Discontinuous Galerkin methods for linear problems; an introduction

Emmanuil H. Georgoulis[1]

Abstract Discontinuous Galerkin (dG) methods for the numerical solution of partial differential equations (PDE) have enjoyed substantial development in recent years. Possible reasons for this are the flexibility in local approximation they offer, together with their good stability properties when approximating convection-dominated problems. Owing to their interpretation both as Galerkin projections onto suitable energy (native) spaces and, simultaneously, as high order versions of classical upwind finite volume schemes, they offer a range of attractive properties for the numerical solution of various classes of PDE problems where classical finite element methods underperform, or even fail. These notes aim to be a gentle introduction to the subject.

1 Introduction

Finite element methods (FEM) have been proven to be extremely useful in the numerical approximation of solutions to self-adjoint or "nearly" self-adjoint elliptic PDE problems and related indefinite PDE systems (e.g., Darcy's equations, Stokes' system, elasticity models), or to their parabolic counterparts.

Possible reasons for the success of FEM are their applicability in very general computational geometries of interest and the availability of tools for their rigorous error analysis. The error analysis is usually based on the variational interpretation of the FEM as a minimisation problem over finite-dimensional sets (or gradient flows of such, in the case of parabolic PDEs). The variational structure is inherited by the corresponding variational interpretation of the underlaying PDE problems, thereby facilitating the use of tools from PDE theory for the error analysis of the FEM.

Department of Mathematics, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom Emmanuil.Georgoulis@mcs.le.ac.uk

However, the use of (classical) FEM for the numerical solution of hyperbolic (or "nearly" hyperbolic) problems and other strongly non-self-adjoint PDE problems is, generally speaking, not satisfactory. These problems do not arise naturally in a variational setting. Indeed, the use of FEM for such problems has been mainly of academic interest in the 1970's and 1980's and for most of the 1990's. Instead, finite volume methods (FVM) have been predominantly used in industrial software packages for the numerical solution of hyperbolic (or "nearly" hyperbolic) systems, especially in the area of Computational Fluid Dynamics.

Nevertheless, in 1971 Reed and Hill [46] proposed a new class of FEM, namely the *discontinuous Galerkin finite element method* (dG method, for short) for the numerical solution of the nuclear transport PDE problem, which involves a linear first-order hyperbolic PDE. This methods was later analysed by LeSaint and Raviart [42] and by Johnson and Pitkäranta [39]. A significant volume of literature on dG methods for hyperbolic problems has since appeared in the literature; suggestively, we mention [17, 16, 14, 19, 24, 13, 8], the volume [15] and the references therein.

In the area of elliptic problems, Nitsche's seminal work on weak imposition of essential boundary conditions [44] for (classical) FEM, allowed for finite element solution spaces that do not satisfy the essential boundary conditions. This was followed up a few years later by Baker [6] who proposed the first modern discontinuous Galerkin method for elliptic problems, later followed by Wheeler [54], Arnold [3] and others. We also mention here the relevant finite element method with penalty of Babuška [5]. Since then, a plethora of DGFEMs have been proposed for a variety of PDE problems: we refer to [15, 34, 48] and the references therein for details.

DG methods exhibit attractive properties for the numerical approximation of problems of hyperbolic or nearly–hyperbolic type, compared to both classical FEM and FVM. Indeed, in contrast with classical FEM, but together with FVM, dG methods are, by construction, locally (or "nearly" locally) conservative with respect to the state variable; moreover, they exhibit enhanced stability properties in the vicinity of sharp gradients (e.g., boundary or interior layers) and/or discontinuities which are often present in the analytical solution of convection/transport dominated PDE problems. Additionally, dG methods offer advantages in the context of automatic local mesh and order adaptivity, such as increased flexibility in the mesh design (irregular grids are admissible) and the freedom to choose the elemental polynomial degrees without the need to enforce any conformity requirements. The implementation of genuinely (locally varying) high-order reconstruction techniques for FVM still remains a computationally difficult task, particularly on general unstructured hybrid grids.

Therefore, dG methods emerge as a very attractive class of arbitrary order methods for the numerical solution of various classes of PDE problems where classical FEM are not applicable and FVM produce typically low order approximations.

The rest of this work is structured as follows. In Section 2, we give a brief revision of the classical FEM for elliptic problems, along with its error analysis, and we discuss its limitations. An introduction to the philosophy of dG methods, along with some basic notation is given in Section 3. Section 4 deals with the construction of the popular interior-penalty dG method for linear elliptic problems, along with derivation of a priori and a posteriori error bounds. In Section 5, we present a dG method for first order linear hyperbolic problems, along with its a priori error analysis. In Section 7, two numerical experiments indicating the good performance of the dG method for PDE problems of mixed type, are given. Section 8 deals with the question of the efficient solution of the large linear systems arising from the discritesation using dG methods, while Section 9 contains some final concluding remarks.

1.1 Sobolev Spaces

We start by recalling the notion of a Sobolev space, based on the Lebesgue space $L^p(\omega)$, $p \in [1, \infty]$, for some open domain $\omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ (for more on Sobolev spaces see, e.g, [1]).

Definition 1. For $k \in \mathbb{N} \cup \{0\}$, we define the *Sobolev space* $W_p^k(\omega)$ over an open domain $\omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, by

$$W_p^k(\omega) := \{u \in L^p(\omega) : D^\alpha u \in L^p(\kappa) \text{ for } |\alpha| \leq k\}, \tag{1}$$

with $\alpha = (\alpha_1, \ldots, \alpha_d)$ being the standard multi-index notation. We also define the associated norm $\|\cdot\|_{W_p^k(\omega)}$ and seminorm $|\cdot|_{W_p^k(\omega)}$ by:

$$\|u\|_{W_p^k(\omega)} := \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p(\omega)}^p \right)^{\frac{1}{p}}, \quad |u|_{W_p^k(\omega)} := \left(\sum_{|\alpha| = k} \|D^\alpha u\|_{L^p(\omega)}^p \right)^{\frac{1}{p}},$$

for $p \in [1, \infty)$, and

$$\|u\|_{W_\infty^k(\omega)} := \max_{|\alpha| \leq s} \|D^\alpha u\|_{L^\infty(\omega)}, \quad |u|_{W_\infty^k(\omega)} := \max_{|\alpha| = k} \|D^\alpha u\|_{L^\infty(\omega)},$$

for $p = \infty$, respectively, for $k \in \mathbb{N} \cup \{0\}$. For $p = 2$, we shall use the abbreviated notation $W_2^k(\omega) \equiv H^s(\omega)$; equipped with the standard inner product, these spaces become Hilbert spaces. For $k = 0$, $p = 2$, we retrieve the standard $L^2(\omega)$ space, whose norm is abbreviated to $\|\cdot\|_\omega$, with associated inner product denoted by $\langle \cdot, \cdot \rangle_\omega$.

Negative and fractional order Sobolev spaces (i.e., where the *Sobolev index* $k \in \mathbb{R}$) are also defined by (standard) duality and function-space interpolation procedures, respectively, (for more on these techniques see, e.g., [1]). Also,

we shall make use of Sobolev spaces on manifolds, as we are interested in the regularity properties of functions on boundaries of domains. These are defined in a standard fashion via diffeomorphisms and partition of unity arguments (see, e.g., [47] for a nice exposition). Finally, we shall denote by $H^1_0(\omega)$ the space

$$H^1_0(\omega) := \{v \in H^1(\omega) : v = 0 \text{ on } \partial\omega\}.$$

2 The Finite Element Method

We illustrate the classical Finite Element Method for linear elliptic problems by considering the Poisson problem with homogeneous Dirichlet boundary conditions over an open bounded polygonal domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$:

$$\begin{aligned} -\Delta u &= f, && \text{in } \Omega \\ u &= 0, && \text{on } \partial\Omega, \end{aligned} \tag{2}$$

where $f : \Omega \to \mathbb{R}$ is a known function.

To simplify the notation, we use the following abbreviations for the L^2-norm and the corresponding inner product when defined over the computational domain Ω: $\|\cdot\|_\Omega \equiv \|\cdot\|$ and $\langle\cdot,\cdot\rangle_\Omega \equiv \langle\cdot,\cdot\rangle$, respectively.

The first step in defining a finite element method is to rewrite the problem (2) in the so-called *weak form* or *variational form*. Let $V = H^1_0(\Omega)$ be the *solution space* and consider a function $v \in V$. Upon multiplication of the PDE in (2) by v (usually, referred to as the *test function*) and integration over the domain Ω, we obtain

$$-\int_\Omega \Delta u\, v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x.$$

Using the divergence theorem to the integral on the left-hand side, and the fact that $v = 0$ on $\partial\Omega$ for all $v \in V$, we arrive to

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x,$$

for all $v \in \mathcal{H}$. Hence, the Poisson problem with homogeneous Dirichlet boundary conditions can be transformed to the following problem in *weak form*:

$$\text{Find } u \in V \text{ such that } a(u, v) = \langle f, v\rangle, \quad \text{for all } v \in V, \tag{3}$$

with the *bilinear form* $a(\cdot,\cdot)$ defined by

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x. \tag{4}$$

The second step is to consider an approximation to the problem (3). To this end, we restrict the (infinite-dimensional) space V of eligible solutions to a finite-dimensional subspace $V_h \subset V$ and we consider the approximation

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = \langle f, v_h \rangle, \quad \text{for all } v_h \in V_h. \tag{5}$$

This procedure is usually referred to as the *Galerkin projection* (also known as the *Ritz projection* when $a(\cdot,\cdot)$ is a symmetric bilinear form, as is the case here).

Setting $v = v_h \in V_h$ in (3) and subtracting (5) from the resulting equation, we arrive to

$$a(u - u_h, v_h) = 0 \quad \text{for all} \quad v_h \in V_h. \tag{6}$$

The identity (6) is usually referred to as the *Galerkin orthogonality property*. Noting that, in this case, the bilinear form $a(\cdot,\cdot)$ satisfies the properties of an inner product on $H_0^1(\Omega)$, the Galerkin orthogonality property states that u_h is the best approximation of u in V_h, with respect to the inner product defined by the bilinear form $a(\cdot,\cdot)$.

We remark that, although $a(\cdot,\cdot)$ may not satisfy the properties of an inner product if it is not symmetric (e.g., as a result of writing a non-self-adjoint PDE problem in weak form), Galerkin orthogonality still holds by construction, as long as long as the approximation is *conforming*, that is as long as $V_h \subset V$.

From the analysis point of view, there is great flexibility in the decision of an appropriate approximation space. The conformity of the approximation space requires $V_h \subset V$. To investigate what other assumptions on V_h are sufficient for (5) to deliver a useful approximation, we consider a family of basis functions ψ_i, with $i = 1, 2, \ldots, N$, for some $N \in \mathbb{N}$, spanning V_h, viz., $V_h = \text{span}\{\psi_i : i = 1, 2, \ldots, N\}$. Due to linearity of the bilinear form, the approximation problem (5) is equivalent to the problem:

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, \psi_i) = \langle f, \psi_i \rangle, \quad \text{for all } i = 1, 2, \ldots, N. \tag{7}$$

Since $u_h \in V_h$, there exist $U_j \in \mathbb{R}$, $j = 1, 2, \ldots, N$, so that $u_h = \sum_{j=1}^{N} U_j \psi_j$, which upon insertion into (7), leads to the linear system

$$A\mathbf{U} = \mathbf{F}, \tag{8}$$

with $A = [A_{ij}]_{i,j=1}^{N}$, $\mathbf{U} = (U_1, \ldots, U_N)^T$ and $\mathbf{F} = (F_1, \ldots, F_N)^T$, where

$$A_{ij} = \int_\Omega \nabla\psi_j \cdot \nabla\psi_i \, \mathrm{d}x, \quad \text{and} \quad F_i = \int_\Omega f\psi_i \, \mathrm{d}x. \tag{9}$$

Notice that the matrix A is symmetric. For the approximation u_h to be well defined, the linear system (8) should have a unique solution. It is, therefore, reasonable to consider a space V_h so that the matrix A is positive definite.

Further restrictions on the choices of "good" subspaces V_h become evident when considering the practical implementation of the Galerkin procedure. In particular, the supports of the basis functions ψ_i should be a covering the computational domain Ω, while being simultaneously relatively simple in shape so that the entries A_{ij} can be computed in an efficient fashion. Also, given that the linear system (8) can be quite large, it would be an advantage if A is a sparse matrix, to reduce the computational cost of solving (8).

The (classical) finite element method (FEM) is defined by the Galerkin procedure described above through a particular choice of the subspace V_h, which we now describe.

We begin by splitting the domain Ω into a covering $\mathcal{T}$, which will be referred to as the *triangulation* or the *mesh*, consisting of open triangles if $d = 2$ or open tetrahedra if $d = 3$, which we shall refer to as the *elements*, with the following properties:

(a) $\Omega = \cup_{T\in\mathcal{T}}\bar{T}$, with $\bar{\cdot}$ denoting the closure of a set in $\mathbb{R}^d$;
(b) for $T, S \in \mathcal{T}$, we only have the possibilities: either $T = S$, or $\bar{T} \cap \bar{S}$ is a common (whole) $d-k$-dimensional face, with $1 \le k \le d$ (i.e., face, edge or vertex, respectively).

In Figure 1, we illustrate a mesh for a domain $\Omega \subset \mathbb{R}^2$.

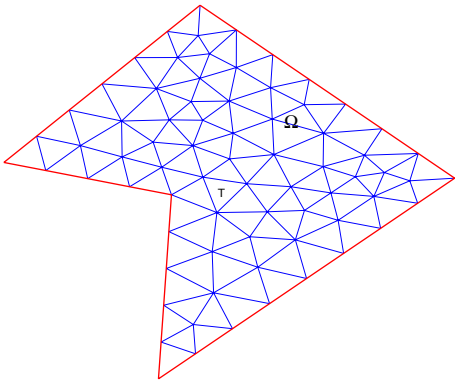


Fig. 1 A mesh in two dimensions

The finite element space V_h^p of degree p is then defined as the space of element-wise d-variate polynomials of degree at most p that are continuous across the inter-element boundaries, viz.,

$$V_h^p := \{w_h \in C(\Omega) : w_h|_T \in \mathcal{P}_p(T),\ T \in \mathcal{T} \ \text{ and } w_h|_{\partial\Omega} = 0\}$$

with $\mathcal{P}_p(T)$ denoting the space of d-variate polynomials of degree at most p. It is evident that $V_h^p \subset H_0^1(\Omega) = V$. The Galerkin procedure with the particular choice of $V_h = V_h^p$ is called the (classical) *finite element method*.

We observe that this choice of finite dimensional subspace is in line with the practical requirements that a "good" subspace should admit. Indeed, the element-wise polynomial functions over simple triangular or tetrahedral domains enable efficient quadrature calculations for the entries of the matrix A. Moreover, choosing carefully a basis for V_h^p (for instance, the *Lagrange elements*, see, e.g., [12, 9] for details) the resulting linear system becomes sparse. This is because the Lagrange elements have very small support consisting of an element together with only *some* of its immediate neighbours sharing a face or an edge or a vertex. Moreover, as we shall see next, the choice of V_h^p yields a positive definite matrix A, thereby it is uniquely solvable.

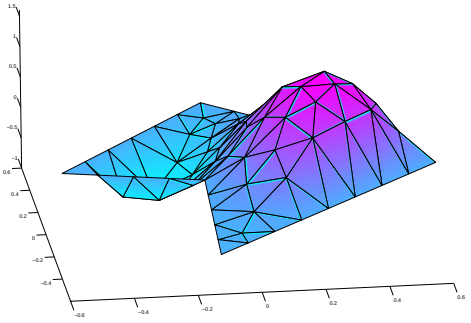


Fig. 2 Example 1. Approximation using the Finite Element Method

Example 1. We use the finite element method to approximate the solution to the Poisson problem with homogeneous Dirichlet boundary conditions:

$$-\Delta u = 100\sin(\pi x), \quad \text{in } \Omega, \quad u = 0, \quad \text{on } \partial\Omega,$$

where Ω is given by the domain in Figure 1, along with the mesh used in the approximation. The finite element approximation for using element-wise linear basis (i.e., $V_h = V_h^1$) is shown in Figure 2.

2.1 Error analysis of the FEM

Let $\|\nabla w\| := \left(\int_\Omega |\nabla w|^2 \,\mathrm{d}x\right)^{1/2}$ for a weakly differentiable scalar function w, and note that $\|\nabla\cdot\|$ is a norm on $V = H_0^1(\Omega)$. It is evident that that

$a(w,w) = \|\nabla w\|^2$ for $w \in V$, i.e., the bilinear form is *coercive in* V. This immediately implies that $a(\cdot,\cdot)$ is also coercive in the closed subspace V_h^p of V. Hence $\|\nabla\cdot\|$ is a norm on V_h^p also and, thus, the corresponding matrix A is positive definite, yielding unique solvability of (8) for $V_h = V_h^p$. The norm for which the bilinear form is coercive for is usually referred to as the *energy norm*.

Coercivity, Galerkin orthogonality (6) and the Cauchy-Schwarz inequality, respectively, imply

$$\begin{aligned}\|\nabla(u-u_h)\|^2 = a(u-u_h,u-u_h) &= a(u-u_h,u-v_h)\\ &\leq \|\nabla(u-u_h)\|\,\|\nabla(u-v_h)\|,\end{aligned}$$

for any $v_h \in V_h^p$, which yields

$$\|\nabla(u-u_h)\| \leq \inf_{v_h\in V_h^p}\|\nabla(u-v_h)\|, \tag{10}$$

that is, the finite element method produces the best approximation of V_h^p for the exact solution $u \in V$ with respect to the energy norm. This result is known in the literature as *Cea's Lemma*. The error analysis of the FEM can be now completed using (10) in conjunction with Jackson-type inequalities (such as the Bramble-Hilbert Lemma, see, e.g., [9]) of the form

$$\inf_{v_h\in V_h^p}\|\nabla(u-v_h)\| \leq Ch^{\min\{p,r-1\}}|u|_{H^r(\Omega)}, \tag{11}$$

for $u \in H^r(\Omega)\cap H_0^1(\Omega)$, where $h = \max_{T\in\mathcal{T}}\mathrm{diam}(T)$ and the constant C is independent of h and of u. Combining now (10) with (11) results to standard a-priori error bounds for the FEM:

$$\|\nabla(u-u_h)\| \leq Ch^{\min\{p,r-1\}}|u|_{H^r(\Omega)}, \tag{12}$$

i.e., as $h \to 0$, the error decreases at an algebraic rate which depends on the local polynomial degree used and the regularity of the solution in the domain Ω.

3 Discontinuous Galerkin methods

The restriction $V_h \subset V$ essentially dictates that the underlying space contains only functions of particular smoothness (e.g., when $V = H_0^1(\Omega)$, we choose $V_h \subset \{v \in C(\bar{\Omega}) : v|_{\partial\Omega} = 0\} \subset H_0^1(\Omega)$). Although the FEM is, generally speaking, well suited for PDE problems related to a variational/minimisation setting, it is well known that this restriction can have a degree of severity in the applicability of FEM for a larger class of PDE problems (e.g. first order hyperbolic PDE problems). There has been a substantial amount of work

in the literature on so-called *non-conforming* FEM, whereby $V_h \nsubseteq V$ since the 1970s. The discontinuous Galerkin (dG) methods described below will admit finite element spaces with "severe" non-conformity, i.e., element-wise discontinuous polynomial spaces, viz.,

$$S_h^p := \{w_h \in L^2(\Omega) : w_h|_T \in \mathcal{P}_p(T),\ T \in \mathcal{T}\}.$$

Let us introduce some notation first. We denote by $\mathcal{T}$ a subdivision of Ω into (triangular or quadrilateral if $d = 2$ and tetrahedral or hexahedral if $d = 3$) elements T. We define $\Gamma := \cup_{T\in\mathcal{T}}\partial T$ the *skeleton* of the mesh (i.e., the union of all $(d-1)$-dimensional element faces) and let $\Gamma_{\rm int} := \Gamma\backslash\partial\Omega$. Let also $\Gamma_{\rm int} := \Gamma\backslash\partial\Omega$, so that $\Gamma = \partial\Omega \cup \Gamma_{\rm int}$.

Let T^+, T^- be two (generic) elements sharing a face $e := T^+ \cap T^- \subset \Gamma_{\rm int}$ with respective outward normal unit vectors $\mathbf{n}^+$ and $\mathbf{n}^-$ on e. For $q : \Omega \to \mathbb{R}$ and $\boldsymbol{\phi} : \Omega \to \mathbb{R}^d$, let $q^\pm := q|_{e\cap\partial T^\pm}$ and $\boldsymbol{\phi}^\pm := \boldsymbol{\phi}|_{e\cap\partial T^\pm}$, and set

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

if $e \subset \partial T \cap \partial\Omega$, we set $\{\boldsymbol{\phi}\}|_e := \boldsymbol{\phi}^+$ and $[\![q]\!]|_e := q^+\mathbf{n}^+$. Finally, we introduce the *meshsize* $h : \Omega \to \mathbb{R}$, defined by $h(x) = \mathrm{diam}(T)$, if $x \in T\backslash\partial T$ and $h(x) = \{h\}$, if $x \in \Gamma$. The subscript e in these definitions will be suppressed when no confusion is likely to occur.

4 Discontinuous Galerkin methods for elliptic problems

Now we are ready to derive the weak form for the Poisson problem (2), which will lead to the discontinuous Galerkin (dG) finite element method.

Since the dG method will be non-conforming, we should work on an extended variational framework, making use of the space $\mathcal{S} := H_0^1(\Omega) + S_h^p$. Assuming for the moment that u is smooth enough, we multiply the equation by a test function $v \in \mathcal{S}$, we integrate over Ω and we split the integrals:

$$-\sum_{T\in\mathcal{T}}\int_T \Delta u v\,\mathrm{d}x = \sum_{T\in\mathcal{T}}\int_T f v\,\mathrm{d}x.$$

Using the divergence theorem on every elemental integral (as v is now element-wise discontinuous), using the anti-clockwise orientation, we have

$$\sum_{T\in\mathcal{T}}\int_T \nabla u\cdot\nabla v\,\mathrm{d}x - \sum_{T\in\mathcal{T}}\int_{\partial T}(\nabla u\cdot n)v\,\mathrm{d}s = \int_\Omega f v\,\mathrm{d}x = \langle f, v\rangle,$$

where n denotes the outward normal to each element edge.

The second term on the left-hand side contains the integrals over the element faces. Thus, when the face is common to two adjacent element, we have two integrals over every interior face.

Now, from standard elliptic regularity estimates (see, e.g., Corollary 8.36 in Gilbarg & Trudinger [32]), we have that $u \in C^1(\Omega')$ for all $\Omega' \subset\subset \Omega$ and, hence, ∇u is continuous across the interior element faces. Thus we can substitute ∇u by $\{\nabla u\}$ for all faces on the skeleton Γ, noting that this is just the definition of $\{\nabla u\}$ on the boundary $\partial\Omega$. Taking into account the orientation convention we have adopted, we can see that this sum can be rewritten as follows:

$$\sum_{T\in\mathcal{T}} \int_T \nabla u \cdot \nabla v \,\mathrm{d}x - \int_\Gamma \{\nabla u\} \cdot [\![v]\!] \,\mathrm{d}s = \langle f, v\rangle. \tag{13}$$

One may now be tempted to define a bilinear form and a linear form from the left- and right-hand sides of (13), respectively and attempt to solve the resulting variational problem. Such an endeavour would be deemed with failure for the reason that the left-hand side does not give rise to a positive-definite operator, (not even a conditionally positive definite operator,) over S_h^p. In other words, there is no coercivity property for such a bilinear form in any relevant norm. There is also the somewhat philosophically discomforting issue that a symmetric variational problem (such as the Poisson problem in variational form) is approximated using a Galerkin procedure based on a *non-symmetric* bilinear form (such as the one stemming from the left-hand side of (13)). This issue may also have practical implications as solving non-symmetric linear systems is usually a far more computationally demanding procedure than solving a symmetric linear system.

To rectify the lack of positivity, we work as follows. We begin by noting that $[\![u]\!] = 0$ on Γ, due to elliptic regularity on Γ_{int} and due to the boundary conditions on $\partial\Omega$. Therefore, we have

$$\int_\Gamma \sigma [\![u]\!] \cdot [\![v]\!] \,\mathrm{d}s = 0, \tag{14}$$

for any positive function $\sigma : \Gamma \to \mathbb{R}$. Note that this term is symmetric with respect to the two arguments u and v and can be arbitrarily large (upon choosing σ arbitrarily large positive function) upon replacing u with a function $v \in \mathcal{S}$. Adding (13) and (14) up, we arrive to

$$\sum_{T\in\mathcal{T}} \int_T \nabla u \cdot \nabla v \,\mathrm{d}x - \int_\Gamma (\{\nabla u\} \cdot [\![v]\!] - \sigma [\![u]\!] \cdot [\![v]\!]) \,\mathrm{d}s = \langle f, v\rangle. \tag{15}$$

Since the term of the left-hand side of (15), stemming from (14) gives rise to a positive-definite term in the bilinear form, which implies that there is a range of (large enough) σ that will render the resulting bilinear form coercive (at least over a finite dimensional subspace of $\mathcal{S}$) to give rise to a positive definite

finite element matrix. The choice of the *discontinuity-penalisation parameter* σ, as it is often called in the literature, will arise from the error analysis of the method.

We note that the left-hand side of (15) is still non-symmetric with respect to the arguments u and v. To rectify this, we observe that also

$$\int_\Gamma \{\nabla v\} \cdot [\![u]\!] \, \mathrm{d}s = 0, \tag{16}$$

assuming that v is smooth enough, which can be subtracted from (15), resulting to

$$\sum_{T\in\mathcal{T}} \int_T \nabla u \cdot \nabla v \, \mathrm{d}x - \int_\Gamma (\{\nabla u\} \cdot [\![v]\!] + \{\nabla v\} \cdot [\![u]\!] - \sigma [\![u]\!] \cdot [\![v]\!]) \, \mathrm{d}s = \langle f, v\rangle, \tag{17}$$

whose left-hand side is now symmetric with respect to the arguments u and v.

The above suggest the following numerical method:

$$\text{Find } u_h \in S_h^p \text{ such that } B(u_h, v_h) = \langle f, v_h\rangle, \quad \text{for all } v_h \in S_h^p, \tag{18}$$

where the bilinear form $B : S_h^p \times S_h^p \to \mathbb{R}$ is defined by

$$B(w, v) := \sum_{T\in\mathcal{T}} \int_T \nabla w \cdot \nabla v \, \mathrm{d}x - \int_\Gamma (\{\nabla w\} \cdot [\![v]\!] + \{\nabla v\} \cdot [\![w]\!] - \sigma [\![w]\!] \cdot [\![v]\!]) \, \mathrm{d}s. \tag{19}$$

This is the so-called *(symmetric) interior penalty discontinuous Galerkin method* for the Poisson problem.

Historically interior penalty methods were the first to appear in the literature [6, 3], but some of the ideas can be traced back to the treatment of non-homogeneous Dirichlet boundary conditions by penalties due to Nitsche [44]. Interior penalty dG methods are, perhaps, the most popular dG methods in the literature and in applications, so they will be our main focus in the present notes. In recent years, a number of other discontinuous Galerkin methods for second order elliptic problems have appeared in the literature; we refer to [4] and the references therein for a discussion on the unifying characteristics of these methods as well as on the particular advantages and disadvantages of each dG method.

4.1 Error analysis of the dG method

In the above derivation of the interior penalty dG method, we have been intentionally relaxed about the smoothness requirements of the arguments of

the bilinear forms. The bilinear form (19) is well defined if the arguments w and v belong to the finite element space S_h^p.

However, it is well known from the theory of Sobolev spaces that functions in $L^2(\Omega)$ do not have a well-defined *trace* on $\partial\Omega$, that is, they are not uniquely defined up to boundary values. Therefore, $\{\nabla w\}$ and $\{\nabla v\}$ are not well defined on $\Gamma_{\rm int}$ in (19) when $w, v \in \mathcal{S}(= H_0^1(\Omega) + S_h^p)$. For the error analysis it is desired that the bilinear form can be applied to the exact solution u. Fortunately, for (standard) a-priori error analysis the exact solution u is assumed to admit at least H^2-regularity which, implies that all the terms in (19) can be taken to be well defined, so this issue does not pose any crucial restriction. For the derivation of a-posteriori error bounds describe below, however, assuming the minimum possible regularity for u is essential for their applicability to the most general setting possible.

It is possible to overcome this hinderance by extending the bilinear form (19) from $S_h^p \times S_h^p$ to $\mathcal{S} \times \mathcal{S}$ in a non-trivial fashion. More specifically, we define

$$\tilde{B}(w,v) := \sum_{T\in\mathcal{T}} \int_T \nabla w \cdot \nabla v \,\mathrm{d}x - \int_\Gamma (\{\Pi\nabla w\}\cdot[\![v]\!] + \{\Pi\nabla v\}\cdot[\![w]\!] - \sigma[\![w]\!]\cdot[\![v]\!])\,\mathrm{d}s, \tag{20}$$

where $\Pi : L^2(\Omega) \to S_h^p$ is the orthogonal L^2-projection with respect to the $\langle\cdot,\cdot,\rangle$- inner product. This way, the face integrals involving the terms $\{\Pi\nabla w\}$ and $\{\Pi\nabla v\}$ are well defined, as these terms are now traces of element-wise polynomial functions from the finite element space. Moreover, it is evident that

$$\tilde{B}(w,v) = B(w,v), \quad \text{if} \quad w, v \in S_h^p,$$

i.e., $\tilde{B}(\cdot,\cdot)$ is an extension of $B(\cdot,\cdot)$ to $\mathcal{S}\times\mathcal{S}$. However, $\tilde{B}(\cdot,\cdot)$ is *inconsistent* with respect to the Poisson problem, that is, it is not a weak form of (2). Indeed, suppose (2) admits a classical solution denoted by u_{cl}. Upon inserting u_{cl} into $\tilde{B}(\cdot,\cdot)$, and integrating by parts, we deduce

$$-\sum_{T\in\mathcal{T}} \int_T \Delta u_{cl} v \,\mathrm{d}x + \int_\Gamma \{\nabla u_{cl}\}\cdot[\![v]\!]\,\mathrm{d}s - \int_\Gamma \{\Pi\nabla u_{cl}\}\cdot[\![v]\!]\,\mathrm{d}s = \langle f, v\rangle,$$

for all $v \in \mathcal{S}$, noting that $[\![u_{cl}]\!] = 0$ on Γ, which implies

$$\int_\Omega (f + \Delta u_{cl}) v \,\mathrm{d}x = \int_\Gamma \{\nabla u_{cl} - \Pi\nabla u_{cl}\}\cdot[\![v]\!]\,\mathrm{d}s,$$

for all $v \in \mathcal{S}$; the right-hand side being a representation of the inconsistency. (If $\tilde{B}(\cdot,\cdot)$ was consistent, the right-hand side should have been equal to zero.) Nevertheless, as we shall see below, the inconsistency is of the same order as the convergence rate of the dG method and it is, therefore, a useful tool in the error analysis.

For the error analysis, we consider the following (natural) quantity:

$$|||w||| := \Big(\sum_{T\in\mathcal{T}} \int_T |\nabla w|^2 \,\mathrm{d}x + \int_\Gamma \sigma |[\![w]\!]|^2 \,\mathrm{d}s \Big)^{1/2},$$

for all $w \in \mathcal{S}$ and for $\sigma > 0$. Note that $|||\cdot|||$ is a norm in $\mathcal{S}$.

We begin the error analysis by assessing the coercivity and the continuity of the bilinear form.

Lemma 1. *Let constant $c > 0$ such that $\mathrm{diam}(T)/\rho_T \le c$ for all $T \in \mathcal{T}$, where ρ_T is the radius of the incircle of T. Let also $\sigma := C_\sigma p^2/h$ with $C_\sigma > 0$ large enough and independent of p, h and of $w, v \in \mathcal{S}$. Then, we have*

$$\frac{1}{2}|||w|||^2 \le \tilde{B}(w,w), \quad \textit{for all} \quad w \in \mathcal{S}, \tag{21}$$

and

$$\tilde{B}(w,v) \le |||w||| \, |||v|||, \quad \textit{for all} \quad w, v \in \mathcal{S}. \tag{22}$$

Proof. We prove (21). For $w \in \mathcal{S}$, we have:

$$\tilde{B}(w,w) = |||w|||^2 - 2\int_\Gamma \{\!\{\Pi\nabla w\}\!\} \cdot [\![w]\!] \,\mathrm{d}s.$$

Now the last term on the right-hand side can be bounded from above as follows:

$$\begin{aligned} 2\int_\Gamma \{\!\{\Pi\nabla w\}\!\} \cdot [\![w]\!] \,\mathrm{d}s =& 2\int_\Gamma \Big(\frac{\sigma}{2}\Big)^{-1/2} \{\!\{\Pi\nabla w\}\!\} \cdot \Big(\frac{\sigma}{2}\Big)^{1/2} [\![w]\!] \,\mathrm{d}s \\ \le& 2\int_\Gamma \sigma^{-1} |\{\!\{\Pi\nabla w\}\!\}|^2 \,\mathrm{d}s + \frac{1}{2}\int_\Gamma \sigma [\![w]\!]^2 \,\mathrm{d}s, \end{aligned} \tag{23}$$

using in the last step an inequality of the form $2\alpha\beta \le \alpha^2 + \beta^2$.

To bound from above the first term on the right-hand side of the last inequality, we make use of the *inverse inequality*:

$$\|v\|_{\partial T}^2 \le C_{\mathrm{inv}} p^2 |\partial T|/|T| \|v\|_T^2, \tag{24}$$

for all $v \in \mathcal{P}_p(T)$, with $C_{\mathrm{inv}} > 0$ independent of p, h and v, with $|\partial T|$ and $|T|$ denoting the $(d-1)$- and d-dimensional volumes of ∂T and T, respectively. (We refer to Theorem 4.76 in [51] for a proof, when T is the reference element; the proof for a general element follows by a standard scaling argument.) To this end, we have

$$2\int_\Gamma \sigma^{-1} |\{\!\{\Pi\nabla w\}\!\}|^2 \,\mathrm{d}s \le \sum_{T\in\mathcal{T}} \int_{\partial T} \sigma^{-1} |\Pi\nabla w|^2 \,\mathrm{d}s, \tag{25}$$

using an inequality of the form $(\alpha+\beta)^2 \le 2\alpha^2 + 2\beta^2$. The right-hand side of (25) can be further bounded using (24), noting that $(\Pi\nabla w)|_T \in \mathcal{P}_p(T)$ for all $T \in \mathcal{T}$, giving

$$\sum_{T\in\mathcal{T}}\int_{\partial T}\sigma^{-1}|\Pi\nabla w|^2\,\mathrm{d}s\leq\sum_{T\in\mathcal{T}}\frac{C_{\mathrm{inv}}p^2|T|}{\sigma|\partial T|}\int_T|\Pi\nabla w|^2\,\mathrm{d}x. \tag{26}$$

The orthogonal L^2-projection operator is stable in the L^2-norm, with $\|\Pi v\|_T\leq\|v\|_T$ for all $v\in L^2(T)$ which, in conjunction with (25) and (26), gives

$$2\int_\Gamma\sigma^{-1}|\{\Pi\nabla w\}|^2\,\mathrm{d}s\leq\sum_{T\in\mathcal{T}}\frac{C_{\mathrm{inv}}p^2}{\sigma\rho_T}\int_T|\nabla w|^2\,\mathrm{d}x, \tag{27}$$

noting that $\rho_T\leq|T|/|\partial T|$. Choosing $C_\sigma\geq 2c^2C_{\mathrm{inv}}$, implies $C_{\mathrm{inv}}p^2h/(\sigma\rho_T)\leq 1/2$, as $h\leq\mathrm{diam}(T)\leq c\rho_T$ and $\rho_T\leq c\rho_{T'}$ for all elements T' sharing a face with T (the latter is due to assumption $\mathrm{diam}(T)/\rho_T\leq c$ for all $T\in\mathcal{T}$). Hence, combining (27) with (23) already implies (21).

The proof (22) uses the Cauchy-Schwarz inequality along with the same tools as above and it is omitted for brevity. □

Remark 1. It can be seen from the proof that the coercivity and continuity constants in the previous result can be different, depending the choice of the penalty parameter σ.

4.1.1 A priori error bounds

Since the bilinear form is now inconsistent, Galerkin orthogonality does not hold for dG (cf. (6) which, in turn complicates slightly the a priori error analysis. To this end, let $u_h\in S_h^p$ be the dG approximation to the exact solution u, arising from solving (18) and consider a $v_h\in S_h^p$. Then, we have

$$\begin{aligned}\frac{1}{2}|||v_h-u_h|||^2\leq&\tilde{B}(v_h-u_h,v_h-u_h)\\=&\tilde{B}(v_h-u,v_h-u_h)+\tilde{B}(u-u_h,v_h-u_h)\\=&\tilde{B}(v_h-u,v_h-u_h)+\tilde{B}(u,v_h-u_h)-\langle f,v_h-u_h\rangle\end{aligned}$$

using coercivity (21), the linearity of the bilinear form and the definition of the dG method (18), respectively. Using the continuity (22) of the bilinear form, and diving by $|||v_h-u_h|||$, we arrive to

$$|||v_h-u_h|||\leq 2|||u-v_h|||+\sup_{w_h\in S_h^p}\frac{|\tilde{B}(u,w_h)-\langle f,w_h\rangle|}{|||w_h|||}, \tag{28}$$

for all $v_h\in S_h^p$. Hence, we can conclude

$$|||v_h-u_h|||\leq 2\inf_{v_h\in S_h^p}|||u-v_h|||+\sup_{w_h\in S_h^p}\frac{|\tilde{B}(u,w_h)-\langle f,w_h\rangle|}{|||w_h|||}, \tag{29}$$

or

$$|||u-u_h||| \leq 3 \inf_{v_h \in S_h^p} |||u-v_h||| + \sup_{w_h \in S_h^p} \frac{|\tilde{B}(u,w_h) - \langle f, w_h\rangle|}{|||w_h|||}, \tag{30}$$

using the triangle inequality. This result is a generalisation of Ceá's Lemma presented above for the case where the Galerkin orthogonality is not satisfied exactly; it is known in the literature as *Strang's Second Lemma* (see, e.g., [53, 12]). Indeed, if the bilinear form is consistent, then the last term on the right-hand side of (30) vanishes.

For the first term on the right-hand side of (30), we can standard use best approximation results (such as the Bramble-Hilbert Lemma, see, e.g., [9]) of the form

$$\inf_{v_h \in V_h^p} |||\nabla(u-v_h)||| \leq C h^{\min\{p,r-1\}} |u|_{H^r(\Omega)}, \tag{31}$$

for $u \in H^r(\Omega) \cap H_0^1(\Omega)$, noting that the parameter σ in the dG-norm $|||\cdot|||$ scales like h^{-1} and using the standard trace estimate

$$\|w\|_{\partial T}^2 \leq C\big(\operatorname{diam}(T)^{-1}\|w\|_T^2 + \operatorname{diam}(T)\|\nabla w\|_T^2\big), \tag{32}$$

on the term of $|||\cdot|||$ involving σ, $T \in \mathcal{T}$.

To bound the second term on the right-hand side (30), we begin by observing that

$$\tilde{B}(u,w_h) - \langle f, w_h\rangle = -\int_\Gamma \{\!\{\nabla u - \Pi\nabla u\}\!\} \cdot [\![w_h]\!]\,\mathrm{d}s,$$

which implies

$$\sup_{w_h \in S_h^p} \frac{|\tilde{B}(u,w_h) - \langle f, w_h\rangle|}{|||w_h|||} \leq \Big(\int_\Gamma \sigma^{-1} |\{\!\{\nabla u - \Pi\nabla u\}\!\}|^2\,\mathrm{d}s\Big)^{1/2}. \tag{33}$$

Letting $\eta := |\nabla u - \Pi\nabla u|$ for brevity and, working as before, the square of right-hand side of (33) can be bounded by

$$\int_\Gamma \sigma^{-1}|\{\!\{\eta\}\!\}|^2\,\mathrm{d}s \leq \frac{1}{2}\sum_{T\in\mathcal{T}}\int_{\partial T}\sigma^{-1}\eta^2\,\mathrm{d}s \leq C\sum_{T\in\mathcal{T}}\big(\|\eta\|_T^2 + \|h\nabla\eta\|_T^2\big), \tag{34}$$

using the trace estimate (32) and the definition of σ. Now, standard best approximation results for the L^2-projection error (see, e.g., [51]) yield

$$\big(\|\eta\|_T^2 + \|h\nabla\eta\|_T^2\big)^{1/2} \leq C h^{\min\{p,r-1\}} |u|_{H^r(\Omega)}, \tag{35}$$

for $u \in H^r(\Omega) \cap H_0^1(\Omega)$. Combining (35), (34), (33), and using the resulting bound together with (31) into (30), we arrive to the *a priori error bound*

$$|||u-u_h||| \leq C h^{\min\{p,r-1\}} |u|_{H^r(\Omega)}, \tag{36}$$

for $u \in H^r(\Omega) \cap H_0^1(\Omega)$, for some $C > 0$ independent of u and of h.

We refer to [6, 3, 15, 50, 36, 4, 31, 28] and the references therein for discussion on a priori error analysis of interior penalty-type dG methods for elliptic problems.

4.1.2 A posteriori error bounds

The above a priori bounds are relevant when we are interested in assessing the asymptotic error behaviour of the dG method. However, since they involve the unknown solution to the boundary-value problem u, they are not of relevance in practice. The derivation of computable bounds, usually referred to the finite element literature as *a posteriori estimates* is therefore, relevant to assess the accuracy of practical computations. Moreover, such bounds can be used to drive automatic mesh-adaptation procedures, usually termed *adaptive algorithms*. A-posteriori bounds for dG methods for elliptic problems have been considered in [7, 40, 35, 11, 22, 23, 41]. Here, we shall only illustrate the main ideas in a simple setting.

We begin by decomposing the discontinuous finite element space S_h^p into the conforming finite element space $V_h^p \subset S_h^p$ and a *non-conforming* remainder part V_d, so as $S_h^p := V_h^p \oplus V_d$, where the uniqueness-of-the-decomposition property in the direct sum can be realised, once an inner product in S_h^p is selected. The approximation of functions in S_h^p by functions in the conforming finite element space V_h^p will play an important role in our derivation of the a posteriori bounds. This can be quantified by the following result, whose proof can be found in [40].

Lemma 2. *For a mesh $\mathcal{T}$, let constant $c > 0$ such that $\mathrm{diam}(T)/\rho_T \leq c$ for all $T \in \mathcal{T}$, where ρ_T is the radius of the incircle of T. Then, for any function $v \in S_h^p$, there exists a function $v_c \in V_h^p$ such that*

$$\|\nabla(v - v_c)\| \leq C_1 \|\sqrt{\sigma}[\![v]\!]\|_\Gamma, \tag{37}$$

where the constant $C_1 > 0$ depends on c, p, but is independent of h, v, and v_c.

Using this lemma it is possible to derive an a posteriori bound for the dG method for the Poisson problem. This is the content of the following result.

Theorem 1. *Let u be the solution (3) and u_h its approximation by the dG method (18). Then, the following bound holds*

$$|||u - u_h||| \leq C\mathcal{E}(u_h, f, \mathcal{T}), \tag{38}$$

with

$$\mathcal{E}(u_h, f, \mathcal{T}) := \left(\|h(f + \Delta u)\|^2 + \|\sqrt{h}[\![\nabla u_h]\!]\|^2_{\Gamma_{\mathrm{int}}} + \|\sqrt{\sigma}[\![u_h]\!]\|^2_\Gamma\right)^{1/2}, \tag{39}$$

where $C > 0$ is independent of u_h, u, h and $\mathcal{T}$.

Proof. Let $u_h^c \in S_c$ the conforming part of u_h as in Lemma 2 and define

$$e := u - u_h = e_c + e_d, \quad \text{where} \quad e_c := u - u_h^c, \quad \text{and} \quad e_d := u_h^c - u_h, \tag{40}$$

yielding $e_c \in H_0^1(\Omega)$. Thus, we have $B(u, e_c) = \langle f, e_c \rangle$. Let $\Pi_0 : L^2(\Omega) \to \mathbb{R}$ denote the orthogonal L^2-projection onto the elementwise constant functions; then $\Pi_0 e_c \in S_h^p$ and we define $\eta := e_c - \Pi_0 e_c$. We then have, respectively,

$$\begin{aligned} B(e, e_c) =& B(u, e_c) - B(u_h, e_c) \\ =& \langle f, e_c \rangle - B(u_h, \eta) - B(u_h, \Pi_0 e_c) \\ =& \langle f, \eta \rangle - B(u_h, \eta), \end{aligned} \tag{41}$$

which, noting that $[\![e_c]\!] = 0$ on Γ, implies

$$\|\nabla e_c\|^2 = B(e_c, e_c) = \langle f, \eta \rangle - B(u_h, \eta) - B(e_d, e_c). \tag{42}$$

For the last term on the right-hand side of (42), we have

$$|B(e_d, e_c)| \leq \|\nabla e_d\| \|\nabla e_c\| + \frac{1}{2} \sum_{e \subset \Gamma} \sum_{T = T^+, T^-} \|\sqrt{h}(\Pi \nabla e_c)|_T\|_e \|h^{-1/2} [\![e_d]\!]\|_e, \tag{43}$$

where κ^+ and κ^- are the (generic) elements having e as common face. Using the inverse inequality (24) and the stability of the L^2-projection, we arrive to

$$|B(e_d, e_c)| \leq \|\nabla e_d\| \|\nabla e_c\| + C \|\nabla e_c\| \|\sqrt{\sigma} [\![e_d]\!]\|_\Gamma. \tag{44}$$

Finally, noting that $[\![e_d]\!] = [\![u_h]\!]$, and making use of (37) we conclude that

$$|B(e_d, e_c)| \leq C \|\nabla e_c\| \, \|\sqrt{\sigma} [\![u_h]\!]\|_\Gamma. \tag{45}$$

To bound the first two terms on the right-hand side of (42), we begin by an element-wise integration by parts yielding

$$\begin{aligned} \langle f, \eta \rangle - B(u_h, \eta) = & \sum_{T \in \mathcal{T}} \int_T \big(f + \Delta u_h\big) \eta - \int_{\Gamma_{\text{int}}} \{\eta\} [\![\nabla u_h]\!] \, \mathrm{d}s \\ & - \int_\Gamma \{\Pi \nabla \eta\} \cdot [\![u_h]\!] \, \mathrm{d}s - \int_\Gamma \sigma [\![u_h]\!] \cdot [\![\eta]\!] \, \mathrm{d}s. \end{aligned} \tag{46}$$

The first term on the right-hand side of (46) can be bounded as follows:

$$\begin{aligned} \Big| \sum_{T \in \mathcal{T}} \int_T \big(f + \Delta u_h\big) \, \eta \Big| \leq & \|h(f + \Delta u_h)\| \, \|h^{-1} \eta\| \\ \leq & C \|h(f + \Delta u_h)\| \, \|\nabla e_c\|, \end{aligned} \tag{47}$$

upon observing that $\|h^{-1}\eta\|_\kappa \leq C\|\nabla e_c\|_\kappa$.

For the second term on the right-hand side of (46), we use the trace estimate (32), the bound $\|h^{-1}\eta\|_\kappa \leq C\|\nabla e_c\|_\kappa$ and the observation that $\nabla\eta = \nabla e_c$, to deduce

$$\Big| \int_{\Gamma_{\rm int}} \{\eta\}[\![\nabla u_h]\!]\,{\rm d}s \Big| \leq C\|\nabla e_c\|\, \|\sqrt{h}[\![\nabla u_h]\!]\|_{\Gamma_{\rm int}}. \tag{48}$$

For the third term on the right-hand side of (46), we use $\nabla\eta = \nabla e_c$ and, working alike to (43), we obtain

$$\Big| \int_{\Gamma} \{\Pi\nabla\eta\} \cdot [\![u_h]\!] \Big| \leq C\|\nabla e_c\|\, \|\sqrt{\sigma}[\![u_h]\!]\|_{\Gamma}, \tag{49}$$

and finally, for the last term on the right-hand side of (46), we get

$$\Big| \int_{\Gamma} \sigma[\![\eta]\!] \cdot [\![u_h]\!] \Big| \leq C\|\nabla e_c\|\, \|\sqrt{\sigma}[\![u_h]\!]\|_{\Gamma}. \tag{50}$$

The result follows combining the above relations. □

5 Discontinuous Galerkin methods for first order hyperbolic problems

The development of dG methods for elliptic problems, introduced above, is an interesting theoretical development and offers a number of advantages in particular cases, for instance, when using irregular meshes or, perhaps, "exotic" basis functions such as wavelets. However, the major argument for using dG methods lies with their ability to provide stable numerical methods for first order PDE problems, for which classical FEM is well known to perform poorly.

We consider the first order Cauchy problem

$$\mathcal{L}_0 u \equiv b\cdot\nabla u + cu = f \quad \text{ in } \Omega, \tag{51}$$

$$u = g \ \text{ on } \partial_-\Omega, \tag{52}$$

where

$$\partial_-\Omega := \{x \in \partial\Omega : b(x)\cdot n(x) < 0\}$$

is the inflow part of the domain boundary $\partial\Omega$, $b := (b_1, \ldots, b_d) \in [C^1(\bar{\Omega})]^d$ and $g \in L^2(\partial_-\Omega)$.

We assume further that there exists a positive constant γ_0 such that

$$c(x) - \frac{1}{2}\nabla\cdot b(x) \geq \gamma_0 \ \text{ for almost every } x \in \Omega, \tag{53}$$

and we define $c_0 := (c - 1/2\nabla \cdot b)^{1/2}$.

Next, we consider a mesh $\mathcal{T}$ of the domain Ω as above, and we define

$$\partial_- T := \{x \in \partial T : b(x) \cdot n(x) < 0\}, \quad \partial_+ T := \{x \in \partial T : b(x) \cdot n(x) > 0\},$$

for each element T; we call these the *inflow* and *outflow* parts of ∂T respectively. For $T \in \mathcal{T}$, and a (possibly discontinuous) element-wise smooth function v, we consider the *upwind jump* across the inflow boundary $\partial_- T$, by

$$\lfloor v \rfloor(x) := \lim_{t \to 0^+} \big(u(x + tb) - u(x - tb)\big),$$

for almost all $x \in \partial_- T$, when $\partial_- T \subset \Gamma_{\text{int}}$, and by $\lfloor v \rfloor(x) := v(x)$ for almost all $x \in \partial_- T$, when $\partial_- T \subset \partial_- \Omega$.

We require some more notation to describe the method. Let $u \in H^1(\Omega, \mathcal{T})$; then, for every element $T \in \mathcal{T}$, we denote by u_T^+ the trace of u on $\partial\kappa$ taken from within the element T (interior trace). We also define the exterior trace u_T^- of $u \in H^1(\Omega, \mathcal{T})$ for almost all $x \in \partial_- T \backslash \Gamma$ to be the interior trace $u_{T'}^+$ of u on the element(s) T' that share the edges contained in $\partial_- T \backslash \Gamma$ of the boundary of element T. Then, the *jump* of u across $\partial_- T \backslash \Gamma$ is defined by

$$\lfloor u \rfloor_T := u_T^+ - u_T^-.$$

We note that this definition of jump is not the same as the one in the pure diffusion case discussed in the previous section; here the sign of the jump depends on the direction of the flow, whereas in the pure diffusion case it only depends on the element-numbering. Since they may genuinely differ up to a sign, we have used different notation for the jumps in the two cases. Again, we note that the subscripts will be suppressed when no confusion is likely to occur.

We shall now describe the construction of the discontinuous Galerkin weak formulation for the problem (51), (52), by imposing "weakly" the value of the solution on an outflow boundary of an element as an inflow boundary for the neighbouring downstream elements, we solve small local problems, until we have found the solution over the complete domain Ω.

We first construct a local weak formulation on every element T that is attached to the inflow boundary of the domain.

We define the space $\mathcal{S}_{\text{adv}} := G_b + S_h^p$, for $p \geq 0$ (note that $p = 0$ is allowed in the dG discretization of first order hyperbolic problems), where

$$G_b := \{w \in L^2(\Omega) : b \cdot \nabla w \in L^2(\Omega)\},$$

is the graph space of the PDE (51). Multiplying with a test function $v \in \mathcal{S}_{\text{adv}}$ and integrating over T we obtain

$$\int_T (\mathcal{L}_0 u) v \, \mathrm{d}x = \int_T f \, v \, \mathrm{d}x. \tag{54}$$

Now we impose the boundary conditions for the local problem. Since $\partial_- T \cap \Gamma_- \neq \emptyset$ we have $u^+ = g$ on $\partial_- T \cap \Gamma_-$. Therefore, after multiplication by $(b \cdot n)v^+$ and integration over $\partial_- T \cap \Gamma_-$, we get

$$\int_{\partial_- T \cap \Gamma_-} (b \cdot n) u^+ v^+ \, \mathrm{d}s = \int_{\partial_- T \cap \Gamma_-} (b \cdot n) g v^+ \, \mathrm{d}s. \tag{55}$$

Upon subtracting (55) from (54) we have

$$\int_T (\mathcal{L}_0 u) v \, \mathrm{d}x - \int_{\partial_- T \cap \Gamma_-} (b \cdot n) u^+ v^+ \, \mathrm{d}s = \int_T f \, v \, \mathrm{d}x - \int_{\partial_- T \cap \Gamma_-} (b \cdot n) g v^+ \, \mathrm{d}s. \tag{56}$$

We shall now deal with the remaining parts of the inflow boundary of the element T. The key idea in the discontinuous Galerkin method is to impose the boundary conditions "weakly", i.e., via integral identities. Therefore, we set as local boundary conditions for the element T on $\partial_- T \backslash \Gamma_-$, the exterior trace of the function u, and we impose them in the same way as the actual inflow boundary part:

$$\int_{\partial_- T \backslash \Gamma_-} (b \cdot n) u^+ v^+ \, \mathrm{d}s = \int_{\partial_- T \backslash \Gamma_-} (b \cdot n) u^- v^+ \, \mathrm{d}s, \tag{57}$$

which is equivalent to

$$\int_{\partial_- T \backslash \Gamma_-} (b \cdot n) \lfloor u \rfloor v^+ \, \mathrm{d}s = 0. \tag{58}$$

In order to justify the validity of (57) we have to resort to the classical theory of characteristics for hyperbolic equations. It is known that the solution of a first-order linear hyperbolic boundary-value problem can only exhibit jump discontinuities across characteristics. Thus the normal flux of the solution $bu \cdot n$ is a continuous function across the element faces $e \subset \Gamma_{\mathrm{int}}$ if $(b \cdot n)|_e \neq 0$, as in that case the element face does not lie on a characteristic. If $(b \cdot n)|_e = 0$, which is the case when e lies on a characteristic, then we have $bu \cdot n = (b \cdot n)u = 0$ on e. Hence in any case we have continuity of the normal flux and therefore (58) and thus (57) hold for all $T \in \mathcal{T}$.

Now, subtracting (57) from (56), we obtain

$$\begin{aligned} &\int_T (\mathcal{L}_0 u) v \, \mathrm{d}x - \int_{\partial_- T \cap \Gamma_-} (b \cdot n) u^+ v^+ \, \mathrm{d}s - \int_{\partial_- T \backslash \Gamma_-} (b \cdot n) \lfloor u \rfloor v^+ \, \mathrm{d}s \\ &= \int_T f v \, \mathrm{d}x - \int_{\partial_- T \cap \Gamma_-} (b \cdot n) g v^+ \, \mathrm{d}s \end{aligned}$$

for all $T \in \mathcal{T}$ such that $\partial_- T \backslash \Gamma_- \neq \emptyset$.

Arguing in the same way as above, we obtain the local weak formulation for the elements whose boundaries do not share any points with the inflow

boundary Γ_- of the computational domain; in this case though the second terms on the left-hand side and the right-hand side of (59) do not appear:

$$\int_T (\mathcal{L}_0 u)v \,\mathrm{d}x - \int_{\partial_- T \setminus \Gamma_-} (b \cdot n) \lfloor u \rfloor v^+ \,\mathrm{d}s = \int_T f v \,\mathrm{d}x, \tag{59}$$

for all $T \in \mathcal{T}$ such that $\partial_- T \cap \Gamma_- = \emptyset$. Adding up all these and setting

$$\begin{aligned} B_{\mathrm{adv}}(u,v) := & \sum_{T\in\mathcal{T}} \int_T (\mathcal{L}_0 u)v \,\mathrm{d}x - \sum_{T\in\mathcal{T}} \int_{\partial_- T \cap \Gamma_-} (b\cdot n) u^+ v^+ \,\mathrm{d}s \\ & - \sum_{T\in\mathcal{T}} \int_{\partial_- T \setminus \Gamma_-} (b\cdot n) \lfloor u \rfloor v^+ \,\mathrm{d}s \\ l_{\mathrm{adv}}(v) := & \sum_{T\in\mathcal{T}} \int_T f v \,\mathrm{d}x - \sum_{T\in\mathcal{T}} \int_{\partial_- T \cap \Gamma_-} (b\cdot n) g v^+ \,\mathrm{d}s \end{aligned} \tag{60}$$

we can write the weak form for the problem (51):

$$\text{Find } u \in G_b \text{ such that } B_{\mathrm{adv}}(u,v) = l_{\mathrm{adv}}(v) \quad \forall v \in \mathcal{S}_{\mathrm{adv}}.$$

The discontinuous Galerkin method for the problem (51) then reads:

$$\text{Find } u_h \in S_h^p \text{ such that } B_{\mathrm{adv}}(u_h, v_h) = l_{\mathrm{adv}}(v_h) \quad \forall v_h \in S_h^p.$$

5.1 Error analysis of the dG method

We define the *energy norm*, denoted again by $|||\cdot|||$, (without causing, hopefully, any confusion) by

$$|||w|||_{\mathrm{adv}} := \Big(\sum_{T\in\mathcal{T}} \|c_0 w\|_\Omega^2 + \frac{1}{2} \|b_n [\![w]\!]\|_\Gamma^2 \Big)^{1/2},$$

where $b_n := \sqrt{|b \cdot \mathbf{n}|}$, with n on ∂T denoting the outward normal to ∂T and σ as above. The choice of the above energy norm is related to the coercivity of the bilinear form $B_{\mathrm{adv}}(\cdot,\cdot)$.

The definition and properties of the dG method for the advection problem may become clearer, by studying the symmetric and the skew-symmetric parts of the bilinear form $B_{\mathrm{adv}}(\cdot,\cdot)$. Indeed, it is possible to rewrite the numerical fluxes as described in the following result.

Lemma 3. *The following identity holds:*

$$
\begin{aligned}
&-\sum_{T\in\mathcal{T}}\left(\int_{\partial_- T\cap\Gamma_-}(b\cdot n)u^+v^+\,\mathrm{d}s+\int_{\partial_- T\setminus\Gamma_-}(b\cdot n)\lfloor u\rfloor v^+\,\mathrm{d}s\right)\\
&=\int_\Gamma\left(\frac{1}{2}|b\cdot n|[\![u]\!]\cdot[\![v]\!]-[\![u]\!]\cdot\{\!\{bv\}\!\}\right)\mathrm{d}s+\frac{1}{2}\int_{\partial\Omega}(b\cdot n)u^+v^+\,\mathrm{d}s.
\end{aligned}
$$

Proof. On each elemental inflow boundary, we have $-(b\cdot n)=|b\cdot n|$. Thus, on each $\partial_- T\setminus\Gamma_-$, we have

$$
\begin{aligned}
-(b\cdot n)\lfloor u\rfloor v^+&=|b\cdot n|\lfloor u\rfloor v^+\\
&=\frac{1}{2}|b\cdot n|\lfloor u\rfloor\lfloor v\rfloor+|b\cdot n|\lfloor u\rfloor\{\!\{v\}\!\}\\
&=\frac{1}{2}|b\cdot n|[\![u]\!]\cdot[\![v]\!]-(b\cdot n)\lfloor u\rfloor\{\!\{v\}\!\}\\
&=\frac{1}{2}|b\cdot n|[\![u]\!]\cdot[\![v]\!]-[\![u]\!]\cdot\{\!\{bv\}\!\}.
\end{aligned}
\tag{61}
$$

Hence,

$$
-\sum_{T\in\mathcal{T}}\int_{\partial_- T\setminus\Gamma_-}(b\cdot n)\lfloor u\rfloor v^+\,\mathrm{d}s=\int_{\Gamma_{\mathrm{int}}}\left(\frac{1}{2}|b\cdot n|[\![u]\!]\cdot[\![v]\!]-[\![u]\!]\cdot\{\!\{bv\}\!\}\right)\mathrm{d}s.
\tag{62}
$$

Recalling the definitions of $[\![\cdot]\!]$ and $\{\!\{\cdot\}\!\}$ on the boundary $\partial\Omega$, along with the identities $-(b\cdot n)=|b\cdot n|$ and $(b\cdot n)=|b\cdot n|$ on the inflow and outflow parts of the boundary, respectively, it is immediate that

$$
\begin{aligned}
&\int_{\partial\Omega}\left(\frac{1}{2}|b\cdot n|[\![u]\!]\cdot[\![v]\!]-[\![u]\!]\cdot\{\!\{bv\}\!\}\right)\mathrm{d}s+\frac{1}{2}\int_{\partial\Omega}(b\cdot n)u^+v^+\,\mathrm{d}s\\
&=-\sum_{T\in\mathcal{T}}\int_{\partial_- T\cap\Gamma_-}(b\cdot n)u^+v^+\,\mathrm{d}s.
\end{aligned}
\tag{63}
$$

By summing (62) and (63), the result follows. □

The above observation shows that the dG method for the advection problem contains a symmetric part on both the face terms and the elemental terms of the bilinear form [18, 10].

Motivated by identity (61), we decompose $B_{\mathrm{adv}}(\cdot,\cdot)$ into symmetric and skew-symmetric components.

Lemma 4. *The bilinear form can be decomposed into symmetric and skew-symmetric parts:*

$$
B_{\mathrm{adv}}(w,v)=B_{\mathrm{adv}}^{\mathrm{symm}}(u,v)+B_{\mathrm{adv}}^{\mathrm{skew}}(u,v)
$$

for all $u,v\in\mathcal{S}_{\mathrm{adv}}$, where

$$B_{\rm adv}^{\rm symm}(u,v) := \sum_{T\in\mathcal{T}} \int_T c_0^2\, u\, v \,\mathrm{d}x + \frac{1}{2}\int_\Gamma |b\cdot n| [\![u]\!]\cdot[\![v]\!]\,\mathrm{d}s$$

and

$$\begin{aligned} B_{\rm adv}^{\rm skew}(u,v) := \frac{1}{2}\sum_{T\in\mathcal{T}} \int_T \big((b\cdot\nabla u)\, v - (b\cdot\nabla v)\, u\,\big)\,\mathrm{d}x \\ + \frac{1}{2}\int_{\Gamma_{\rm int}} \big([\![v]\!]\cdot\{\!\{bu\}\!\} - [\![u]\!]\cdot\{\!\{bv\}\!\}\big)\,\mathrm{d}s. \end{aligned}$$

Proof. By adding and subtracting $1/2\sum_{T\in\mathcal{T}}\int_T \nabla\cdot buv\,\mathrm{d}x$ to the bilinear form, a straightforward calculation yields

$$B_{\rm adv}(u,v) = B_{\rm adv}^{\rm symm}(u,v) + \sum_{T\in\mathcal{T}}\int_T \Big((b\cdot\nabla u)\, v + \frac{\nabla\cdot b}{2}\, u\, v\Big)\,\mathrm{d}x - \int_\Gamma [\![u]\!]\cdot\{\!\{bv\}\!\}\,\mathrm{d}s, \tag{64}$$

with $B_{\rm adv}^{\rm symm}(u,v)$ as defined in (64). Integration by parts of the second term in the first integral on the right-hand side of (64) yields

$$\begin{aligned} \sum_{T\in\mathcal{T}}\int_T \frac{\nabla\cdot b}{2}\, u\, v\,\mathrm{d}x = -\frac{1}{2}\sum_{T\in\mathcal{T}}\int_T \big((b\cdot\nabla u)\, v + (b\cdot\nabla v)\, u\,\big)\,\mathrm{d}x \\ + \frac{1}{2}\sum_{T\in\mathcal{T}}\int_{\partial T} (b\cdot n) u^+ v^+\,\mathrm{d}s. \end{aligned} \tag{65}$$

The result follows by making use of the (standard) identity (see, e.g., [4])

$$\sum_{T\in\mathcal{T}}\int_{\partial T}(b\cdot n)u^+v^+\,\mathrm{d}s = \int_\Gamma [\![u]\!]\cdot\{\!\{bv\}\!\}\,\mathrm{d}s + \int_{\Gamma_{\rm int}} \{\!\{u\}\!\}[\![bv]\!]\,\mathrm{d}s \tag{66}$$

and by observing that $\{\!\{u\}\!\}[\![bv]\!] = \{\!\{bu\}\!\}\cdot[\![v]\!]$. □

Remark 2. We observe the coercivity of the bilinear form:

$$B_{\rm adv}(w,w) = |\!|\!| w |\!|\!|_{\rm adv}^2, \tag{67}$$

for all $w\in\mathcal{S}_{\rm adv}$, as $B_{\rm adv}^{\rm symm}(w,w) = |\!|\!| w |\!|\!|_{\rm adv}^2$ and $B_{\rm adv}^{\rm skew}(w,w) = 0$.

To prove a priori error bounds for the dG method, we begin by observing the Galerkin orthogonality property

$$B_{\rm adv}(u-u_h, v_h) = 0, \tag{68}$$

for all $v_h\in\mathcal{S}_{\rm adv}$, coming from subtracting the dG method from the weak form of the problem, tested again functions from the finite element space.

For simplicity of the presentation, we shall assume in the sequel that

$$b\cdot\nabla v_h \in S_h^p. \tag{69}$$

Results for more general wind b are available, e.g., in [36, 27].

Using (67) and (68), we get the identity

$$|||v_h - u_h|||^2_{\rm adv} = B_{\rm adv}(v_h - u_h, v_h - u_h) = -B_{\rm adv}(u - v_h, v_h - u_h), \quad (70)$$

for all $v_h \in s_h^p$.

The next step is to bound the bilinear form from above by a multiple of $|||v_h - u_h|||_{\rm adv}$. To this end, we work as follows. Integrating by parts the first term in the integrant of the first term on the right-hand side of (64) and using the standard identity (66), we can arrive to

$$\begin{aligned} B_{\rm adv}(u - v_h, v_h - u_h) =& B^{\rm symm}_{\rm adv}(u - v_h, v_h - u_h) \\ & - \sum_{T\in\mathcal{T}} \int_T (b\cdot\nabla(v_h - u_h))\,(u - v_h)\,{\rm d}x \\ & + \int_\Gamma [\![v_h - u_h]\!] \cdot \{\!\{ b(u - v_h) \}\!\}\,{\rm d}s. \end{aligned} \quad (71)$$

Setting $v_h = \Pi u$, where, as above, $\Pi : L^2(\Omega) \to S - h^p$ is the orthogonal L^2-projection operator onto the finite element space, we observe that the second term on the right-hand side of (71) vanishes in view of (69). The Cauchy-Schwarz inequality then yields

$$B_{\rm adv}(u-\Pi u, \Pi u - u_h) \le 2|||\Pi u - u_h|||_{\rm adv}\big(|||u - \Pi u|||^2_{\rm adv} + \|\{\!\{u - \Pi u\}\!\}\|^2_\Gamma\big)^{1/2}, \quad (72)$$

which can be used on (70) to deduce

$$|||\Pi u - u_h|||_{\rm adv} \le 2\big(|||u - \Pi u|||^2_{\rm adv} + \|\{\!\{u - \Pi u\}\!\}\|^2_\Gamma\big)^{1/2}, \quad (73)$$

which, using triangle inequality and the approximation properties of the L^2-projection (see, e.g., [36] for details), yields the a priori error bound

$$|||u - u_h|||_{\rm adv} \le Ch^{\min\{p+1,r\}-1/2}|u|_{H^r(\Omega)}, \quad (74)$$

for $p \ge 0$ and $r \ge 1$.

6 Problems with non-negative characteristic form

Having considered the dG method for self-adjoint elliptic and first order hyperbolic problems respectively, we are now in position to combine the ideas presented above and present a dG method for a wide class of linear PDE problems.

Let Ω be a bounded open (curvilinear) polygonal domain in $\mathbb{R}^d$, and let $\partial\Omega$ signify the union of its $(d-1)$-dimensional open edges, which are assumed

to be sufficiently smooth (in a sense defined rigorously later). We consider the convection-diffusion-reaction equation

$$\mathcal{L}u \equiv -\nabla \cdot (\bar{\mathbf{a}}\nabla u) + b \cdot \nabla u + cu = f \quad \text{in } \Omega, \tag{75}$$

where $f \in L^2(\Omega)$, $c \in L^\infty(\Omega)$, b is a vector function whose entries are Lipschitz continuous real-valued functions on $\bar{\Omega}$, and $\bar{\mathbf{a}}$ is the *symmetric* diffusion tensor whose entries are bounded, piecewise continuous real-valued functions defined on $\bar{\Omega}$, with

$$\zeta^T \bar{\mathbf{a}}(x)\zeta \geq 0 \quad \forall \zeta \in \mathbb{R}^d, \quad x \in \bar{\Omega}. \tag{76}$$

Under this hypothesis, (75) is termed a *partial differential equation with a nonnegative characteristic form*. By n we denote the unit outward normal vector to $\partial\Omega$. We define

$$\Gamma_0 = \left\{ x \in \partial\Omega : n(x)^T \bar{\mathbf{a}}(x) n(x) > 0 \right\},$$

$$\Gamma_- = \{ x \in \partial\Omega \backslash \Gamma_0 : b(x) \cdot n(x) < 0 \}, \quad \Gamma_+ = \{ x \in \partial\Omega \backslash \Gamma_0 : b(x) \cdot n(x) \geq 0 \}.$$

The sets Γ_- and Γ_+ are referred to as the *inflow* and *outflow* boundary, respectively. We can also see that $\partial\Omega = \Gamma_0 \cup \Gamma_- \cup \Gamma_+$. If Γ_0 has a positive $(d-1)$-dimensional Hausdorff measure, we also decompose Γ_0 into two parts Γ_D and Γ_N, and we impose Dirichlet and Neumann boundary conditions, respectively, via

$$\begin{aligned} u &= g_\mathrm{D} \text{ on } \Gamma_\mathrm{D} \cup \Gamma_-, \\ (\bar{\mathbf{a}}\nabla u) \cdot n &= g_\mathrm{N} \text{ on } \Gamma_\mathrm{N}, \end{aligned} \tag{77}$$

where we adopt the (physically reasonable) hypothesis that $b \cdot n \geq 0$ on Γ_N, whenever the latter is nonempty.

For a discussion on the physical models that are described by the above family of boundary-value problems, we refer to [36] and the references therein. The existence and uniqueness of solutions (in various settings) has been considered in [45, 25, 26, 37], under the standard assumption (53).

Then the *interior penalty dG method* for the problem (75), (77) is defined as follows:

$$\text{Find } u_h \in S_h^p \text{ such that } B(u_h, v_h) = l(v_h) \quad \forall v_h \in S_h^p, \tag{78}$$

where

$$B(w,v) := \sum_{T\in\mathcal{T}} \int_T \left(\bar{\mathbf{a}}\nabla w\cdot\nabla v + (b\cdot\nabla w)\, v + c\, w\, v\right)\,\mathrm{d}x$$
$$- \sum_{T\in\mathcal{T}} \int_{\partial_- T\cap(\Gamma_-\cup\Gamma_{\mathrm{D}})} (b\cdot n) w^+ v^+\,\mathrm{d}s - \sum_{T\in\mathcal{T}} \int_{\partial_- T\setminus\partial\Omega} (b\cdot n)\lfloor w\rfloor v^+\,\mathrm{d}s$$
$$+ \int_{\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{int}}} \left(\theta\{\bar{\mathbf{a}}\nabla v\}\cdot[\![w]\!] - \{\bar{\mathbf{a}}\nabla w\}\cdot[\![v]\!] + \sigma[\![w]\!]\cdot[\![v]\!]\right)\,\mathrm{d}s$$

and

$$l(v) := \sum_{T\in\mathcal{T}} \int_T fv\,\mathrm{d}x - \sum_{T\in\mathcal{T}} \int_{\partial_- T\cap(\Gamma_-\cup\Gamma_{\mathrm{D}})} (b\cdot n) g_{\mathrm{D}} v^+\,\mathrm{d}s$$
$$+ \int_{\Gamma_{\mathrm{D}}} (\theta\bar{\mathbf{a}}\nabla v\cdot n + \sigma v) g_{\mathrm{D}}\,\mathrm{d}s + \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}} v\,\mathrm{d}s \tag{79}$$

for $\theta\in\{-1,1\}$, with the function σ defined by

$$\sigma|_e := C_\sigma \left\{\frac{\mathtt{a}p^2}{h}\right\},$$

where $\mathtt{a}:\Omega\to\mathbb{R}$, with $\mathtt{a}|_T = \|(|\sqrt{\bar{\mathbf{a}}}|_2)^2\|_{L^\infty(T)}$, $T\in\mathcal{T}$, with $|\cdot|_2$ denoting the matrix-2-norm, and C_σ is a positive constant. We refer to the dG method with $\theta=-1$ as the *symmetric interior penalty dG method*, whereas $\theta=1$ yields the *nonsymmetric interior penalty dG method*. This terminology stems from the fact that when $b\equiv\mathbf{0}$, the bilinear form $B(\cdot,\cdot)$ is symmetric if and only if $\theta=-1$.

Various types of error analysis for the variants of interior penalty DGFEMs can be found, e.g., in [6, 3, 15, 49, 36, 4, 27, 31, 29, 28, 21, 20, 38], along with an extensive discussion on the properties of this family of methods.

7 Numerical examples

7.1 Example 1

We consider the first IAHR/CEGB problem (devised by workers at the CEGB for an IAHR workshop in 1981 as a benchmark steady-state convection-diffusion problem). For

$$b = \Big(2y(1-x^2), -2x(1-y^2)\Big)$$

and $0\le\epsilon\ll 1$, we consider the convection-diffusion equation

$$-\epsilon\Delta u + b\cdot\nabla u = 0 \quad \text{for} \quad (x_1,x_2)\in(-1,1)\times(0,1),$$

subject to Dirichlet boundary conditions

$$u(-1,x_2) = u(x_1,1) = u(1,x_2) = 1 - \tanh(\alpha), \quad -1 \le x_1 \le 1,\ 0 \le x_2 \le 1,$$

on the tangential boundaries, with $\alpha > 0$ parameter, and inlet boundary condition

$$u(x_1,0) = 1 + \tanh\Big(\alpha(2x+1)\Big), \quad -1 \le x_1 \le 0. \tag{80}$$

Finally, a homogeneous Neumann boundary condition is imposed at the outlet $0 < x_1 \le 1$, $x_2 = 0$.

We remark that this choice of convective velocity field b does not satisfy assumption (53). On the other hand b is incompressible, that is $\nabla \cdot b = 0$, and, therefore, $c_0 = 0$.

The inlet profile (80) involves the presence of a steep interior layer centred at $(-1/2, 0)$, whose steepness depends on the value of the parameter α. This layer travels clockwise circularly due to the convection and exits at the outlet.

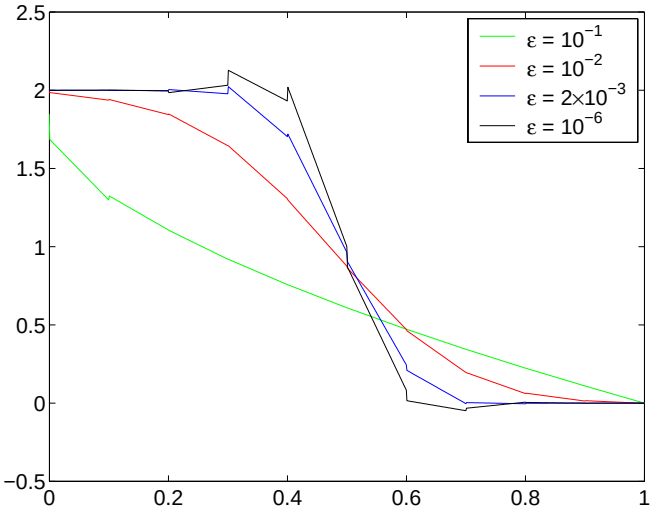


Fig. 3 Example 1. Outlet profiles for different values of ϵ.

Following MacKenzie & Morton [43] (cf. also Smith & Hutton [52]) we have chosen to work with $\alpha = 10$ on a uniform mesh of 20×10 elements, and for $\epsilon = 10^{-6}, 2 \times 10^{-3}, 10^{-2}, 10^{-1}$, respectively.

In Figure 3 the profiles of the outlet boundary $0 < x_1 \le 1$, $x_2 = 0$ are plotted for different values of ϵ, and for $p = 1$. Note that the vertical line segments in the profiles correspond to the discontinuities across the element interfaces. To address the question of accuracy of the computation, in Figure

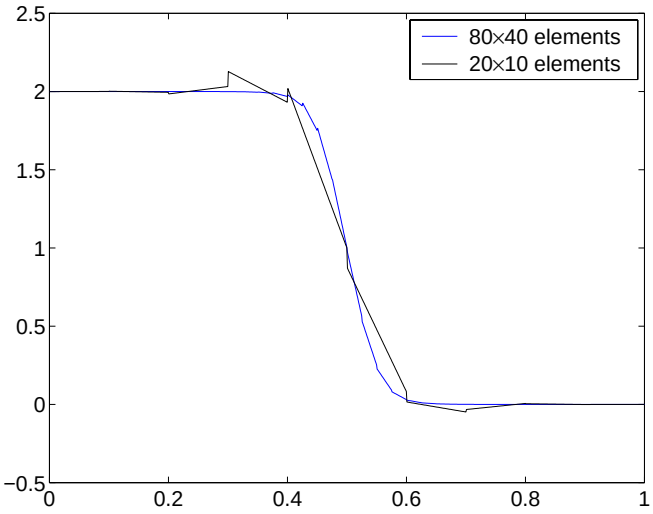


Fig. 4 Example 1. Outlet profile for $\epsilon = 10^{-6}$ when 20×10 elements and 80×40 elements are used.

4 we compare the profile for $\epsilon = 10^{-6}$ (drawn in black in Figure 3) on the 20×10 mesh with the corresponding profile on a much finer mesh containing 80×40 elements. Also, in Figure 5 we present the computed outlet profiles when we use of uniform polynomial degrees $p = 1, \ldots, 4$ on the 20×10-mesh. Note that the quality of the approximation is better for $p = 4$ on the 20×10-mesh (DOF= 5000), than the computed outlet profile for $p = 1$ on the 80×40 mesh (DOF= 12800).

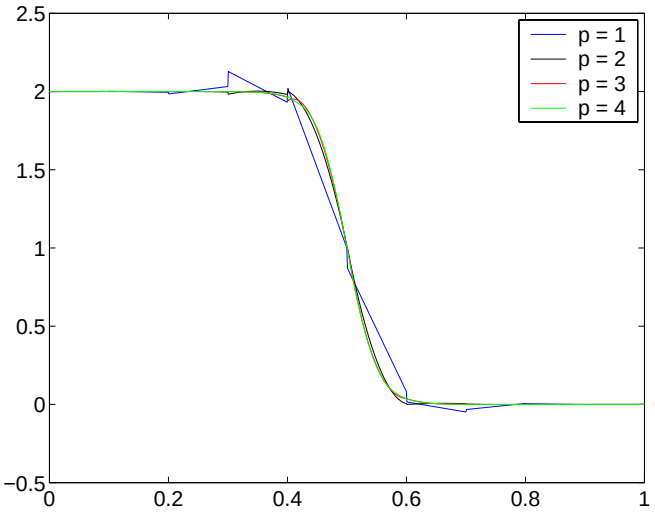
Finally, in Figure 6 we present the computed solutions on the 20×10-mesh for the different values of ϵ. We note that the quality of the approximations is remarkably good considering the computationally demanding features of the solutions.

7.2 Example 2

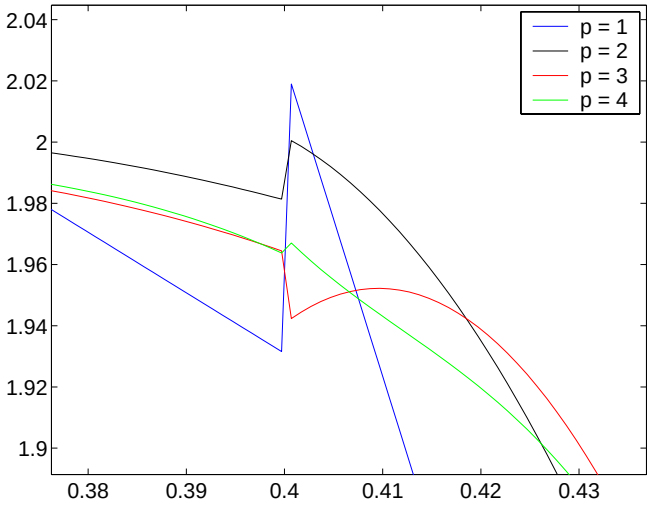
We consider the following equation on $\Omega = (-1, 1)^2$

$$\begin{aligned} -x_1^2 u_{x_2 x_2} + u_{x_1} + u = 0, \quad \text{for } -1 \leq x_1 \leq 1, \ x_2 > 0, \\ u_{x_1} + u = 0, \quad \text{for } -1 \leq x_1 \leq 1, \ x_2 \leq 0, \end{aligned}$$

whose analytical solution is

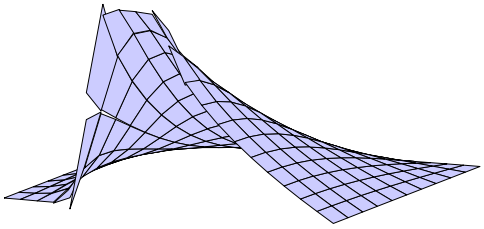


(a) Outlet profile for $\epsilon = 10^{-6}$ for $p = 1, \ldots, 4$

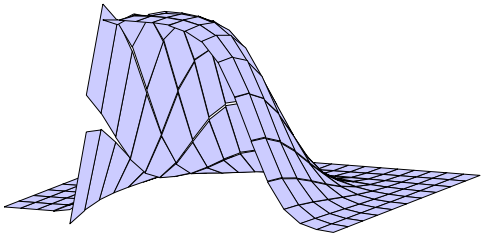


(b) Detail of (a)

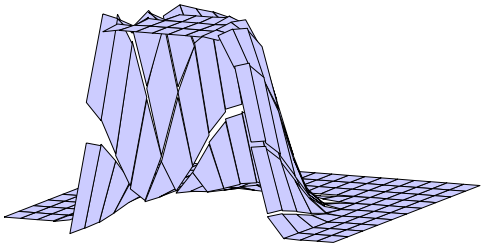
Fig. 5 Example 1. Outlet profile for $\epsilon = 10^{-6}$ on the 20×10 mesh for $p = 1, \ldots, 4$.



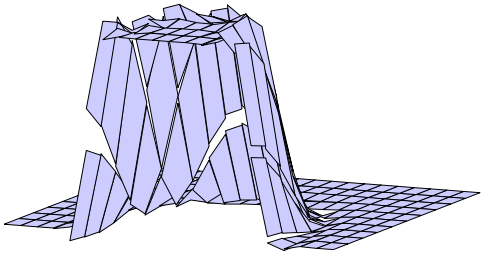
(a) $\epsilon = 10^{-1}$



(b) $\epsilon = 10^{-2}$



(c) $\epsilon = 2 \times 10^{-3}$



(d) $\epsilon = 10^{-6}$

Fig. 6 Example 1. Numerical solutions on the 20×10-mesh for $p = 1$ and for $\epsilon = 10^{-1}, 10^{-2}, 2 \times 10^{-3}, 10^{-6}$, respectively.

$$u(x_1,x_2)=\begin{cases}\sin\left(\frac{1}{2}\pi(1+x_2)\right)\exp\left(-\left(x_1+\frac{\pi^2}{4}\frac{x_1^3}{3}\right)\right), & \text{if } x_1\in[-1,1], x_2>0;\\ \sin\left(\frac{1}{2}\pi(1+x_2)\right)\exp(-x_1), & \text{if } x_1\in[-1,1], x_2\le 0,\end{cases}$$

along with an appropriate Dirichlet boundary condition. This problem is of changing-type, as there exists a second order term for $x_2 > 0$, which is no longer present for $x_2 \le 0$. Moreover, we can easily verify that its analytical solution u exhibits a discontinuity along $x_2 = 0$, although the derivative of u, in the direction normal to this line of discontinuity in u, is continuous across $x_2 = 0$. We test the performance of dG method by employing various meshes. We have to modify the method by setting $\sigma_e = 0$ for all element edges $e \subset (-1, 1)\times\{0\}$, where σ_e denotes the discontinuity-penalisation parameter; this is done in order to avoid penalising physical discontinuities. Note that the diffusive flux $(\bar{\mathbf{a}}\nabla u)\cdot n$ is still continuous across $x_2 = 0$, and thus the method still applies.

When subdivisions with $(-1, 1)\times\{0\} \subset \bar{\Gamma}$ are used, the method appears to converge at exponential rates under p-enrichment. In Figure 7, we can see the convergence history for various such meshes. The reason for this excellent behaviour of the method, in a problem where standard conforming finite element methods would only provide us with low algebraic rates of convergence, lies in the fact that merely element-wise regularity is required for dG methods, as opposed to global regularity hypothesis that is needed for conforming methods to produce such results. If $(-1, 1)\times\{0\}$ is not a subset of $\bar{\Gamma}$, the method produces results inferior to the ones described for the case $(-1, 1)\times\{0\} \subset \bar{\Gamma}$, as the solution is then discontinuous within certain elements.

8 Solving the linear system

FEM and dG methods lead to large linear systems of the form $AU = F$, where usually the condition number $\kappa(A)$ of the matrix A increases as $h \to 0$; for the case of second order PDE problems we normally have $\kappa(A) = O(h^{-d})$. This is particularly inconvenient in the context of iterative methods for solving the linear system. Therefore, the construction of preconditioning strategies for the resulting linear system is of particular importance. Here we follow [30], where scalable solvers for linear systems arising from dG methods have been considered.

The classical preconditioning approach consists of designing a matrix P, called the *preconditioner*, such that the matrix $P^{-1}A$ is "well" conditioned compared to A (i.e., $\kappa(P^{-1}A) << \kappa(A)$) while at the same time P is cheaply inverted numerically. These two requirements are competing, constituting the construction of efficient preconditioners a challenge. Once such a precondi-

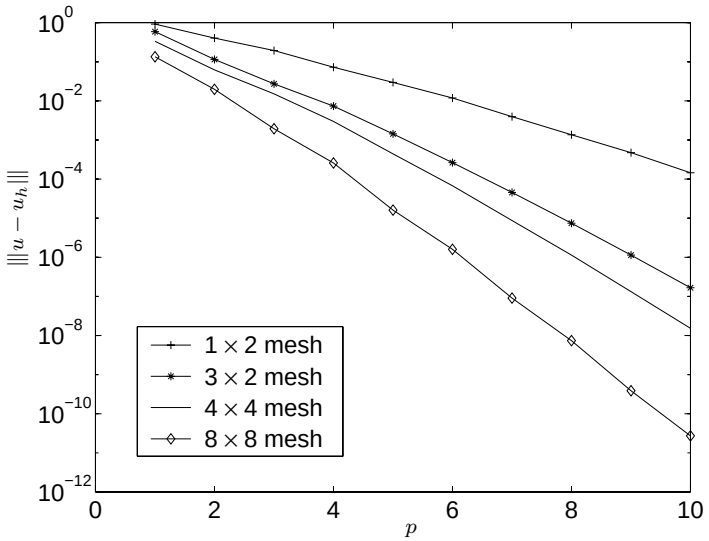


Fig. 7 Example 2. Convergence of the dG method in the dG-norm under p-enrichment.

tioner P is known, one can solve recursively the linear systems $Py = F$ and $P^{-1}Ax = y$ efficiently to find x.

In [30] it is proposed to use preconditioned GMRES iterative solver (see, e.g., [33]) with preconditioner

$$A_s := \frac{A + A^T}{2},$$

i.e., the symmetric part of the stiffness matrix A. This choice has a number of implications as we shall now see.

From an implementation point of view, employing GMRES with system matrix

$$A_s^{-\frac{1}{2}} A A_s^{-\frac{1}{2}},$$

is equivalent to running GMRES in the A_s-inner product and using A_s as a left-preconditioner.

It is not hard to see that

$$A_s^{-\frac{1}{2}} A A_s^{-\frac{1}{2}} = I + S,$$

where S is a skew-symmetric matrix. It is known that applying the GMRES algorithm applied to a matrix of the form $I + S$, where S is skew symmetric, is a 3-term recurrence, i.e., there is no need to compute using the whole Krylov subspace [2], but only the last two Krylov subspace vectors. Hence

using A_s as preconditioner within a GMRES algorithm leads to significant computational and storage savings.

Moreover, it is possible to show that the resulting preconditioned GMRES algorithm is *scalable* with respect to the size of the matrix, i.e., the number of GMRES iterations does not increase as the meshsize $h \to 0$ and/or as the polynomial degree $p \to \infty$. This property is shown theoretically in [30], but here we shall illustrate it via a numerical example.

To this end, we consider the convection-diffusion problem

$$-\epsilon \Delta u + u_x + u_y = f \quad \text{for } (x, y) \in (0, 1)^2,$$

subject to Dirichlet boundary conditions, and right-hand side f, such that the analytical solution is given by

$$u(x, y) = x + y(1 - x) + \frac{\mathrm{e}^{-\frac{1}{\epsilon}} - \mathrm{e}^{-\frac{(1-x)(1-y)}{\epsilon}}}{1 - \mathrm{e}^{-\frac{1}{\epsilon}}}$$

The solution exhibits boundary layer behaviour along $x = 1$ and $y = 1$, and the layers become steeper as $\epsilon \to 0$. We solve the problem for a range of ϵ using the dG method for a range of uniform meshsizes h and polynomial degrees p. The results are presented Table 1. As predicted by the theory in [30], the number of iterations is independent of discretization parameters.

For comparison purposes, we included the corresponding GMRES runs for the choice of a black-box preconditioner such as ILU. The results are presented in Table 2. We can see that, while the number of iterations is low for some values of the parameters, the overall convergence behavior is quite undesirable, with iteration counts growing with both discretization parameters. Thus, while the number of iterations appears to be decreasing with ϵ, it is exactly for this range that the discretization parameters have to be increased in order to resolve layers. The resulting convergence behavior becomes rapidly too costly to implement in practice. We note here that the ILU

Table 1 GMRES iterations for DGFEM discretization of the convection-diffusion problem with constant wind $\mathbf{b}^T = (1, 1)$ and with preconditioner A_s.

p	n	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.05$	$\epsilon = 0.01$
1	2,500	7	15	22	77
	10,000	7	15	22	80
	40,000	7	14	22	80
2	5,625	7	14	22	80
	22,500	6	14	22	80
	90,000	6	14	21	78
3	10,000	6	14	22	79
	40,000	6	14	22	78
	160,000	6	13	21	78

Table 2 GMRES iterations for dG discretization of the convection-diffusion problem with ILU(10^{-2}) preconditioning.

p	n	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.01$
1	2,500	12	13	7
	10,000	36	40	29
	40,000	124	117	69
2	5,625	18	17	12
	22,500	61	59	60
	90,000	235	231	137
3	10,000	39	29	23
	40,000	112	114	100
	160,000	> 300	> 300	> 300

preconditioner is implemented with a standard *full* GMRES routine, which means that the storage increases with every iteration.

9 Concluding remarks

These notes aim at giving a gentle introduction to discontinuous Galerkin methods used for the numerical solution of linear PDE problems of mixed type. The material is presented in a simple fashion in an effort to maximise accessbility. Indeed, this note is far from being exhaustive in any of the topics presented and, indeed, it is *not* meant to be a survey of the ever-growing subject of discontinuous Galerkin methods. For more material on dG methods we refer to the volumes [15, 34, 48] and the references therein.

References

1. R. A. Adams and J. J. F. Fournier, *Sobolev spaces*, vol. 140 of Pure and Applied Mathematics (Amsterdam), Elsevier/Academic Press, Amsterdam, second ed., 2003.
2. M. Arioli, D. Loghin, and A. J. Wathen, *Stopping criteria for iterations in finite element methods*, Numer. Math., 99 (2005), pp. 381–410.
3. D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
4. D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001/02), pp. 1749–1779 (electronic).
5. I. Babuška, *The finite element method with penalty*, Math. Comp., 27 (1973), pp. 221–228.
6. G. A. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

7. R. Becker, P. Hansbo, and M. G. Larson, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 723–733.
8. K. S. Bey and T. Oden, *hp-version discontinuous Galerkin methods for hyperbolic conservation laws*, Comput. Methods Appl. Mech. Engrg., 133 (1996), pp. 259–286.
9. S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.
10. F. Brezzi, L. D. Marini, and E. Süli, *Discontinuous Galerkin methods for first-order hyperbolic problems*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1893–1903.
11. C. Carstensen, T. Gudi, and M. Jensen, *A unifying theory of a posteriori error control for discontinuous Galerkin FEM*, Numer. Math., 112 (2009), pp. 363–379.
12. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.
13. B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, in High-order methods for computational physics, Springer, Berlin, 1999, pp. 69–224.
14. B. Cockburn, S. Hou, and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.
15. B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin methods*, Springer-Verlag, Berlin, 2000. Theory, computation and applications, Papers from the 1st International Symposium held in Newport, RI, May 24–26, 1999.
16. B. Cockburn, S. Y. Lin, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. III. One-dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.
17. B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework*, Math. Comp., 52 (1989), pp. 411–435.
18. ———, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463 (electronic).
19. ———, *The Runge-Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.
20. A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems. I. General theory*, SIAM J. Numer. Anal., 44 (2006), pp. 753–778.
21. A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order elliptic PDEs*, SIAM J. Numer. Anal., 44 (2006), pp. 2363–2388.
22. A. Ern and A. F. Stephansen, *A posteriori energy-norm error estimates for advection-diffusion equations approximated by weighted interior penalty methods*, J. Comp. Math., 26 (2008), pp. 488–510.
23. S. A. F. Ern, A. and P. Zunino, *A discontinuous galerkin method with weighted averages for advectiondiffusion equations with locally small and anisotropic diffusivity*, IMA J. Numer. Anal., 29 (2009), pp. 235–256.
24. R. S. Falk and G. R. Richter, *Local error estimates for a finite element method for hyperbolic and convection-diffusion equations*, SIAM J. Numer. Anal., 29 (1992), pp. 730–754.
25. G. Fichera, *Sulle equazioni differenziali lineari ellittico-paraboliche del secondo ordine*, Atti Accad. Naz. Lincei. Mem. Cl. Sci. Fis. Mat. Nat. Sez. I. (8), 5 (1956), pp. 1–30.
26. ———, *On a unified theory of boundary value problems for elliptic-parabolic equations of second order*, in Boundary problems in differential equations, Univ. of Wisconsin Press, Madison, 1960, pp. 97–120.
27. E. H. Georgoulis, *Discontinuous Galerkin methods on shape-regular and anisotropic meshes*, D.Phil. Thesis, University of Oxford, (2003).
28. E. H. Georgoulis, E. Hall, and J. M. Melenk, *On the suboptimality of the p-version interior penalty discontinuous Galerkin method*, J. Sci. Comput., 42 (2010), pp. 54–67.

29. E. H. Georgoulis and A. Lasis, *A note on the design of hp-version interior penalty discontinuous Galerkin finite element methods for degenerate problems.*, IMA J. Numer. Anal., 26 (2006), pp. 381–390.
30. E. H. Georgoulis and D. Loghin, *Norm preconditioners for discontinuous Galerkin hp-finite element methods*, SIAM J. Sci. Comput., 30 (2008), pp. 2447–2465.
31. E. H. Georgoulis and E. Süli, *Optimal error estimates for the hp–version interior penalty discontinuous Galerkin finite element method*, IMA J. Numer. Anal., 25 (2005), pp. 205–220.
32. D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, vol. 224 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, second ed., 1983.
33. G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, third ed., 1996.
34. J. S. Hesthaven and T. Warburton, *Nodal discontinuous Galerkin methods*, vol. 54 of Texts in Applied Mathematics, Springer, New York, 2008. Algorithms, analysis, and applications.
35. P. Houston, D. Schötzau, and T. P. Wihler, *Energy norm a posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems*, Math. Models Methods Appl. Sci., 17 (2007), pp. 33–62.
36. P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163 (electronic).
37. P. Houston and E. Süli, *Stabilised hp-finite element approximation of partial differential equations with nonnegative characteristic form*, Computing, 66 (2001), pp. 99–119. Archives for scientific computing. Numerical methods for transport-dominated and related problems (Magdeburg, 1999).
38. M. Jensen, *Discontinuous Galerkin methods for friedrichs systems*, D.Phil. Thesis, University of Oxford, (2005).
39. C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.
40. O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399 (electronic).
41. ———, *Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems*, SIAM J. Numer. Anal., 45 (2007), pp. 641–665 (electronic).
42. P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical aspects of finite elements in partial differential equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974), Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974, pp. 89–123. Publication No. 33.
43. J. A. Mackenzie and K. W. Morton, *Finite volume solutions of convection-diffusion test problems*, Math. Comp., 60 (1993), pp. 189–220.
44. J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Uni. Hamburg, 36 (1971), pp. 9–15.
45. O. A. Oleĭnik and E. V. Radkevič, *Second order equations with nonnegative characteristic form*, Plenum Press, New York, 1973. Translated from the Russian by Paul C. Fife.
46. W. H. Reed and T. R. Hill, *Triangular mesh methods for the neutron transport equation.*, Technical Report LA-UR-73-479 Los Alamos Scientific Laboratory, (1973).
47. M. Renardy and R. C. Rogers, *An introduction to partial differential equations*, vol. 13 of Texts in Applied Mathematics, Springer-Verlag, New York, 1993.

48. B. Rivière, *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, vol. 35 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and implementation.
49. B. Rivière, M. F. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. I*, Comput. Geosci., 3 (1999), pp. 337–360 (2000).
50. ———, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931 (electronic).
51. C. Schwab, *p- and hp- finite element methods: Theory and applications in solid and fluid mechanics*, Oxford University Press: Numerical mathematics and scientific computation, 1998.
52. R. M. Smith and A. G. Hutton, *The numericla treatment of convetion–a performance/comparison of current methods*, Numer. Heat Transfer, 5 (1982), pp. 439–461.
53. G. Strang and G. J. Fix, *An analysis of the finite element method*, Prentice-Hall Inc., Englewood Cliffs, N. J., 1973. Prentice-Hall Series in Automatic Computation.
54. M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.