\mathbf{u}, \nabla\mathbf{u}) = 0 \quad \text{in } \Omega. \tag{440}$$

The vector of conservative variables $\mathbf{u}$ and the convective fluxes $\mathbf{f}_k^c$, $k = 1, 2$, are given by (395). Furthermore, the viscous fluxes $\mathbf{f}_k^v$, $k = 1, 2$, are defined by

$$\mathbf{f}_1^v(\mathbf{u}, \nabla\mathbf{u}) = \begin{bmatrix} 0 \\ \tau_{11} \\ \tau_{21} \\ \tau_{1j}v_j + \mathcal{K}T_{x_1} \end{bmatrix} \quad \text{and} \quad \mathbf{f}_2^v(\mathbf{u}, \nabla\mathbf{u}) = \begin{bmatrix} 0 \\ \tau_{12} \\ \tau_{22} \\ \tau_{2j}v_j + \mathcal{K}T_{x_2} \end{bmatrix},$$

respectively, where $\mathcal{K}$ is the thermal conductivity coefficient. Finally, the viscous stress tensor is defined by

$$\tau = \mu\left(\nabla\mathbf{v} + (\nabla\mathbf{v})^\top - \tfrac{2}{3}(\nabla\cdot\mathbf{v})I\right),$$

where μ is the dynamic viscosity coefficient, and the temperature T is given by $e = c_v T$; thus

$$\mathcal{K}T = \tfrac{\mu\gamma}{Pr}\left(E - \tfrac{1}{2}\mathbf{v}^2\right),$$

where $Pr = 0.72$ is the Prandtl number.

For the purposes of discretization, we rewrite the compressible Navier–Stokes equations (440) in the following (equivalent) form:

$$\frac{\partial}{\partial x_k}\left(\mathbf{f}_k^c(\mathbf{u}) - G_{kl}(\mathbf{u})\frac{\partial\mathbf{u}}{\partial x_l}\right) = 0 \quad \text{in } \Omega.$$

Here, the matrices $G_{kl}(\mathbf{u}) = \partial\mathbf{f}_k^v(\mathbf{u}, \nabla\mathbf{u})/\partial u_{x_l}$, for $k, l = 1, 2$, are the homogeneity tensors defined by $\mathbf{f}_k^v(\mathbf{u}, \nabla\mathbf{u}) = G_{kl}(\mathbf{u})\partial\mathbf{u}/\partial x_l$, $k = 1, 2$, where

$$G_{11} = \frac{\mu}{\rho}\begin{pmatrix} 0 & 0 & 0 & 0 \\ -\frac{4}{3}v_1 & \frac{4}{3} & 0 & 0 \\ -v_2 & 0 & 1 & 0 \\ -\left(\frac{4}{3}v_1^2 + v_2^2 + \frac{\gamma}{Pr}\left(E - \mathbf{v}^2\right)\right) & \left(\frac{4}{3} - \frac{\gamma}{Pr}\right)v_1 & \left(1 - \frac{\gamma}{Pr}\right)v_2 & \frac{\gamma}{Pr} \end{pmatrix},$$

$$G_{12} = \frac{\mu}{\rho}\begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{2}{3}v_2 & 0 & -\frac{2}{3} & 0 \\ -v_1 & 1 & 0 & 0 \\ -\frac{1}{3}v_1v_2 & v_2 & -\frac{2}{3}v_1 & 0 \end{pmatrix}, G_{21} = \frac{\mu}{\rho}\begin{pmatrix} 0 & 0 & 0 & 0 \\ -v_2 & 0 & 1 & 0 \\ \frac{2}{3}v_1 & -\frac{2}{3} & 0 & 0 \\ -\frac{1}{3}v_1v_2 & -\frac{2}{3}v_2 & v_1 & 0 \end{pmatrix},$$

$$G_{22} = \frac{\mu}{\rho}\begin{pmatrix} 0 & 0 & 0 & 0 \\ -v_1 & 1 & 0 & 0 \\ -\frac{4}{3}v_2 & 0 & \frac{4}{3} & 0 \\ -\left(v_1^2 + \frac{4}{3}v_2^2 + \frac{\gamma}{Pr}\left(E - \mathbf{v}^2\right)\right) & \left(1 - \frac{\gamma}{Pr}\right)v_1 & \left(\frac{4}{3} - \frac{\gamma}{Pr}\right)v_2 & \frac{\gamma}{Pr} \end{pmatrix}.$$

Like for the compressible Euler equations we consider supersonic and subsonic inflow and outflow boundary conditions. Furthermore, we destinguish between *isothermal* and *adiabatic* wall boundary conditions. To this end, decomposing $\Gamma_W = \Gamma_{\text{iso}} \cup \Gamma_{\text{adia}}$, we set

$$\mathbf{v} = \mathbf{0} \;\text{ on }\; \Gamma_W, \qquad T = T_{\text{wall}} \;\text{ on }\; \Gamma_{\text{iso}}, \qquad \mathbf{n} \cdot \nabla T = 0 \;\text{ on }\; \Gamma_{\text{adia}}, \tag{441}$$

where T_{wall} is a given wall temperature.

9.2 DG discretizations of the compressible Navier-Stokes equations

The derivation of discontinuous Galerkin discretizations of the compressible Navier-Stokes equations is similar to the derivation for Poisson's equation. Starting point is the compressible Navier-Stokes equations written in terms of the homogeneity tensors $G(\mathbf{u})$. In the previous section we have concentrated on the discretization of the convective flux $\nabla \cdot \mathcal{F}^c(\mathbf{u}) = \frac{\partial}{\partial x_k}\mathbf{f}_k^c(\mathbf{u})$ representing the inviscid Euler part of the equations. Therefore, in the following, we can ignore the convective part and concentrate on the remaining viscous part, i.e. we consider the discretization of

$$-\nabla \cdot (G(\mathbf{u})\nabla \mathbf{u}) = -\frac{\partial}{\partial x_k}\left(G_{kl}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x_l}\right) = 0 \quad \text{in } \Omega, \tag{442}$$

subject to boundary conditions given above. In index notation this writes,

$$\partial_{x_k}\left((G(\mathbf{u})_{kl})_{ij}\; \partial_{x_l} u_j \right) = 0 \quad \text{in } \Omega.$$

Like for Poisson's equation we rewrite this problem as a first-order system:

$$\underline{\sigma} = G(\mathbf{u})\nabla \mathbf{u}, \qquad\qquad -\nabla \cdot \underline{\sigma} = 0 \quad \text{in } \Omega,$$

i.e. $\sigma_{ik} = (G(\mathbf{u})_{kl})_{ij}\; \partial_{x_l} u_j$. Multiplying the first and second equations by test functions $\underline{\tau}$ and v, respectively, integrating on an element $\kappa \in \mathcal{T}_h$, and integrating by parts, we obtain

$$\begin{aligned} \int_\kappa \underline{\sigma} : \underline{\tau}\, \mathrm{d}\mathbf{x} &= -\int_\kappa \mathbf{u}\nabla \cdot \left(G^\top(\mathbf{u})\underline{\tau}\right)\, \mathrm{d}\mathbf{x} + \int_{\partial\kappa} \mathbf{u}\left(G^\top(\mathbf{u})\underline{\tau}\right)\cdot \mathbf{n}\, \mathrm{d}s, \\ \int_\kappa \underline{\sigma} : \nabla \mathbf{v}\, \mathrm{d}\mathbf{x} &= \int_{\partial\kappa} \underline{\sigma}\cdot \mathbf{n}\mathbf{v}\, \mathrm{d}s, \end{aligned} \tag{443}$$

where $\mathbf{n}$ is the unit outward normal vector to $\partial\kappa$. Here, we used

$$\int_\kappa \sigma_{ik}\tau_{ik}\, \mathrm{d}\mathbf{x} = \int_\kappa (G(\mathbf{u})_{kl})_{ij}\; \partial_{x_l} u_j \tau_{ik}\, \mathrm{d}\mathbf{x} = \int_\kappa \partial_{x_l} u_j\, (G(\mathbf{u})_{kl})_{ij}\, \tau_{ik}\, \mathrm{d}\mathbf{x} = \int_\kappa \nabla \mathbf{u} : \left(G^\top(\mathbf{u})\underline{\tau}\right)\, \mathrm{d}\mathbf{x}.$$

In addition to the vector-valued discrete function space $\mathbf{V}_{h,p}^d$ defined in (399) we now introduce the tensor-valued discrete function space $\underline{\Sigma}_{h,p}^d$ consisting tensor-valued polynomial functions of degree $p \geq 0$, defined by

$$\begin{aligned} \underline{\Sigma}_{h,p}^d = \{ \underline{\tau} \in \left[L^2(\Omega)\right]^{4\times 2} :& \underline{\tau}|_\kappa \circ \sigma_\kappa \in [Q_p(\hat{\kappa})]^{4\times 2} \text{ if } \hat{\kappa} \text{ is the unit hypercube, and} \\ & \underline{\tau}|_\kappa \circ \sigma_\kappa \in [P_p(\hat{\kappa})]^{4\times 2} \text{ if } \hat{\kappa} \text{ is the unit simplex}, \kappa \in \mathcal{T}_h \}. \end{aligned}$$

If we sum (443) over all elements $\kappa \in \mathcal{T}_h$ and replace $\mathbf{u}$, $\mathbf{v}$, $\underline{\sigma}$ and $\underline{\tau}$ by discrete functions $\mathbf{u}_h, \mathbf{v}_h \in \mathbf{V}_{h,p}^d$ and $\underline{\sigma}_h, \underline{\tau}_h \in \underline{\Sigma}_{h,p}^d$ we obtain following discretization in the so-called *flux formulation*: find $\mathbf{u}_h \in \mathbf{V}_{h,p}^d$ and $\underline{\sigma}_h \in \underline{\Sigma}_{h,p}^d$ such that

$$\int_\Omega \underline{\sigma}_h : \underline{\tau}_h\, \mathrm{d}\mathbf{x} = -\int_\Omega \mathbf{u}_h\nabla_h \cdot \left(G^\top(\mathbf{u}_h)\underline{\tau}_h\right)\, \mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa} \hat{\mathbf{u}}_h\left(G^\top(\mathbf{u}_h)\underline{\tau}_h\right)\cdot \mathbf{n}\, \mathrm{d}s, \tag{444}$$

$$\int_\Omega \underline{\sigma}_h : \nabla_h \mathbf{v}_h\, \mathrm{d}\mathbf{x} = \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa} \hat{\underline{\sigma}}_h \cdot \mathbf{n}\mathbf{v}_h\, \mathrm{d}s, \tag{445}$$

for all $\underline{\tau}_h \in \underline{\Sigma}_{h,p}^d$ and $\mathbf{v}_h \in \mathbf{V}_{h,p}^d$. Here, the *numerical fluxes* $\hat{\mathbf{u}}_h$ and $\underline{\hat{\sigma}}_h$ are approximations to $\mathbf{u}$ and $\underline{\sigma} = \nabla \mathbf{u}$, respectively. Depending on the particular choice of $\hat{\mathbf{u}}_h$ and $\underline{\hat{\sigma}}_h$ several different DG methods can be derived, each with specific properties with respect to stability and accuracy.

The flux formulation (444) represents the discretization of a first order system with unknowns $\mathbf{u}_h \in \mathbf{V}_{h,p}^d$ and $\underline{\sigma}_h \in \underline{\Sigma}_{h,p}^d$. However, this is $(d+1)$ times the size of a problem involving $\mathbf{u}_h \in \mathbf{V}_{h,p}^d$, only. In order to reduce the problem size, the auxiliary variable $\underline{\sigma}_h$ in (444) and (445) is usually eliminated to gain a *primal formulation* involving only the primal variable $\mathbf{u}_h$. To this end, we perform a second integration by parts on each element κ in (444) and set $\underline{\tau}_h = \nabla_h \mathbf{v}_h$ which gives us

$$\int_\Omega \underline{\sigma}_h : \nabla_h \mathbf{v}_h \,\mathrm{d}\mathbf{x} = \int_\Omega G(\mathbf{u}_h) \nabla_h \mathbf{u}_h : \nabla_h \mathbf{v}_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} (\hat{\mathbf{u}}_h - \mathbf{u}_h) \left(G^\top(\mathbf{u}_h) \nabla \mathbf{v}_h \right) \cdot \mathbf{n} \,\mathrm{d}s. \quad (446)$$

Substituting (446) into (445) we obtain following problem: find $\mathbf{u}_h \in \mathbf{V}_{h,p}^d$ such that

$$\hat{N}_h^v(\mathbf{u}_h, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_{h,p}^d, \quad (447)$$

where the semilinear form $\hat{N}_h^v(\cdot,\cdot) : [H^1(\mathcal{T}_h)]^m \times [H^1(\mathcal{T}_h)]^m \to \mathbb{R}$ which is nonlinear in its first and linear in its second argument, is defined by

$$\hat{N}_h^v(\mathbf{u}_h, \mathbf{v}_h) = \int_\Omega G(\mathbf{u}_h) \nabla_h \mathbf{u}_h : \nabla_h \mathbf{v}_h \,\mathrm{d}\mathbf{x} - \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} \underline{\hat{\sigma}}_h \cdot \mathbf{n} \mathbf{v}_h \,\mathrm{d}s + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} (\hat{\mathbf{u}}_h - \mathbf{u}_h) \left(G^\top(\mathbf{u}_h) \nabla \mathbf{v}_h \right) \cdot \mathbf{n} \,\mathrm{d}s. \quad (448)$$

Here, we use the notation $\hat{N}_h^v$ instead of $\hat{N}_h$ to underline that $\hat{N}_h^v$ includes the discretization of the viscous part of the compressible Navier-Stokes equations only. Later in Section 9.2 we consider the discretization of the complete compressible Navier-Stokes equations, including convective and viscous parts, which will then be denoted by $\hat{N}_h$.

We call (447) the *primal formulation* of the method and call $\hat{N}_h(\cdot,\cdot)$ the *primal form*. This bilinear form is denoted by $\hat{N}_h$ (and not N_h) as it includes the (still unspecified) numerical fluxes $\hat{\mathbf{u}}_h$ and $\underline{\hat{\sigma}}_h$. Furthermore, $\hat{B}$ includes – through the specification of $\hat{\mathbf{u}}_h$ and $\underline{\hat{\sigma}}_h$ on the boundary — all boundary data terms. We note, that (448) is a generalization of (172) for Poisson's equation to the case of an elliptic system of equations with diffusion tensor $G(\mathbf{u})$.

Finally, we note that $\hat{N}_h$ in (448) is the *cell-based* primal form, i.e. it is given in terms of $\sum_\kappa \int_{\partial\kappa}$. This means that each interior face $e = \Gamma_\mathcal{I}$ occurs twice in the sum over all elements κ (once in $\int_{\partial\kappa}$ and once in $\int_{\partial\kappa'}$ for $e = \partial\kappa \cap \partial\kappa' \neq 0$).

In the following, we transfer the cell-based primal form into a *face-based* primal form, i.e. we rewrite $\hat{N}_h$ in terms of $\int_{\Gamma_\mathcal{I}}$ where each interior face is treated only once. However, before doing so, we introduce some more notation.

Mean value and jump operators First we recall the definition of mean values and jumps operating on scalar and vector-valued functions given in Section 5.1.

Definition 9.1 *Let $e \in \Gamma_\mathcal{I}$ be an interior edge between two adjacent elements κ^+ and κ^- with unit outward normal vectors, $\mathbf{n}^+, \mathbf{n}^- \in \mathbb{R}^d$, respectively. Let $q \in T(\mathcal{T}_h)$ and $\boldsymbol{\phi} \in [T(\mathcal{T}_h)]^d$ be the traces of a scalar and a vector valued function, respectively. Then, we define the mean value and jump operators, $\{\!\{\cdot\}\!\}$ and $[\![\cdot]\!]$, as follows*

$$\{\!\{q\}\!\} = \tfrac{1}{2}(q^+ + q^-), \qquad [\![q]\!] = q^+ \mathbf{n}^+ + q^- \mathbf{n}^-,$$
$$\{\!\{\boldsymbol{\phi}\}\!\} = \tfrac{1}{2}(\boldsymbol{\phi}^+ + \boldsymbol{\phi}^-), \qquad [\![\boldsymbol{\phi}]\!] = \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ + \boldsymbol{\phi}^- \cdot \mathbf{n}^-.$$

Definition 9.2 *On boundary edges $e \in \Gamma$ the mean value and jump operators are defined by*

$$\begin{aligned} \{\!\{q\}\!\} &= q^+, & [\![q]\!] &= q^+\mathbf{n}^+, \\ \{\!\{\boldsymbol{\phi}\}\!\} &= \boldsymbol{\phi}^+, & [\![\boldsymbol{\phi}]\!] &= \boldsymbol{\phi}^+ \cdot \mathbf{n}^+. \end{aligned}$$

Furthermore, we recall Lemma 5.4 which has been used to transfer between cell-based and face-based integrals:

Lemma 9.3 *Again, let $q \in T(\mathcal{T}_h)$ and $\boldsymbol{\phi} \in [T(\mathcal{T}_h)]^d$, then*

$$\sum_\kappa \int_{\partial\kappa\setminus\Gamma} \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+ \, ds = \int_{\Gamma_\mathcal{I}} \{\!\{\boldsymbol{\phi}\}\!\} \cdot [\![q]\!] \, ds + \int_{\Gamma_\mathcal{I}} [\![\boldsymbol{\phi}]\!] \, \{\!\{q\}\!\} \, ds, \tag{449}$$

$$\sum_\kappa \int_{\partial\kappa} \boldsymbol{\phi}^+ \cdot \mathbf{n}^+ \, q^+ \, ds = \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\boldsymbol{\phi}\}\!\} \cdot [\![q]\!] \, ds + \int_{\Gamma_\mathcal{I}} [\![\boldsymbol{\phi}]\!] \, \{\!\{q\}\!\} \, ds. \tag{450}$$

Proof: See Lemma 5.4. □

Next we give the definition of tensor mean value and jump operators.

Definition 9.4 *Let $e \in \Gamma_\mathcal{I}$ be an interior edge between two adjacent elements κ^+ and κ^- with unit outward normal vectors, $\mathbf{n}^+, \mathbf{n}^- \in \mathbb{R}^d$, respectively. Let $\mathbf{v} \in [T(\mathcal{T}_h)]^m$ and $\underline{\tau} \in [T(\mathcal{T}_h)]^{m\times d}$ be the traces of a vector-valued and tensor-valued function, respectively. Then, we define the mean value and jump operators, $\{\!\{\cdot\}\!\}$ and $[\![\cdot]\!]$, as follows*

$$\begin{aligned} \{\!\{\underline{\tau}\}\!\} &= \tfrac{1}{2}(\underline{\tau}^+ + \underline{\tau}^-) && \text{on } \Gamma_\mathcal{I}, & \{\!\{\underline{\tau}\}\!\} &= \underline{\tau}^+ && \text{on } \Gamma, \\ [\![\underline{\tau}]\!] &= \underline{\tau}^+\mathbf{n}^+ + \underline{\tau}^-\mathbf{n}^- && \text{on } \Gamma_\mathcal{I}, & [\![\underline{\tau}]\!] &= \underline{\tau}^+\mathbf{n}^+ && \text{on } \Gamma, \\ \underline{[\![\mathbf{v}]\!]} &= \mathbf{v}^+ \otimes \mathbf{n}^+ + \mathbf{v}^- \otimes \mathbf{n}^- && \text{on } \Gamma_\mathcal{I}, & \underline{[\![\mathbf{v}]\!]} &= \mathbf{v}^+ \otimes \mathbf{n}^+ && \text{on } \Gamma. \end{aligned}$$

Now we can generalize Lemma 9.3 to systems of equations:

Lemma 9.5 *Let $\mathbf{v} \in [T(\mathcal{T}_h)]^m$ and $\underline{\tau} \in [T(\mathcal{T}_h)]^{m\times d}$, then*

$$\sum_\kappa \int_{\partial\kappa\setminus\Gamma} \left(\underline{\tau}^+\mathbf{n}^+\right) \cdot \mathbf{v}^+ \, ds = \int_{\Gamma_\mathcal{I}} \{\!\{\underline{\tau}\}\!\} : \underline{[\![\mathbf{v}]\!]} \, ds + \int_{\Gamma_\mathcal{I}} [\![\underline{\tau}]\!] \cdot \{\!\{\mathbf{v}\}\!\} \, ds, \tag{451}$$

$$\sum_\kappa \int_{\partial\kappa} \left(\underline{\tau}^+\mathbf{n}^+\right) \cdot \mathbf{v}^+ \, ds = \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\underline{\tau}\}\!\} : \underline{[\![\mathbf{v}]\!]} \, ds + \int_{\Gamma_\mathcal{I}} [\![\underline{\tau}]\!] \cdot \{\!\{\mathbf{v}\}\!\} \, ds. \tag{452}$$

Proof: Employing Lemma 9.3 for any $i = 1, \ldots, m$ we obtain

$$\begin{aligned} \sum_\kappa \int_{\partial\kappa\setminus\Gamma} \tau_{ik}^+ n_k^+ \, v_i^+ \, \mathrm{d}s &= \int_{\Gamma_\mathcal{I}} \{\!\{\tau_{ik}\}\!\} \left(v_i^+ n_k^+ + v_i^- n_k^-\right) \mathrm{d}s + \int_{\Gamma_\mathcal{I}} \left(\tau_{ik}^+ n_k^+ + \tau_{ik}^- n_k^-\right) \{\!\{v_i\}\!\} \, \mathrm{d}s \\ &= \int_{\Gamma_\mathcal{I}} \{\!\{\underline{\tau}\}\!\} : \underline{[\![\mathbf{v}]\!]} \, \mathrm{d}s + \int_{\Gamma_\mathcal{I}} [\![\underline{\tau}]\!] \cdot \{\!\{\mathbf{v}\}\!\} \, \mathrm{d}s, \end{aligned}$$

thus (451). Use Definition 9.4 of mean value and jump operators on Γ for (452). □

Face-based form of DG discretizations We now proceed in transferring the cell-based form (448) into a face-based form. To this end, we use Equation (452) twice (once for $\underline{\tau} = \underline{\hat{\sigma}}_h$ and $\mathbf{v} = \mathbf{v}_h$, and once for $\underline{\tau} = G^\top(\mathbf{u}_h)\nabla\mathbf{v}_h$, i.e. $\tau_{jl} = (G(\mathbf{u})_{kl})_{ij}\,\partial_{x_k} v_i$, and $\mathbf{v} = \hat{\mathbf{u}}_h - \mathbf{u}_h$) and rewrite (448) as follows

$$\begin{aligned} \hat{N}_h^v(\mathbf{u}_h, \mathbf{v}_h) = &\int_\Omega G(\mathbf{u}_h)\nabla_h\mathbf{u}_h : \nabla_h\mathbf{v}_h \, \mathrm{d}\mathbf{x} - \int_{\Gamma_\mathcal{I}\cup\Gamma} \{\!\{\underline{\hat{\sigma}}_h\}\!\} : \underline{[\![\mathbf{v}_h]\!]} \, \mathrm{d}s - \int_{\Gamma_\mathcal{I}} [\![\underline{\hat{\sigma}}_h]\!] \cdot \{\!\{\mathbf{v}_h\}\!\} \, \mathrm{d}s \\ &+ \int_{\Gamma_\mathcal{I}\cup\Gamma} \underline{[\![\hat{\mathbf{u}}_h - \mathbf{u}_h]\!]} : \{\!\{G^\top(\mathbf{u}_h)\nabla\mathbf{v}_h\}\!\} \, \mathrm{d}s + \int_{\Gamma_\mathcal{I}} \{\!\{\hat{\mathbf{u}}_h - \mathbf{u}_h\}\!\} \cdot [\![G^\top(\mathbf{u}_h)\nabla\mathbf{v}_h]\!] \, \mathrm{d}s, \end{aligned}$$

which results in following *face-based* primal form:

$$\hat{N}_h^v(\mathbf{u}_h,\mathbf{v}_h) = \int_\Omega G(\mathbf{u}_h)\nabla_h\mathbf{u}_h : \nabla_h\mathbf{v}_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_\mathcal{I}\cup\Gamma} \underline{[\![\hat{\mathbf{u}}_h - \mathbf{u}_h]\!]} : \{\!\{G^\top(\mathbf{u}_h)\nabla\mathbf{v}_h\}\!\} - \{\!\{\underline{\hat{\sigma}}_h\}\!\} : \underline{[\![\mathbf{v}_h]\!]}\,\mathrm{d}s + \int_{\Gamma_\mathcal{I}} \{\!\{\hat{\mathbf{u}}_h - \mathbf{u}_h\}\!\}\cdot[\![G^\top(\mathbf{u}_h)\nabla\mathbf{v}_h]\!] - [\![\underline{\hat{\sigma}}_h]\!]\cdot\{\!\{\mathbf{v}_h\}\!\}\,\mathrm{d}s. \quad (453)$$

Derivation of various DG discretization methods Depending on the specific choice of numerical fluxes $\hat{\mathbf{u}}_h$ and $\underline{\hat{\sigma}}_h$ several different discontinuous Galerkin discretizations of the compressible Navier-Stokes equations can be derived. First we note, that for the discretizations to be consistent and adjoint consistent there are requirements on the numerical fluxes $\hat{\mathbf{u}}_h$ and $\underline{\hat{\sigma}}_h$ analog to Theorems 5.7 and 5.12 for Poisson's equations. In fact, we have

Theorem 9.6 *Let $\hat{N}_h^v(\cdot,\cdot)$ be given by* (453)*. Then the discretization: find $\mathbf{u}_h \in \mathbf{V}_{h,p}^d$ such that*

$$\hat{N}_h^v(\mathbf{u}_h,\mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_{h,p}^d, \quad (454)$$

of a homogeneous Dirichlet problem is consistent *if and only if the numerical fluxes $\hat{\mathbf{u}}$ and $\underline{\hat{\sigma}}$ are consistent, i.e.*

$$\hat{\mathbf{u}}(\mathbf{v}) = \mathbf{v}, \qquad \underline{\hat{\sigma}}(\mathbf{v},\nabla\mathbf{v}) = G(\mathbf{v})\nabla\mathbf{v} \quad \text{on } \Gamma_\mathcal{I}\cup\Gamma, \quad (455)$$

holds for all functions $\mathbf{v} \in [H^2(\Omega)]^m$. Furthermore, the discretization (454) *is* adjoint consistent *if and only if the numerical fluxes $\hat{\mathbf{u}}$ and $\underline{\hat{\sigma}}$ are conservative, i.e.*

$$\underline{[\![\hat{\mathbf{u}}(\mathbf{v})]\!]} = 0, \qquad [\![\underline{\hat{\sigma}}(\mathbf{v},\nabla\mathbf{v})]\!] = 0 \quad \text{on } \Gamma_\mathcal{I}\cup\Gamma, \quad (456)$$

holds for all functions $\mathbf{v} \in [H^2(\Omega)]^m$.

Proof: Analog to the proofs of Theorems 5.7 and 5.12. □

We recall from the discretization of Poisson's equation that the interior face terms of the symmetric interior penalty method, SIPG, and of the modified DG discretization of Bassi and Rebay, BR2, are adjoint consistent. In contrast to that the method of Baumann-Oden, BO, and the nonsymmetric interior penalty method, NIPG, are adjoint inconsistent and will thus not be considered in the following.

For SIPG and BR2 let us choose the fluxes $\hat{\mathbf{u}}_h$ and $\underline{\hat{\sigma}}_h$ to be given by

$$\hat{\mathbf{u}}_h = \{\!\{\mathbf{u}_h\}\!\}, \qquad \underline{\hat{\sigma}}_h = \{\!\{G(\mathbf{u}_h)\nabla_h\mathbf{u}_h\}\!\} - \boldsymbol{\delta}(\mathbf{u}_h) \qquad \text{on } \Gamma_\mathcal{I},$$

where the penalization term $\boldsymbol{\delta}(\mathbf{u}_h)$ is given by

$$\begin{aligned}
\boldsymbol{\delta}(\mathbf{u}_h) &= \boldsymbol{\delta}^{\mathrm{ips}}(\mathbf{u}_h) = C_{\mathrm{IP}}\tfrac{p^2}{h_e}\mu\underline{[\![\mathbf{u}_h]\!]} && \text{for IP [31]},\\
\boldsymbol{\delta}(\mathbf{u}_h) &= \boldsymbol{\delta}^{\mathrm{ip}}(\mathbf{u}_h) = C_{\mathrm{IP}}\tfrac{p^2}{h_e}\{\!\{G(\mathbf{u}_h)\}\!\}\underline{[\![\mathbf{u}_h]\!]} && \text{for IP [33]},\\
\boldsymbol{\delta}(\mathbf{u}_h) &= \boldsymbol{\delta}^{\mathrm{br2}}(\mathbf{u}_h) = C_{\mathrm{BR2}}\{\!\{\mathbf{L}_0^e(\mathbf{u}_h)\}\!\} && \text{for BR2 [7, 8]},
\end{aligned} \quad (457)$$

where the local lifting operator $\mathbf{L}_0^e(\mathbf{u}_h) \in \underline{\Sigma}_{h,p}^d$ is defined by:

$$\int_{\Omega_e} \mathbf{L}_0^e(\mathbf{u}_h) : \underline{\tau}\,\mathrm{d}\mathbf{x} = \int_e \underline{[\![\mathbf{u}_h]\!]} : \{\!\{G^\top(\mathbf{u}_h)\underline{\tau}\}\!\}\,\mathrm{d}s \qquad \forall \underline{\tau} \in \underline{\Sigma}_{h,p}^d,$$

where $\Omega_e = \kappa_e^+ \cup \kappa_e^-$ with $e = \partial\kappa_e^+ \cap \partial\kappa_e^-$. Then,

$$\begin{aligned}
\underline{[\![\hat{\mathbf{u}}_h]\!]} &= \underline{[\![\{\!\{\mathbf{u}_h\}\!\}]\!]} = 0,\\
\{\!\{\hat{\mathbf{u}}_h\}\!\} &= \{\!\{\{\!\{\mathbf{u}_h\}\!\}\}\!\} = \{\!\{\mathbf{u}_h\}\!\},\\
\{\!\{\underline{\hat{\sigma}}_h\}\!\} &= \{\!\{\{\!\{G(\mathbf{u}_h)\nabla_h\mathbf{u}_h\}\!\}\}\!\} - \{\!\{\boldsymbol{\delta}(\mathbf{u}_h)\}\!\} = \{\!\{G(\mathbf{u}_h)\nabla_h\mathbf{u}_h\}\!\} - \boldsymbol{\delta}(\mathbf{u}_h),\\
[\![\underline{\hat{\sigma}}_h]\!] &= [\![\{\!\{G(\mathbf{u}_h)\nabla_h\mathbf{u}_h\}\!\}]\!] - [\![\boldsymbol{\delta}(\mathbf{u}_h)]\!] = 0,
\end{aligned}$$

the last term in (453) vanishes and thus (453) reduces to

$$N_h^v(\mathbf{u}_h,\mathbf{v}_h) = \int_\Omega G(\mathbf{u}_h)\nabla_h\mathbf{u}_h : \nabla_h\mathbf{v}_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_\mathcal{I}} \underline{[\![\mathbf{u}_h]\!]} : \{\!\{G^\top(\mathbf{u}_h)\nabla\mathbf{v}_h\}\!\} \,\mathrm{d}s \\ - \int_{\Gamma_\mathcal{I}} \{\!\{G(\mathbf{u}_h)\nabla_h\mathbf{u}_h\}\!\} : \underline{[\![\mathbf{v}_h]\!]} + \int_{\Gamma_\mathcal{I}} \boldsymbol{\delta}(\mathbf{u}_h) : \underline{[\![\mathbf{v}_h]\!]} \,\mathrm{d}s + \hat{N}_{\Gamma,h}(\mathbf{u}_h,\mathbf{v}_h),$$

where the boundary term $N_{\Gamma,h}^v(\mathbf{u}_h,\mathbf{v}_h)$ will be specified in the following section.

Discretization of viscous boundary terms On boundary edges we choose

$$\hat{\mathbf{u}}_h = \mathbf{u}_\Gamma(\mathbf{u}_h^+), \qquad \underline{\hat{\sigma}}_h = \mathcal{F}_\Gamma^v(\mathbf{u}_h^+,\nabla\mathbf{u}_h^+) - \boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+),$$

where

$$\mathcal{F}_\Gamma^v(\mathbf{u}_h,\nabla\mathbf{u}_h) = \mathcal{F}^v(\mathbf{u}_\Gamma(\mathbf{u}_h),\nabla\mathbf{u}_h) = G_\Gamma(\mathbf{u}_h)\nabla\mathbf{u}_h = G(\mathbf{u}_\Gamma(\mathbf{u}_h))\nabla\mathbf{u}_h \quad \text{on } \Gamma,$$

and on $\Gamma_{\mathrm{W,adia}}$, the viscous flux $\mathcal{F}_\Gamma^v$ and the corresponding homogeneity tensor G_Γ are modified such that $\mathbf{n}\cdot\nabla T = 0$, i.e.

$$\mathbf{n}\cdot\mathcal{F}_\Gamma^v(\mathbf{u}_h,\nabla\mathbf{u}_h) = (0,\tau_{1j}n_{x_j},\tau_{2j}n_{x_j},\tau_{ij}v_jn_{x_i})^\top.$$

The penalization term $\boldsymbol{\delta}_\Gamma(\mathbf{u}_h)$ on Γ is given by

$$\begin{aligned}
&\boldsymbol{\delta}_\Gamma(\mathbf{u}_h) = \boldsymbol{\delta}_\Gamma^{\mathrm{ips}}(\mathbf{u}_h) = C_{\mathrm{IP}}\tfrac{p^2}{h_e}\mu\,(\mathbf{u}_h - \mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n} && \text{for IP [31]},\\
&\boldsymbol{\delta}_\Gamma(\mathbf{u}_h) = \boldsymbol{\delta}_\Gamma^{\mathrm{ip}}(\mathbf{u}_h) = C_{\mathrm{IP}}\tfrac{p^2}{h_e}G_\Gamma(\mathbf{u}_h)\,(\mathbf{u}_h - \mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n} && \text{for IP [33]},\\
&\boldsymbol{\delta}_\Gamma(\mathbf{u}_h) = \boldsymbol{\delta}_\Gamma^{\mathrm{br2}}(\mathbf{u}_h) = C_{\mathrm{BR2}}\mathbf{L}_\Gamma^e(\mathbf{u}_h) && \text{for BR2 [7, 8]},
\end{aligned} \tag{458}$$

where the local lifting operator $\mathbf{L}_\Gamma^e(\mathbf{u}_h) \in \underline{\Sigma}_{h,p}^d$ on Γ is defined by:

$$\int_\kappa \mathbf{L}_\Gamma^e(\mathbf{u}_h) : \underline{\tau}\,\mathrm{d}\mathbf{x} = \int_e (\mathbf{u}_h - \mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n} : \left(G_\Gamma^\top(\mathbf{u}_h)\underline{\tau}\right) \mathrm{d}s \qquad \forall\underline{\tau}\in\underline{\Sigma}_{h,p}^d$$

for κ such that $\partial\kappa\cap\Gamma = e$. Thereby the boundary term $N_{\Gamma,h}^v(\mathbf{u}_h,\mathbf{v}_h)$ is given by

$$\begin{aligned}
N_{\Gamma,h}^v(\mathbf{u}_h,\mathbf{v}_h) = &-\int_\Gamma \mathcal{F}_\Gamma^v(\mathbf{u}_h^+,\nabla\mathbf{u}_h^+) : \mathbf{v}_h^+\otimes\mathbf{n}\,\mathrm{d}s\\
&-\int_\Gamma \left(G_\Gamma^\top(\mathbf{u}_h^+)\nabla\mathbf{v}_h^+\right) : (\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+))\otimes\mathbf{n}\,\mathrm{d}s + \int_\Gamma \boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+) : \mathbf{v}_h^+\otimes\mathbf{n}\,\mathrm{d}s.
\end{aligned} \tag{459}$$

Here, the boundary value function $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$ on supersonic and subsonic inflow and outflow boundary conditions is given like for inviscid flows in Section 8.4. Additionally, for no-slip wall boundaries with $\mathbf{v} = 0$ we distinguish *adiabatic* boundary conditions, $\mathbf{n}\cdot\nabla T = 0$, where

$$\mathbf{u}_\Gamma(\mathbf{u}) = (u_1,0,0,u_4)^\top \quad \text{on} \quad \Gamma_{\mathrm{W,adia}}, \tag{460}$$

and *isothermal* boundary conditions, $T = T_{\mathrm{wall}}$, where

$$\mathbf{u}_\Gamma(\mathbf{u}) = (u_1,0,0,u_1c_vT_{\mathrm{wall}})^\top \quad \text{on} \quad \Gamma_{\mathrm{W,iso}}. \tag{461}$$

Having derived DG discretizations of the compressible Euler equations and various DG discretizations of the viscous part of the compressible Navier-Stokes equations we now can combine the discretizations of convective and viscous parts to form the DG discretization of the compressible Navier-Stokes equations: find $\mathbf{u}_h\in\mathbf{V}_{h,p}^d$ such that

$$N_h(\mathbf{u}_h,\mathbf{v}_h) = 0 \quad \forall\mathbf{v}_h\in\mathbf{V}_{h,p}^d, \tag{462}$$

where the semilinear form $N_h(\cdot,\cdot) : [H^1(\mathcal{T}_h)]^m \times [H^1(\mathcal{T}_h)]^m \to \mathbb{R}$ is given by

$$\begin{aligned} N_h(\mathbf{u}_h, \mathbf{v}_h) = &- \int_\Omega \mathcal{F}^c(\mathbf{u}_h) : \nabla_h \mathbf{v}_h \,\mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} \mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}) \cdot \mathbf{v}_h^+ \,\mathrm{d}s \\ &\int_\Omega G(\mathbf{u}_h) \nabla_h \mathbf{u}_h : \nabla_h \mathbf{v}_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_\mathcal{I}} \underline{[\![\mathbf{u}_h]\!]} : \{\!\{ G^\top(\mathbf{u}_h) \nabla \mathbf{v}_h \}\!\} \,\mathrm{d}s \\ &- \int_{\Gamma_\mathcal{I}} \{\!\{ G(\mathbf{u}_h) \nabla_h \mathbf{u}_h \}\!\} : \underline{[\![\mathbf{v}_h]\!]} \,\mathrm{d}s + \int_{\Gamma_\mathcal{I}} \boldsymbol{\delta}(\mathbf{u}_h) : \underline{[\![\mathbf{v}_h]\!]} \,\mathrm{d}s + N_{\Gamma,h}(\mathbf{u}_h, \mathbf{v}_h), \end{aligned}$$

and the boundary term $N_{\Gamma,h}(\mathbf{u}_h, \mathbf{v}_h)$ is given by

$$\begin{aligned} N_{\Gamma,h}(\mathbf{u}_h, \mathbf{v}_h) = &\int_\Gamma \mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}) \cdot \mathbf{v}_h^+ \,\mathrm{d}s - \int_\Gamma \mathcal{F}_\Gamma^v(\mathbf{u}_h^+, \nabla \mathbf{u}_h^+) \,:\, \mathbf{v}_h^+ \otimes \mathbf{n} \,\mathrm{d}s \\ &- \int_\Gamma \left(G_\Gamma^\top(\mathbf{u}_h^+) \nabla \mathbf{v}_h^+ \right) : \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+) \right) \otimes \mathbf{n} \,\mathrm{d}s + \int_\Gamma \boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+) : \mathbf{v}_h^+ \otimes \mathbf{n} \,\mathrm{d}s. \end{aligned} \tag{463}$$

Furthermore, on Γ we have

$$\begin{aligned} \mathcal{H}_\Gamma(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}) &= \mathbf{n} \cdot \mathcal{F}_\Gamma^c(\mathbf{u}_h^+) = \mathbf{n} \cdot \mathcal{F}^c(\mathbf{u}_\Gamma(\mathbf{u}_h^+)), \\ \mathcal{F}_\Gamma^v(\mathbf{u}_h, \nabla \mathbf{u}_h) &= \mathcal{F}^v(\mathbf{u}_\Gamma(\mathbf{u}_h), \nabla \mathbf{u}_h) = G_\Gamma(\mathbf{u}_h) \nabla \mathbf{u}_h, \\ G_\Gamma(\mathbf{u}_h) \nabla \mathbf{u}_h &= G(\mathbf{u}_\Gamma(\mathbf{u}_h)) \nabla \mathbf{u}_h, \end{aligned}$$

and $\boldsymbol{\delta}(\mathbf{u}_h)$ and $\boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+)$ are as given in (457) and (458), respectively. Finally, $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$ is given like in Section 8.4 and in Equations (460) and (461).

9.3 Adjoint consistency analysis of DG for the compressible Navier-Stokes equations

In this section we analyze the consistency and adjoint consistency property of the compressible Navier-Stokes equations. In particular, we derive target functional modifications which result in an adjoint consistent discretization.

9.3.1 The continuous adjoint problem to the compressible Navier-Stokes equations

The most important target quantities in viscous compressible flows are the total (i.e. the pressure induced plus viscous) drag and lift coefficients, c_d and c_l, defined by

$$J(\mathbf{u}) = \int_\Gamma j(\mathbf{u}) \,\mathrm{d}s = \frac{1}{C_\infty} \int_{\Gamma_W} (p\,\mathbf{n} - \underline{\tau}\,\mathbf{n}) \cdot \boldsymbol{\psi} \,\mathrm{d}s = \frac{1}{C_\infty} \int_{\Gamma_W} (p\,n_i - \tau_{ij} n_j)\, \psi_i \,\mathrm{d}s, \tag{464}$$

where C_∞ and $\boldsymbol{\psi}$ are as in (421). In order to derive the adjoint problem, we multiply the left hand side of (440) by $\mathbf{z}$, integrate by parts and linearize about $\mathbf{u}$ to obtain

$$\begin{aligned} &\left(\nabla \cdot \left(\mathcal{F}_\mathbf{u}^c \mathbf{w} - \mathcal{F}_\mathbf{u}^v \mathbf{w} - \mathcal{F}_{\nabla\mathbf{u}}^v \nabla \mathbf{w}\right), \mathbf{z}\right)_\Omega \\ &\qquad = -\left(\left(\mathcal{F}_\mathbf{u}^c - \mathcal{F}_\mathbf{u}^v\right) \mathbf{w} - \mathcal{F}_{\nabla\mathbf{u}}^v \nabla \mathbf{w}, \nabla \mathbf{z}\right)_\Omega + \left(\mathbf{n} \cdot \left(\mathcal{F}_\mathbf{u}^c \mathbf{w} - \mathcal{F}_\mathbf{u}^v \mathbf{w} - \mathcal{F}_{\nabla\mathbf{u}}^v \nabla \mathbf{w}\right), \mathbf{z}\right)_\Gamma, \end{aligned}$$

where $\mathcal{F}_\mathbf{u}^v := \partial_\mathbf{u} \mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u}) = G'[\mathbf{u}]\nabla\mathbf{u}$ and $\mathcal{F}_{\nabla\mathbf{u}}^v := \partial_{\nabla\mathbf{u}} \mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u}) = G(\mathbf{u})$ denote the derivatives of $\mathcal{F}^v$ with respect to $\mathbf{u}$ and $\nabla\mathbf{u}$, respectively. Using integration by parts once more, we obtain the following variational formulation of the continuous adjoint problem: find $\mathbf{z}$ such that

$$\begin{aligned} &- \left(\mathbf{w}, \left(\mathcal{F}_\mathbf{u}^c - \mathcal{F}_\mathbf{u}^v\right)^\top \nabla \mathbf{z}\right)_\Omega - \left(\mathbf{w}, \nabla \cdot \left(\left(\mathcal{F}_{\nabla\mathbf{u}}^v\right)^\top \nabla \mathbf{z}\right)\right)_\Omega + \left(\mathbf{w}, \mathbf{n} \cdot \left(\left(\mathcal{F}_{\nabla\mathbf{u}}^v\right)^\top \nabla \mathbf{z}\right)\right)_\Gamma \\ &\qquad + \left(\mathbf{w}, \left(\mathbf{n} \cdot \left(\mathcal{F}_\mathbf{u}^c - \mathcal{F}_\mathbf{u}^v\right)\right)^\top \mathbf{z}\right)_\Gamma - \left(\nabla \mathbf{w}, \left(\mathbf{n} \cdot \mathcal{F}_{\nabla\mathbf{u}}^v\right)^\top \mathbf{z}\right)_\Gamma = J'[\mathbf{u}](\mathbf{w}) \quad \forall \mathbf{w} \in \mathbf{V}. \end{aligned}$$

Given that

$$\begin{aligned} J'[\mathbf{u}](\mathbf{w}) =& \frac{1}{C_\infty}\int_{\Gamma_W} (p_{\mathbf{u}}[\mathbf{u}]\,\mathbf{n} - \underline{\tau}_{\mathbf{u}}[\mathbf{u}]\,\mathbf{n})\cdot\boldsymbol{\psi}\,\mathbf{w} - (\underline{\tau}_{\nabla\mathbf{u}}[\mathbf{u}]\,\mathbf{n})\cdot\boldsymbol{\psi}\,\nabla\mathbf{w}\,\mathrm{d}s \\ =& \left(\mathbf{w}, \tfrac{1}{C_\infty}\,(p_{\mathbf{u}}\,\mathbf{n} - \underline{\tau}_{\mathbf{u}}\,\mathbf{n})\cdot\boldsymbol{\psi}\right)_{\Gamma_W} - \left(\nabla\mathbf{w}, \tfrac{1}{C_\infty}\,(\underline{\tau}_{\nabla\mathbf{u}}\,\mathbf{n})\cdot\boldsymbol{\psi}\right)_{\Gamma_W}, \end{aligned} \tag{465}$$

we see, that the adjoint solution $\mathbf{z}$ satisfies following equation

$$-\left(\mathcal{F}^c_{\mathbf{u}} - \mathcal{F}^v_{\mathbf{u}}\right)^\top \nabla\mathbf{z} - \nabla\cdot\left(\left(\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top\nabla\mathbf{z}\right) = 0, \tag{466}$$

subject to the boundary conditions on $\Gamma_W = \Gamma_{\text{iso}}\cup\Gamma_{\text{adia}}$,

$$\left(\mathbf{n}\cdot\left(\mathcal{F}^c_{\mathbf{u}} - \mathcal{F}^v_{\mathbf{u}}\right)\right)^\top\mathbf{z} + \mathbf{n}\cdot\left(\left(\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top\nabla\mathbf{z}\right) = \tfrac{1}{C_\infty}\,(p_{\mathbf{u}}\,\mathbf{n} - \underline{\tau}_{\mathbf{u}}\,\mathbf{n})\cdot\boldsymbol{\psi}, \tag{467}$$

$$\left(\mathbf{n}\cdot\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top\mathbf{z} = \tfrac{1}{C_\infty}\,(\underline{\tau}_{\nabla\mathbf{u}}\,\mathbf{n})\cdot\boldsymbol{\psi}. \tag{468}$$

At wall boundaries Γ_W, where $\mathbf{v} = (v_1, v_2)^\top = 0$, the normal viscous flux reduces to $\mathbf{n}\cdot\mathcal{F}^v(\mathbf{u},\nabla\mathbf{u}) = (0, (\tau\mathbf{n})_1, (\tau\mathbf{n})_2, \mathcal{K}\mathbf{n}\cdot\nabla T)^\top$. Hence, (468) is fulfilled provided $\mathbf{z}$ satisfies

$$\begin{pmatrix} 0 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_1\, z_2 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_2\, z_3 \\ \mathcal{K}\mathbf{n}\cdot\nabla T_{\nabla\mathbf{u}}\, z_4 \end{pmatrix} = \frac{1}{C_\infty}\begin{pmatrix} 0 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_1\,\psi_1 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_2\,\psi_2 \\ 0 \end{pmatrix}, \tag{469}$$

which reduces to the conditions $z_2 = \frac{1}{C_\infty}\psi_1$ on Γ_W, $z_3 = \frac{1}{C_\infty}\psi_2$ on Γ_W, and $z_4 = 0$ on Γ_{iso}. At adiabatic boundaries we have $\mathbf{n}\cdot\nabla T = 0$ and the last condition in (469) vanishes. Substituted into (467) we obtain $\mathbf{n}\cdot((\mathcal{F}^v_{\nabla\mathbf{u}})^\top\nabla\mathbf{z}) = 0$ on Γ_W which at adiabatic boundaries reduces to $\mathbf{n}\cdot\nabla z_4 = 0$. On isothermal boundaries no additional boundary condition is obtained. In summary, the boundary conditions of the adjoint problem (466) to the compressible Navier-Stokes equations are given by

$$z_2 = \tfrac{1}{C_\infty}\psi_1,\ z_3 = \tfrac{1}{C_\infty}\psi_2 \ \text{ on } \Gamma_W, \qquad z_4 = 0 \ \text{ on } \Gamma_{\text{iso}}, \qquad \mathbf{n}\cdot\nabla z_4 = 0 \ \text{ on } \Gamma_{\text{adia}}. \tag{470}$$

9.3.2 Primal residual form of DG for the compressible Navier-Stokes equations

Using integration by parts in (462) we obtain

$$\begin{aligned} N_h(\mathbf{u}_h,\mathbf{v}_h) \;\equiv\; & \int_\Omega (\nabla_h\cdot\mathcal{F}^c(\mathbf{u}_h))\cdot\mathbf{v}_h\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\left(\mathcal{H}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+) - \mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h^+)\right)\cdot\mathbf{v}_h^+\,\mathrm{d}s \\ & - \int_\Omega (\nabla_h\cdot\mathcal{F}^v(\mathbf{u}_h,\nabla_h\mathbf{u}_h))\cdot\mathbf{v}_h\,\mathrm{d}\mathbf{x} + \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}[\![\mathcal{F}^v(\mathbf{u}_h,\nabla_h\mathbf{u}_h)]\!]\cdot\mathbf{v}_h^+\,\mathrm{d}s \\ & -\frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} G^\top(\mathbf{u}_h)\nabla_h\mathbf{v}_h : \underline{[\![\mathbf{u}_h]\!]}\,\mathrm{d}s + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\boldsymbol{\delta}(\mathbf{u}_h) : \mathbf{v}_h^+\otimes\mathbf{n}^+\,\mathrm{d}s \\ & - \int_\Gamma\left(\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{F}^v(\mathbf{u}_h^+,\nabla\mathbf{u}_h^+)\right) : \mathbf{v}_h^+\otimes\mathbf{n}\,\mathrm{d}s + N_{\Gamma,h}(\mathbf{u}_h,\mathbf{v}_h) = 0, \end{aligned}$$

which can be expressed in the primal residual form as follows: find $\mathbf{u}_h\in\mathbf{V}^d_{h,p}$ such that

$$\begin{aligned} \int_\Omega \mathbf{R}(\mathbf{u}_h)\cdot\mathbf{v}_h\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathbf{r}(\mathbf{u}_h)\cdot\mathbf{v}_h^+ + \boldsymbol{\rho}(\mathbf{u}_h) : \nabla\mathbf{v}_h^+\,\mathrm{d}s & \\ + \int_\Gamma \mathbf{r}_\Gamma(\mathbf{u}_h)\cdot\mathbf{v}_h^+ + \boldsymbol{\rho}_\Gamma(\mathbf{u}_h) : \nabla\mathbf{v}_h^+\,\mathrm{d}s = 0 \quad \forall\mathbf{v}_h\in\mathbf{V}^d_{h,p}, & \end{aligned}$$

where the primal residuals are given by

$$
\begin{aligned}
\mathbf{R}(\mathbf{u}_h) =& -\nabla\cdot\mathcal{F}^c(\mathbf{u}_h) + \nabla\cdot\mathcal{F}^v(\mathbf{u}_h,\nabla_h\mathbf{u}_h) && \text{in } \kappa,\kappa\in\mathcal{T}_h,\\
\mathbf{r}(\mathbf{u}_h) =& \mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{H}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+) - \frac{1}{2}[\![\mathcal{F}^v(\mathbf{u}_h,\nabla_h\mathbf{u}_h)]\!] - \mathbf{n}\cdot\boldsymbol{\delta}(\mathbf{u}_h),\\
\boldsymbol{\rho}(\mathbf{u}_h) =& \frac{1}{2}\left(G(\mathbf{u}_h)\underline{[\![\mathbf{u}_h]\!]}\right)^\top && \text{on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h,\\
\mathbf{r}_\Gamma(\mathbf{u}_h) =& \mathbf{n}\cdot\left(\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{F}_\Gamma^c(\mathbf{u}_h^+) - \mathcal{F}^v(\mathbf{u}_h^+,\nabla\mathbf{u}_h^+) + \mathcal{F}_\Gamma^v(\mathbf{u}_h^+,\nabla\mathbf{u}_h^+) - \boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+)\right),\\
\boldsymbol{\rho}_\Gamma(\mathbf{u}_h) =& \left(G_\Gamma^\top(\mathbf{u}_h^+) : \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)\otimes\mathbf{n}\right)^\top && \text{on } \Gamma.
\end{aligned}
$$

We see that the exact solution $\mathbf{u}$ to (440) satisfies

$$
\mathbf{R}(\mathbf{u}) = 0, \qquad \mathbf{r}(\mathbf{u}) = 0, \qquad \boldsymbol{\rho}(\mathbf{u}) = 0, \qquad \mathbf{r}_\Gamma(\mathbf{u}) = 0, \qquad \boldsymbol{\rho}_\Gamma(\mathbf{u}) = 0,
$$

where we used consistency of the numerical flux, $\mathcal{H}(\mathbf{w},\mathbf{w},\mathbf{n}) = \mathbf{n}\cdot\mathcal{F}^c(\mathbf{w})$, continuity of $\mathbf{u}$, and the consistency of the boundary function, i.e. $\mathbf{u}$ satisfies $\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ on Γ. We conclude that the discretization given in Section 9.2 is consistent.

9.3.3 Adjoint residual form of DG for the compressible Navier-Stokes equations

Given the target quantity $J(\cdot)$ defined in (464) with Fréchet derivative (465), we consider following modification of $J(\cdot)$

$$
\tilde{J}(\mathbf{u}_h) = J(\mathbf{i}(\mathbf{u}_h)) + \int_\Gamma r_J(\mathbf{u}_h)\,\mathrm{d}s = J_\Gamma(\mathbf{u}_h) + \int_\Gamma r_J(\mathbf{u}_h)\,\mathrm{d}s. \tag{471}
$$

As in Section 8 for the compressible Euler equations, here we set $\mathbf{i}(\mathbf{u}_h) = \mathbf{u}_\Gamma(\mathbf{u}_h)$ and $J_\Gamma(\mathbf{u}_h) = J(\mathbf{u}_\Gamma(\mathbf{u}_h))$; $r_J(\mathbf{u}_h)$ will be specified later. Noting that $\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ holds for the exact solution $\mathbf{u}$, $\tilde{J}(\cdot)$ in (471) is a consistent modification of $J(\cdot)$ provided that $\mathbf{u}$ satisfies $r_J(\mathbf{u}) = 0$, see also (288).

The discrete adjoint problem is given by: find $\mathbf{z}_h\in\mathbf{V}_{h,p}^d$ such that

$$
N_h'[\mathbf{u}_h](\mathbf{w}_h,\mathbf{z}_h) = \tilde{J}'[\mathbf{u}_h](\mathbf{w}_h) \quad \forall\mathbf{w}_h\in\mathbf{V}_{h,p}^d, \tag{472}
$$

where $N_h'[\mathbf{u}](\mathbf{w},\mathbf{z})$ is given by

$$
\begin{aligned}
N_h'[\mathbf{u}](\mathbf{w},\mathbf{z}) \;=\; & -\int_\Omega (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}]\mathbf{w}) : \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathcal{H}_{\mathbf{u}^+}'(\mathbf{u}^+,\mathbf{u}^-,\mathbf{n}^+)\mathbf{w}^+[\![\mathbf{z}]\!]\cdot\mathbf{n}\,\mathrm{d}\mathbf{s}\\
& + \int_\Omega \left(G'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{u}\right) : \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x} + \int_\Omega (G(\mathbf{u})\nabla_h\mathbf{w}) : \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x}\\
& - \int_{\Gamma_\mathcal{I}} \{\!\{G'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{u}\}\!\} : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s - \int_{\Gamma_\mathcal{I}} \{\!\{G(\mathbf{u})\nabla_h\mathbf{w}\}\!\} : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s\\
& - \int_{\Gamma_\mathcal{I}} \{\!\{\left(G^\top\right)'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{z}\}\!\} : \underline{[\![\mathbf{u}]\!]}\,\mathrm{d}s - \int_{\Gamma_\mathcal{I}} \{\!\{G^\top(\mathbf{u})\nabla_h\mathbf{z}\}\!\} : \underline{[\![\mathbf{w}]\!]}\,\mathrm{d}s\\
& + \int_{\Gamma_\mathcal{I}} \boldsymbol{\delta}'[\mathbf{u}](\mathbf{w}) : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s + N_{\Gamma,h}'[\mathbf{u}](\mathbf{w},\mathbf{z}).
\end{aligned}
$$

Using integration by parts this can be rewritten as follows

$$
\begin{aligned}
&-\int_\Omega \mathbf{w}\left(\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}]\right)^\top \nabla_h \mathbf{z}\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathbf{w}^+\left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}^+,\mathbf{u}^-,\mathbf{n}^+)\right)^\top [\![\mathbf{z}]\!]\cdot\mathbf{n}\,\mathrm{d}\mathbf{s}\\
&+\int_\Omega \mathbf{w}\left(G'[\mathbf{u}]\nabla_h\mathbf{u}\right)^\top \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x} - \int_\Omega \mathbf{w}\nabla_h\cdot\left(G^\top(\mathbf{u})\nabla_h\mathbf{z}\right)\,\mathrm{d}\mathbf{x}\\
&-\frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\left(G'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{u}\right):\underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s - \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}(G(\mathbf{u})\nabla_h\mathbf{w}):\underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s\\
&-\frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\left(\left(G^\top\right)'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{z}\right):\underline{[\![\mathbf{u}]\!]}\,\mathrm{d}s + \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathbf{w}[\![G^\top(\mathbf{u})\nabla_h\mathbf{z}]\!]\,\mathrm{d}s\\
&+\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\boldsymbol{\delta}'[\mathbf{u}](\mathbf{w})[\![\mathbf{z}]\!]\,\mathrm{d}\mathbf{s} + \int_{\mathbf{\Gamma}}(\mathbf{w}\otimes\mathbf{n}):\left(\mathbf{G}^\top(\mathbf{u})\nabla_{\mathbf{h}}\mathbf{z}\right)\,\mathrm{d}\mathbf{s} + \mathbf{N}'_{\mathbf{\Gamma},\mathbf{h}}[\mathbf{u}](\mathbf{w},\mathbf{z}).
\end{aligned}
$$

Hence, the discrete adjoint problem (472) in adjoint residual form is given as follows: find $\mathbf{z}_h\in\mathbf{V}^d_{h,p}$ such that

$$
\begin{aligned}
\int_\Omega \mathbf{w}\cdot\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathbf{w}\cdot\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) + \nabla\mathbf{w}:\boldsymbol{\rho}^*[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}s\\
+\int_\Gamma \mathbf{w}\cdot\mathbf{r}^*_\Gamma[\mathbf{u}_h](\mathbf{z}_h) + \nabla\mathbf{w}:\boldsymbol{\rho}^*_\Gamma[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}s = 0 \quad \forall\mathbf{w}\in\mathbf{V}^d_{h,p},
\end{aligned}
\tag{473}
$$

where the adjoint residuals are given by

$$
\begin{aligned}
\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h) &= \left(\mathcal{F}^c_{\mathbf{u}}(\mathbf{u}_h) - G'[\mathbf{u}_h]\nabla\mathbf{u}_h\right)^\top\nabla_h\mathbf{z}_h + \nabla_h\cdot\left(G^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right) \quad \text{in } \kappa,\kappa\in\mathcal{T}_h,\\
\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) &= -\left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+)\right)^\top[\![\mathbf{z}_h]\!]\cdot\mathbf{n} - \frac{1}{2}[\![G^\top(\mathbf{u}_h)\nabla\mathbf{z}_h]\!] - \left(\boldsymbol{\delta}'[\mathbf{u}_h]\right)^\top[\![\mathbf{z}_h]\!]\\
&\quad + \frac{1}{2}\left(G'[\mathbf{u}_h]\nabla\mathbf{u}_h\right)^\top\underline{[\![\mathbf{z}_h]\!]} + \frac{1}{2}\left(G'[\mathbf{u}_h]\underline{[\![\mathbf{u}_h]\!]}\right)^\top\nabla_h\mathbf{z}_h \quad \text{on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h,\\
\boldsymbol{\rho}^*[\mathbf{u}_h](\mathbf{z}_h) &= \frac{1}{2}G^\top[\mathbf{u}_h]\underline{[\![\mathbf{z}_h]\!]} \qquad\qquad \text{on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h.
\end{aligned}
\tag{474}
$$

The adjoint boundary residuals $\mathbf{r}^*_\Gamma$ and $\boldsymbol{\rho}^*_\Gamma$ will be specified below. Recalling that $\mathcal{F}^v_{\mathbf{u}} = G'[\mathbf{u}]\nabla\mathbf{u}$ and $\mathcal{F}^v_{\nabla\mathbf{u}} = G(\mathbf{u})$ we see that the exact solution $\mathbf{z}$ to the continuous adjoint problem (466) satisfies $\mathbf{R}^*[\mathbf{u}](\mathbf{z}) = 0$. In the two lines in (474) representing the face residual term $\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h)$ we recognize the jump $-(\mathcal{H}'_{\mathbf{u}^+})^\top[\![\mathbf{z}_h]\!]\cdot\mathbf{n}$ due to the convective part of the equations, cf. (439), furthermore the second term in the first line corresponding to the adjoint face residuals of the Poisson's equation, cf. (309), and finally the two terms in the second line due to the nonlinearity of the compressible Navier-Stokes equations. Whereas the last term in the second line vanishes for a smooth exact primal solution $\mathbf{u}$, all other terms vanish for the exact solution $\mathbf{z}$ to the adjoint problem (466). Thereby, the adjoint solution $\mathbf{z}$ satisfies $\mathbf{r}^*[\mathbf{u}](\mathbf{z}) = 0$. Furthermore, $\mathbf{z}$ satisfies $\boldsymbol{\rho}^*[\mathbf{u}](\mathbf{z}) = 0$. In summary, we see that, as for the Poisson's equation, the element and interior face terms of the SIPG discretization are adjoint consistent.

The boundary terms of the discrete adjoint problem are given by

$$\begin{aligned}
N'_{\Gamma,h}[\mathbf{u}_h](\mathbf{w},\mathbf{z}_h)+&\int_\Gamma (\mathbf{w}\otimes\mathbf{n}) : \left(G_\Gamma^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right)\,\mathrm{d}s \equiv\\
&+\int_\Gamma \mathbf{n}\cdot\left(\mathcal{F}^c_{\Gamma,\mathbf{u}}[\mathbf{u}_h](\mathbf{w})\right)\mathbf{z}_h\,\mathrm{d}s+\int_\Gamma \boldsymbol{\delta}'_\Gamma[\mathbf{u}_h](\mathbf{w})\cdot\mathbf{z}\,\mathrm{d}s,\\
&-\int_\Gamma \mathbf{n}\cdot\left(\mathcal{F}^v_{\Gamma,\mathbf{u}}[\mathbf{u}_h,\nabla_h\mathbf{u}_h](\mathbf{w})+\mathcal{F}^v_{\Gamma,\nabla\mathbf{u}}[\mathbf{u}_h,\nabla_h\mathbf{u}_h](\nabla_h\mathbf{w})\right)\mathbf{z}_h\,\mathrm{d}s\\
&-\int_\Gamma \left(\left(\left(G_\Gamma^\top\right)'[\mathbf{u}_h]\mathbf{w}\right)\nabla_h\mathbf{z}_h\right) : (\mathbf{u}_h-\mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n}\,\mathrm{d}s\\
&-\int_\Gamma \left(G_\Gamma^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right) : \left(\mathbf{w}-\mathbf{u}'_\Gamma[\mathbf{u}_h]\mathbf{w}\right)\otimes\mathbf{n}\,\mathrm{d}s\\
&+\int_\Gamma (\mathbf{w}\otimes\mathbf{n}) : \left(G_\Gamma^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right)\,\mathrm{d}s=\tilde{J}'[\mathbf{u}_h](\mathbf{w}).
\end{aligned}$$

Thus the adjoint boundary residuals in (473) on Γ_W are given by

$$\begin{aligned}
\mathbf{r}^*_\Gamma[\mathbf{u}_h](\mathbf{z}_h) =&\tfrac{1}{C_\infty}\left(p_\mathbf{u}\,\mathbf{n}-\underline{\tau}_\mathbf{u}\,\mathbf{n}\right)\cdot\psi-\left(\mathbf{n}\cdot\left(\mathcal{F}^c_{\Gamma,\mathbf{u}}-\mathcal{F}^v_{\Gamma,\mathbf{u}}\right)\right)^\top\mathbf{z}_h-\mathbf{n}\cdot\left(G_\Gamma^\top\nabla\mathbf{z}_h\right)\\
&+r'_J[\mathbf{u}_h]-\left(\boldsymbol{\delta}'_\Gamma[\mathbf{u}_h]\right)^\top\mathbf{z}_h+\left(G'_\Gamma[\mathbf{u}_h] : (\mathbf{u}_h-\mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_h\\
&+\left(G_\Gamma(\mathbf{u}_h) : \left(I-\mathbf{u}'_\Gamma[\mathbf{u}_h]\right)\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_h, \qquad (475)\\
\boldsymbol{\rho}^*_\Gamma[\mathbf{u}_h](\mathbf{z}_h) =&-\frac{1}{C_\infty}\left(\underline{\tau}_{\nabla\mathbf{u}}\mathbf{n}\right)\cdot\psi+\left(\mathbf{n}\cdot\mathcal{F}^v_{\Gamma,\nabla\mathbf{u}}\right)^\top\mathbf{z}_h. \qquad (476)
\end{aligned}$$

We recall (468), $\mathcal{F}^v_{\Gamma,\nabla\mathbf{u}}=G_\Gamma(\mathbf{u})$, and see that the exact solutions $\mathbf{u}$ and $\mathbf{z}$ to the primal problem (440) and the continuous adjoint problem (466)-(470) satisfy $\boldsymbol{\rho}^*_\Gamma[\mathbf{u}](\mathbf{z})=0$.

We now choose the modification $r_J(\mathbf{u}_h)$ of the target functional in (471) as follows

$$r_J(\mathbf{u}_h)=\boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+):\mathbf{z}_\Gamma\otimes\mathbf{n}-\left(G_\Gamma^\top(\mathbf{u}_h^+)\nabla_h\mathbf{z}_\Gamma\right):\left(\mathbf{u}_h^+-\mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)\otimes\mathbf{n}, \qquad (477)$$

with Fréchet derivative

$$\begin{aligned}
r'_J[\mathbf{u}_h](\mathbf{w})=\boldsymbol{\delta}'_\Gamma[\mathbf{u}_h](\mathbf{w}):\mathbf{z}_\Gamma\otimes\mathbf{n}-\left(G'_\Gamma[\mathbf{u}_h] : (\mathbf{u}-\mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_\Gamma\\
-\left(G_\Gamma(\mathbf{u}_h) : \left(I-\mathbf{u}'_\Gamma[\mathbf{u}_h]\right)\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_\Gamma.
\end{aligned}$$

As the exact solution $\mathbf{u}$ to the primal problem satisfies $\mathbf{u}_\Gamma(\mathbf{u})=\mathbf{u}$, we have $r_J(\mathbf{u})=0$. Hence, (477) is a consistent modification of the target functional. Recalling (467), we see that the exact solutions $\mathbf{u}$ and $\mathbf{z}$ satisfy

$$\begin{aligned}
\mathbf{r}^*_\Gamma[\mathbf{u}](\mathbf{z})=\left(\mathbf{n}\cdot\boldsymbol{\delta}'_\Gamma[\mathbf{u}]\right)^\top(\mathbf{z}_\Gamma-\mathbf{z})-\left(G'[\mathbf{u}] : (\mathbf{u}-\mathbf{u}_\Gamma(\mathbf{u}))\otimes\mathbf{n}\right)^\top(\nabla\mathbf{z}_\Gamma-\nabla\mathbf{z})\\
-\left(G(\mathbf{u}) : \left(I-\mathbf{u}'_\Gamma[\mathbf{u}]\right)\otimes\mathbf{n}\right)^\top(\nabla\mathbf{z}_\Gamma-\nabla\mathbf{z}).
\end{aligned}$$

Furthermore, setting $\mathbf{z}_\Gamma=\mathbf{z}$ on Γ_W we obtain $\mathbf{r}^*_\Gamma[\mathbf{u}](\mathbf{z})=0$ and conclude that the discretization of boundary terms is adjoint consistent.

Due to $\mathbf{n}\cdot\left(G_\Gamma^\top(\mathbf{u}_h^+)\nabla\mathbf{z}\right)=\mathbf{n}\cdot((\mathcal{F}^v_{\nabla\mathbf{u}})^\top\nabla\mathbf{z})=0$ on Γ_W the second term in (477) vanishes. Furthermore, on adiabatic boundaries Γ_{adia} we have $(\mathbf{u}_h^+-\mathbf{u}_\Gamma(\mathbf{u}_h^+))_i=0$, $i=1,4$, and on isothermal boundaries Γ_{iso} we have $(\mathbf{u}_h^+-\mathbf{u}_\Gamma(\mathbf{u}_h^+))_1=0$. Together with (470), the consistent modification (477) reduces to

$$\begin{aligned}
r_J(\mathbf{u}_h)&=\boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+):\mathbf{z}_\Gamma\otimes\mathbf{n}\\
&=\left(\boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+)\right)_2\tfrac{1}{C_\infty}\psi_1 n_1+\left(\boldsymbol{\delta}_\Gamma(\mathbf{u}_h^+)\right)_3\tfrac{1}{C_\infty}\psi_2 n_2,
\end{aligned} \qquad (478)$$

which completes the adjoint consistency analysis of the interior penalty discontinuous Galerkin discretization of the compressible Navier-Stokes equations. Finally, we note that the consistent modification $r_J(\mathbf{u}_h)$ given in (478) for SIPG reduces to $r_J(\mathbf{u}_h) = \delta \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right) \cdot \mathbf{z}_\Gamma$ which corresponds to the IP modification of target functionals for the Poisson's equation, where $r_J(u_h) = \delta(u_h - g_D)z_\Gamma$, with $z_\Gamma = -j_D$, see (311).

In summary, we have shown that the adjoint element and interior residuals $\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h)$, $\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h)$ and $\boldsymbol{\rho}^*[\mathbf{u}_h](\mathbf{z}_h)$, see (474), vanish for the exact solutions $\mathbf{u}$ and $\mathbf{z}$ to (440) and (466), respectively. Additionally, using an adjoint consistent treatment of convective and diffusive boundary fluxes,

$$\mathbf{n} \cdot \mathcal{F}_\Gamma^c(\mathbf{u}_h^+) = \mathbf{n} \cdot \mathcal{F}^c(\mathbf{u}_\Gamma(\mathbf{u}_h^+)), \qquad \mathbf{n} \cdot \mathcal{F}_\Gamma^v(\mathbf{u}_h^+) = \mathbf{n} \cdot \mathcal{F}^v(\mathbf{u}_\Gamma(\mathbf{u}_h^+), \nabla_h \mathbf{u}_h^+), \tag{479}$$

and using the following consistent modification of the target functional,

$$\tilde{J}(\mathbf{u}_h) = J(\mathbf{u}_\Gamma(\mathbf{u}_h)) + \int_{\Gamma_W} \boldsymbol{\delta}_\Gamma(\mathbf{u}_h) \, : \mathbf{z}_\Gamma \otimes \mathbf{n} \, \mathrm{d}s, \tag{480}$$

with $\mathbf{z}_\Gamma = \frac{1}{C_\infty}(0, \psi_1, \psi_2, 0)^\top$, for $J(\cdot)$ representing a total force coefficient defined in (464), the adjoint boundary residuals $\mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h)$ and $\boldsymbol{\rho}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h)$, see (475) and (476), vanish for the exact solutions $\mathbf{u}$ and $\mathbf{z}$. Thereby, using the modifications given in (479) and (480), we recover an adjoint consistent discontinuous Galerkin discretizations of the compressible Navier-Stokes equations in conjunction with total force coefficients.

We note that arguments given in [36] *en route* to obtaining an adjoint consistent discretization based on the BR2 scheme [7] can also be covered within the presented framework and lead to analogous modifications. Furthermore, we note that numerical experiments in [27] have confirmed that in contrast to the original formulation in [31] the discrete adjoint solution to the adjoint consistent discretization is entirely smooth. Furthermore, numerical tests on globally refined meshes have shown that the adjoint consistent discretization is by a factor of 2-400 more accurate measured in terms of viscous force coefficients than the original formulation. Also a significantly improved order of convergence has been observed. Finally, we note that in [33], see also Section 9.4, the interior penalty DG discretization with adjoint consistent discretization of boundary conditions and an improved penalty term (457) has shown to be of optimal order. In fact, the accuracy of the IP discretization in [33] is comparable to the accuracy of the BR2 discretization in [7, 8] while the residual computation of IP is significantly cheaper than that of BR2 which requires the additional evaluation of local lifting operators.

9.4 Numerical results

Flow over a flat plate We begin by investigating the accuracy of higher order DG discretizations in resolving laminar boundary layers. To this end, we consider a Mach 0.01 flow with Reynolds number 10000 horizontally passing over a flat plate of length $l = 2$. The boundary layer solution to this problem can be approximated using Blasius' solution, see [39], for example. In Figure 14, by [31], we compare the numerical solution computed with the DG(p) method for $1 \leq p \leq 3$, at $x = \frac{l}{2} = 1$ and a local Reynolds number $Re_x = 5000$, with the Blasius solution ($\eta = y\sqrt{u_\infty/(\nu x)} = \frac{y}{x}\sqrt{Re_x}$ versus u/u_∞, cf. [39]) on a sequence of rather coarse computational meshes. On the coarsest mesh, which has about one or two elements within the boundary layer, we see that the DG solution computed with $p = 1, 2$ are not very close to the Blasius solution; increasing the polynomial order to $p = 3$ clearly yields a dramatic improvement in the underlying computed numerical solution. On the next finer mesh, where three elements are placed within the boundary layer, the bilinear approximation is still not very accurate, though now both the computed solution with $p = 2, 3$ are in excellent agreement with the Blasius solution. On the subsequent two meshes we clearly observe that the DG approximation with bilinear elements ($p = 1$) finally starts to coincide with the Blasius solution, at least on a macroscopic level. A more detailed view of the numerical solution on these

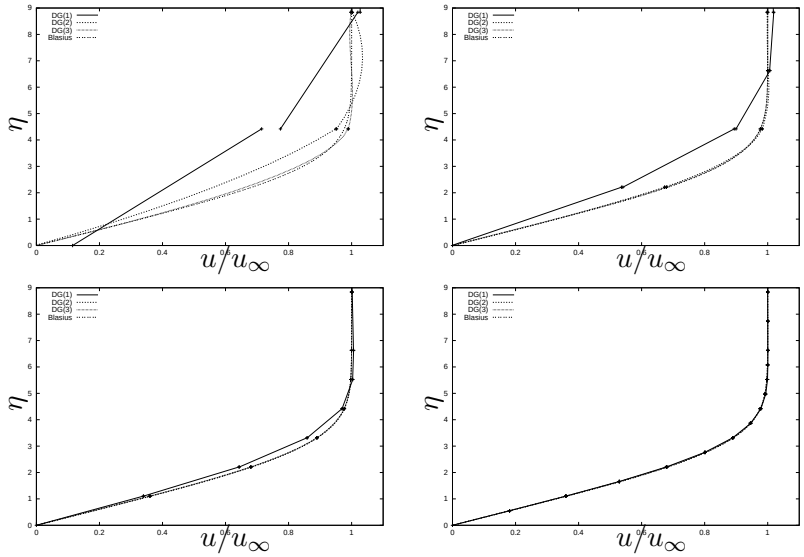


Figure 14: DG(p), $1 \leq p \leq 3$, solutions in comparison with the Blasius solution ($\eta = y\sqrt{u_\infty/(\nu x)} = \frac{y}{x}\sqrt{Re_x}$ versus u/u_∞) on a sequence of meshes with an increasing number of elements, [31].

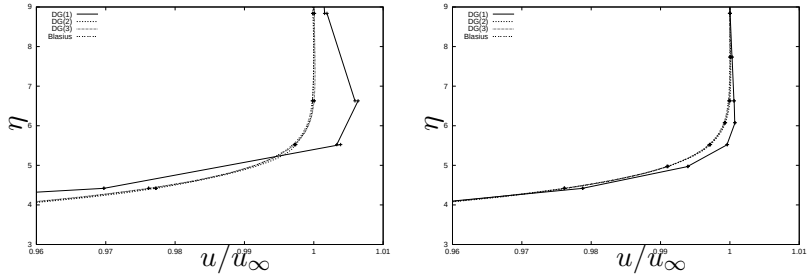


Figure 15: Zoom of the DG(p), $1 \leq p \leq 3$, solutions on the two finest grids, [31].

latter two finer meshes is shown in the zoom depicted in Figure 15. Here, we see that there is still a significant difference between the Blasius solution and the computed discontinuous Galerkin solution with $p = 1$. Indeed, these figures clearly highlight the substantial gains in accuracy attained when higher–order polynomial degrees are employed with the DG method. This is further highlighted in Table 1, where we summarize the number of elements and the number of degrees of freedom, orthogonal to the wall, which are required by the DG method for each polynomial degree in order to resolve the boundary layer to a sufficient accuracy that the error in computed viscous stress forces exerted on the wall are within 5% of that computed with the Blasius solution.

	DG(1)	DG(2)	DG(3)
elements	36	5	3
DoFs	72	15	12

Table 1: Number of elements and degrees of freedom in the boundary layer required by DG(p), $1 \leq p \leq 3$, discretizations for approximating the viscous force up to 5%, [31].

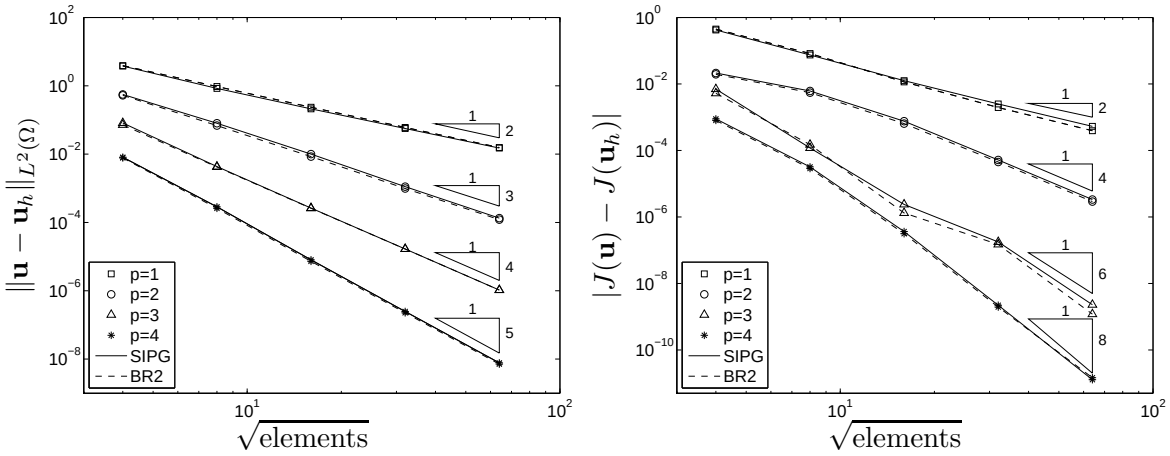


Figure 16: Flow in a square domain. Comparison of the SIPG and BR2 methods when the error is measured in terms of: (left) $L^2(\Omega)$-norm; (right) Weighted mean-value functional $J(\cdot)$, [33].

Flow in a square domain In the following we examine the experimental order of convergence of the interior penalty DG discretization, [33]. To this end we consider following model problem: let $\Omega = (0,\pi)^2$, and supplement the compressible Navier–Stokes equations (440) with an inhomogeneous forcing function $\mathbf{f}$, which is chosen so that the analytical solution to (440) is given by

$$\mathbf{u}(\mathbf{x}) = (\sin(2(x_1+x_2))+4, \sin(2(x_1+x_2))/5+4, \sin(2(x_1+x_2))/5+4, (\sin(2(x_1+x_2))+4)^2)^\top,$$

where the dynamic viscosity coefficient μ has been set to $1/10$. This represents a modification of the (unsteady) test problem employed in the article [19]. In this section we shall be interested in measuring the discretization error in terms of both the $L^2(\Omega)$-norm as well as in terms of a given target functional $J(\cdot)$. In the latter case, we consider the weighted mean-value of the density, i.e.,

$$J(\mathbf{u}) \equiv J_\Omega(\mathbf{u}) = \int_\Omega u_1 \psi \, \mathrm{d}\mathbf{x},$$

where $\psi = \sin(\pi x)\sin(\pi y)$; thereby, the true value of the functional is given by $J(\mathbf{u}) = 1.1685876486$.

In Figure 16(a) we present a comparison of the error in the $L^2(\Omega)$-norm with the (square root of the) number of elements for $p = 1,2,3,4$, employing both the SIPG method with $C_{\mathrm{IP}} = 10$ and the Bassi–Rebay method (BR2) with $C_{\mathrm{BR2}} = 4$. In both cases, we observe that $\|\mathbf{u}-\mathbf{u}_h\|_{L^2(\Omega)}$ converges to zero at the expected optimal rate $\mathcal{O}(h^{p+1})$ as the mesh is refined for each fixed p. Moreover, from Figure 16(b) we observe that the error in the computed target functional $J(\cdot)$ behaves (approximately) like $\mathcal{O}(h^{2p})$, for each fixed p, as the mesh is uniformly refined for both of the discretization schemes considered. These rates of convergence for both the $L^2(\Omega)$-norm of the error and the error in the computed target functional $J(\cdot)$ are in complete agreement with the corresponding convergence behavior we would expect for the SIPG and BR2 methods when applied to a linear convection–diffusion problem; see [23], for example, for the analysis of general interior penalty DGFEMs for second–order partial differential equations with non-negative characteristic form. We remark that in terms of accuracy, for a given number of elements, or equivalently, for a fixed number of degrees of freedom, both the SIPG scheme and the BR2 method perform in a comparable manner, with the latter scheme being, in general, slightly more accurate. However, in terms of computational resources, the time required to assemble the residual vector of the BR2 method, which is the most computationally intensive part of the flow solver, when explicit time-stepping schemes are employed, is significantly more expensive than the computation

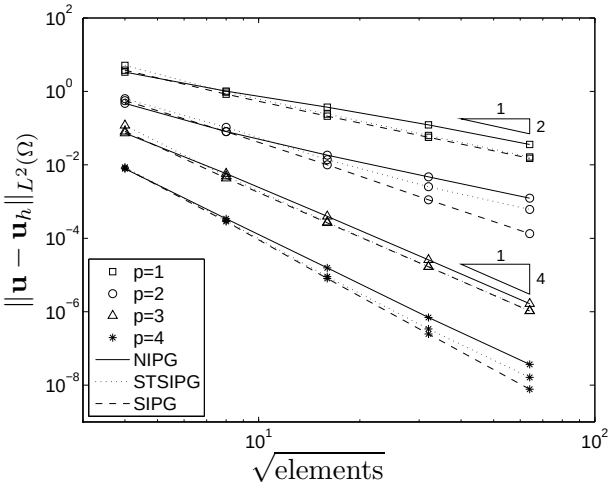


Figure 17: Flow in a square domain. Comparison of the SIPG, NIPG, and STSIPG methods when the error is measured in terms of the $L^2(\Omega)$-norm, [33].

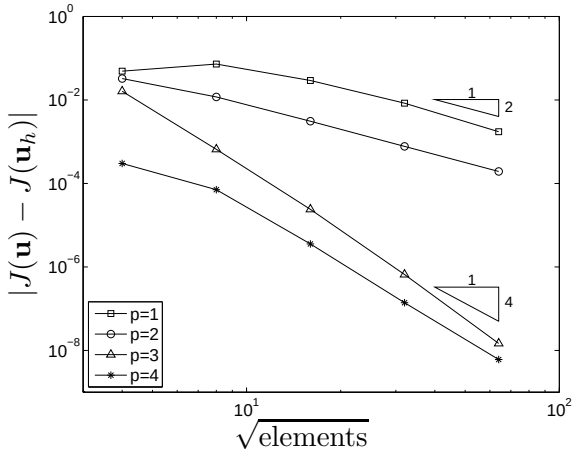


Figure 18: Flow in a square domain. Convergence of the NIPG scheme with respect to $J(\cdot)$ with h–refinement, [33].

of the corresponding quantity when the SIPG scheme is employed. More precisely, for (bi)-linear, elements, i.e., $p = 1$, the BR2 method is around 38% more expensive than the SIPG scheme; this overhead increases as the underlying polynomial degree is enriched. Indeed, for $p = 2$, the BR2 method is approximately 47% more expensive, and for $p = 3$ and $p = 4$ the additional work rises to around 55%. This increase in the cpu times when the BR2 method is employed is attributed to the computation of the lifting operator on each face of the computational mesh.

Finally, in this section we compare the performance of the SIPG method with both the corresponding NIPG formulation of the underlying scheme, together with the interior penalty method outlined in the article [31]; we shall refer to this latter scheme as the standard SIPG (STSIPG) method. To this end, in Figure 17 we plot the $L^2(\Omega)$-norm of the error against the (square root of the) number of elements for $p = 1, 2, 3, 4$ using each of the above schemes. In contrast to the SIPG and BR2 methods, we now observe that $\|\mathbf{u}-\mathbf{u}_h\|_{L^2(\Omega)}$ behaves like $\mathcal{O}(h^{p+1})$ for odd p and like $\mathcal{O}(h^p)$ for even p when either the NIPG method or the STSIPG scheme are employed. The sub-optimal convergence observed when employing these two schemes is attributed to the lack of smoothness

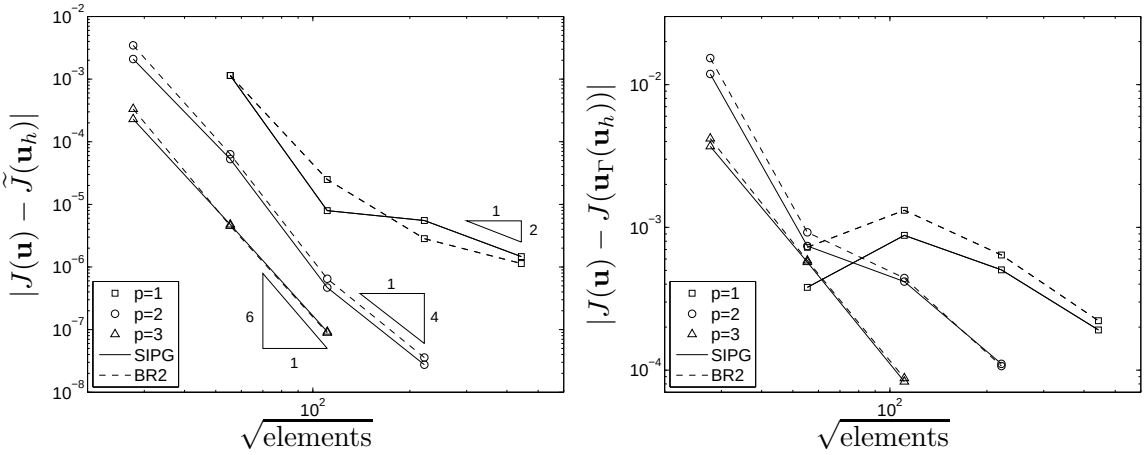


Figure 19: Viscous flow around NACA0012 airfoil. Comparison of the SIPG and BR2 methods employing: (left) Adjoint consistent reformulation of the drag functional; (right) Adjoint consistent reformulation of the drag functional excluding the penalty terms, [33].

in the resulting adjoint problems, cf. [23, 28]. Moreover, the same behavior is also observed in the functional setting; indeed, for the NIPG scheme, from Figure 18 we see that $|J(\mathbf{u}) - J(\mathbf{u}_h)|$ tends to zero at (approximately) the rate $\mathcal{O}(h^{p+1})$ for odd p and $\mathcal{O}(h^p)$ for even p, as the mesh is uniform refined. Analogous behavior is also observed when the error in the computed target functional $J(\cdot)$ is evaluated using the STSIPG method; for brevity, these numerics have been omitted.

Viscous flow around a NACA0012 airfoil In this example, we consider a subsonic viscous flow around a NACA0012 airfoil. At the farfield (inflow) boundary we specify a Mach 0.5 flow at a zero angle of attack, i.e. $\alpha = 0^\circ$, with Reynolds number $\mathrm{Re} = 5000$; on the walls of the airfoil geometry, we impose a zero heat flux (adiabatic) no-slip boundary condition. This is a standard laminar test case which has been investigated by many other authors, cf. [5, 31], for example. The solution to this problem consists of a strictly subsonic flow which is symmetric about the x-axis.

Here, we consider the estimation of the drag coefficient c_{d}; i.e., the target functional is given by

$$J(\mathbf{u}) = \int_\Gamma j(\mathbf{u})\,\mathrm{d}s = \frac{1}{C_\infty}\int_{\Gamma_W} (p\,\mathbf{n} - \underline{\tau}\,\mathbf{n})\cdot\psi_d\,\mathrm{d}s,$$

where $j(\mathbf{u}) = \frac{1}{C_\infty} p\,\mathbf{n}\cdot\psi_d$ on Γ_W and $j(\mathbf{u}) \equiv 0$ elsewhere, cf. (464) and (421). We remark that the adjoint consistency of the SIPG scheme is based on the consistent reformulation of $J(\cdot)$ defined in (480). With this in mind, in Figure 19(a) we present a comparison of the error in the computed target functional with the (square root of the) number of elements for $p = 1, 2, 3$, employing both the SIPG method with $C_{\mathrm{IP}} = 10$ and the Bassi–Rebay method (BR2) with $C_{\mathrm{BR2}} = 4$. In both cases, we observe that, asymptotically, at least, $|J(\mathbf{u}) - \widetilde{J}(\mathbf{u}_h)|$ converges to zero at the expected optimal rate $\mathcal{O}(h^{2p})$ as the mesh is refined for each fixed p. Moreover, as before, we note that in terms of accuracy, for a given number of elements, or equivalently, for a fixed number of degrees of freedom, both the SIPG scheme and the BR2 method perform in a comparable manner, though as already noted, the SIPG scheme requires less computational effort to attain the computed solution. To highlight the necessity of the consistent reformulation of the original target functional $J(\cdot)$ through the additional of the term involving the penalty function $\underline{\delta}_\Gamma(\cdot)$, cf. (480) for the definition of $\widetilde{J}(\cdot)$, in Figure 19(b) we present a comparison of $|J(\mathbf{u}) - J(\mathbf{u}_\Gamma(\mathbf{u}_h))|$ with the (square root of the) number of elements for $p = 1, 2, 3$ employing both the SIPG and BR2 schemes. In this case, we now observe that there is a significant deterioration of the error for a given mesh size and polynomial

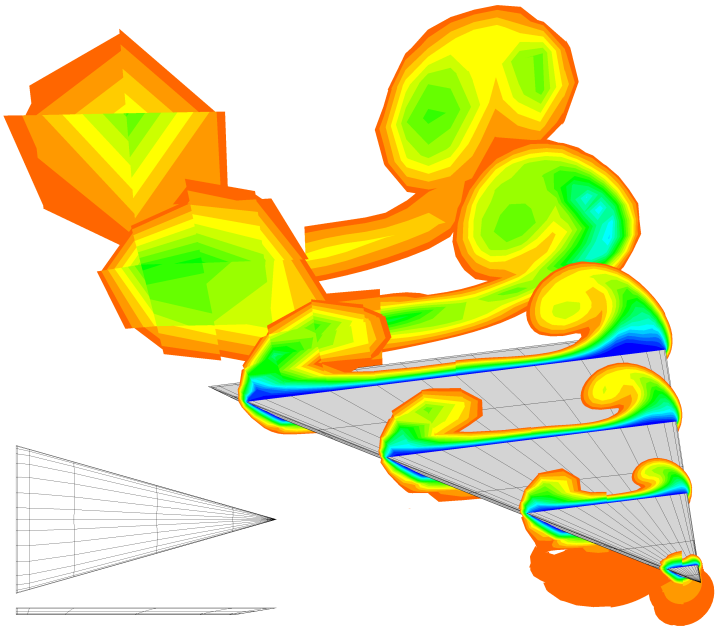


Figure 20: Laminar flow around a delta wing. Geometry of the delta wing and the Mach number isolines on several slices of the flow field computed based on a DG(p) discretization with $p = 1$ (2nd order) on the left wing and with $p = 4$ (5th order) on the right wing.

order when compared to the corresponding results when the penalty function modification of the target functional has been included. Indeed, comparing Figures 19(a) and 19(b), we see that the inclusion of the penalty function modification in the definition of $\widetilde{J}(\cdot)$ leads to around 2–3 orders of magnitude improvement in the computed error in the drag.

Laminar flow around a delta wing As a final example we consider a laminar flow around a delta wing. At the farfield (inflow) boundary we specify a Mach 0.3 flow at an angle $\alpha = 12.5^\circ$ of attack with Reynolds number $\mathrm{Re} = 4000$. On the walls of the delta wing we impose an isothermal wall boundary condition. This is the BTC3 test case of the EU-project ADIGMA [1]. Figure 20 shows the Mach number isolines on several slices of the flow field. The flow is computed on a coarse mesh of 3264 elements. The corresponding surface mesh is depicted on the wing geometry. The flow field on the left part of the delta wing is based on a (2nd order) DG discretization with $p = 1$ and the right part is based on a (5th order) DG discretization with $p = 4$. We see that the 5th order flow solution provides a good resolution of the primary and secondary vortices. Furthermore, the vortices are tracked over some distance behind the wing. In contrast to that, the primary and secondary vortices are almost indistinguishable in the 2nd order flow solution. Here, the vortices merge and are damped out far too early. Already after a short distance behind the wing the original vortex system is lost due to numerical viscosity. Figure 21 compares the DG(p) solutions for $p = 1, 2, 3$. Here, the error is given in terms of the drag, lift and moment coefficients, c_l, c_d and c_m, respectively, and is plotted against the number of degrees of freedoms (DoFs) per equation. The horizontal line in each error plot in Figure 21 represents the error tolerances,

$$
\begin{aligned}
|J_{c_l}(\mathbf{u}) - J_{c_l}(\mathbf{u})| &\leq 10^{-2}, \\
|J_{c_d}(\mathbf{u}) - J_{c_d}(\mathbf{u})| &\leq 10^{-3}, \\
|J_{c_m}(\mathbf{u}) - J_{c_m}(\mathbf{u})| &\leq 10^{-3},
\end{aligned}
$$

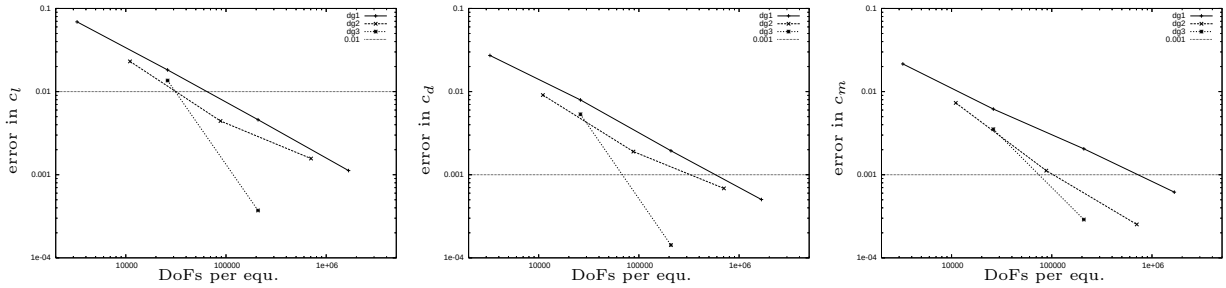


Figure 21: Laminar flow around a delta wing. Comparison of the DG(p), $p = 1, 2, 3$, solutions. The error is given in terms of the drag, lift and moment coefficients, c_l, c_d and c_m, respectively.

as defined in ADIGMA project. Here, we see a clear advantage of using higher order DG approximations over 2nd order approximations. In fact, in terms of DoFs the DG(3) discretization is about a factor of 10 more efficient than DG(1). This advantage further increases when stronger accuracy requirements, i.e. smaller error tolerances, are imposed.

Acknowledgments

The author acknowledges the financial support by the President's Initiative and Networking Fund of the Helmholtz Association of German Research Centres, and the financial support of the European Union, under the `ADIGMA` project, [1]. If not indicated differently, computations have been performed using the DG flow solver `PADGE` [29] based on the `deal.II` library [3, 4].

References

[1] ADIGMA. Adaptive higher-order variational methods for aerodynamic applications in industry. A specific targeted research project of the sixth European community framework programme. See http://www.dlr.de/as/desktopdefault.aspx/tabid-2035.

[2] D. Arnold, F. Brezzi, B. Cockburn, and L. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.

[3] W. Bangerth, R. Hartmann, and G. Kanschat. deal.II – A general purpose object oriented finite element library. *ACM Transactions on Mathematical Software*, 33(4), Aug. 2007.

[4] W. Bangerth, R. Hartmann, and G. Kanschat. `deal.II` *Differential Equations Analysis Library, Technical Reference*. http://www.dealii.org/, 6.1 edition, May 2008. First edition 1999.

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

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

[7] F. Bassi and S. Rebay. GMRES discontinuous Galerkin solution of the compressible Navier-Stokes equations. In B. Cockburn, G. Karniadakis, and C.-W. Shu, editors, *Discontinuous Galerkin Methods*, volume 11, pages 197–208. Springer, 1999.

[8] F. Bassi and S. Rebay. Numerical evaluation of two discontinuous Galerkin methods for the compressible Navier-Stokes equations. *Int. J. Numer. Meth. Fluids*, 40:197–207, 2002.

[9] C. Baumann and J. Oden. A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations. *International Journal for Numerical Methods in Fluids*, 31:79–95, 1999.

[10] C. Baumann and J. Oden. An adaptive-order discontinuous Galerkin method for the solution of the Euler equations of gas dynamics. *International Journal for Numerical Methods in Engineering*, 47:61–73, 2000.

[11] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Num. Meth. Part. Diff. Eq.*, 16(4):365–378, 2000.

[12] F. Brezzi, D. Marini, and E. Süli. Residual-free bubbles for advection-diffusion problems: the general error analysis. *Numerische Mathematik*, 85(1):31–47, 2000.

[13] F. Brezzi, L. D. Marini, and E. Süli. Discontinuous Galerkin methods for first-order hyperbolic problems. *Math. Models and Methods in Appl. Sci.*, 14(12):1893–1903, 2004.

[14] A. Cangiani and E. Süli. Enhanced residual-free bubble method for convection-diffusion problems. *Int. J. Numer. Meth. Fluids*, 47(10–11):1307–1313, 2005.

[15] G. Chiocchia. Exact solutions to transonic and supersonic flows. Technical Report AR-211, AGARD, 1985.

[16] V. Dolejsi. On the discontinuous Galerkin method for the numerical solution of the Navier-Stokes equations. *Int. J. Numer. Meth. Fluids*, 45:1083–1106, 2004.

[17] K. J. Fidkowski and D. L. Darmofal. A triangular cut-cell adaptive method for high-order discretizations of the compressible Navier-Stokes equations. *J. Comput. Physics*, 225:1653–1672, 2007.

[18] K. J. Fidkowski, T. A. Oliver, J. Lu, and D. L. Darmofal. *p*-multigrid solution of high-order discontinuous Galerkin discretizations of the compressible Navier-Stokes equations. *J. Comp. Phys.*, 207(1):92–113, July 2005.

[19] G. Gassner, F. Lörcher, and C.-D. Munz. A discontinuous Galerkin scheme based on a space-time expansion II. Viscous flow equations in multi-dimensions. *J. Sci. Comput.*, 34(3):260–286, 2008.

[20] M. Giles and N. Pierce. Adjoint equations in CFD: duality, boundary conditions and solution behaviour. *AIAA*, 97-1850, 1997.

[21] J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN*, 33:1293–1316, 1999.

[22] K. Harriman, D. Gavaghan, and E. Süli. The importance of adjoint consistency in the approximation of linear functionals using the discontinuous Galerkin finite element method. Technical report, Oxford University Computing Laboratory, 2004.

[23] K. Harriman, P. Houston, B. Senior, and E. Süli. *hp*–Version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form. In *Recent Advances in Scientific Computing and Partial Differential Equations*, volume 330 of *Contemporary Mathematics*, pages 89–119. AMS, 2003.

[24] R. Hartmann. *Adaptive Finite Element Methods for the Compressible Euler Equations*. PhD thesis, University of Heidelberg, 2002.

[25] R. Hartmann. The role of the Jacobian in the adaptive Discontinuous Galerkin method for the compressible Euler equations. In G. Warnecke, editor, *Analysis and Numerics for Conservation Laws*, pages 301–316. Springer, 2005.

[26] R. Hartmann. Derivation of an adjoint consistent discontinuous Galerkin discretization of the compressible Euler equations. In G. Lube and G. Rapin, editors, *Proceedings of the BAIL 2006 conference*, 2006.

[27] R. Hartmann. Adjoint consistency analysis of discontinuous Galerkin discretizations. *SIAM J. Numer. Anal.*, 45(6):2671–2696, 2007.

[28] R. Hartmann. Error estimation and adjoint based refinement for an adjoint consistent DG discretization of the compressible Euler equations. *Int. J. Computing Science and Mathematics*, 1(2–4):207–220, 2007.

[29] R. Hartmann, J. Held, T. Leicht, and F. Prill. *PADGE, Parallel Adaptive Discontinuous Galerkin Environment, Technical reference*. DLR, Braunschweig, 2008. In preparation.

[30] R. Hartmann and P. Houston. Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations. *J. Comput. Phys.*, 183(2):508–532, 2002.

[31] R. Hartmann and P. Houston. Symmetric interior penalty DG methods for the compressible Navier–Stokes equations I: Method formulation. *Int. J. Num. Anal. Model.*, 3(1):1–20, 2006.

[32] R. Hartmann and P. Houston. Symmetric interior penalty DG methods for the compressible Navier–Stokes equations II: Goal–oriented a posteriori error estimation. *Int. J. Num. Anal. Model.*, 3(2):141–162, 2006.

[33] R. Hartmann and P. Houston. An optimal order interior penalty discontinuous Galerkin discretization of the compressible Navier–Stokes equations. *J. Comput. Phys.*, 227(22):9670–9685, 2008.

[34] P. Houston, J. Mackenzie, E. Süli, and G. Warnecke. A posteriori error analysis for numerical approximations of Friedrichs systems. *Numerische Mathematik*, 82:433–470, 1999.

[35] P. Houston and E. Süli. *hp*–adaptive discontinuous Galerkin finite element methods for first–order hyperbolic problems. *SIAM J. Sci. Comp.*, 23(4):1226–1252, 2002.

[36] J. Lu. *An a posteriori Error Control Framework for Adaptive Precision Optimization using Discontinuous Galerkin Finite Element Method*. PhD thesis, M.I.T., 2005.

[37] T. E. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.*, 28(1):133–140, 1991.

[38] S. Prudhomme, F. Pascal, J. Oden, and A. Romkes. Review of *a priori* error estimation for discontinuous Galerkin methods. TICAM Report 00-27, University of Texas, 2000.

[39] H. Schlichting and K. Gersten. *Boundary-Layer Theory*. Springer, 2003.

[40] C. Schwab. *p- and hp-Finite Element methods. Theory and Applications to Solid and Fluid Mechanics*. Oxford University Press, 1998.

[41] J. van der Vegt and H. van der Ven. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows, I. General formulation. *J. Comp. Phys.*, 182:546–585, 2002.

[42] K. G. van der Zee. An $H^1(P^h)$-coercive discontinuous Galerkin formulation for the Poisson problem: 1-d analysis. Master's thesis, TU Delft, 2004.