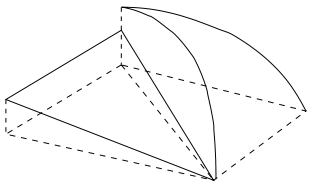

Chapter 2
Higher dimensional problem

This chapter deals with the formulation and analysis of the primal DG methods NIPG, SIPG, and IIPG in two and three dimensions for a general elliptic equation. The chapter also includes a brief description of the local discontinuous Galerkin (LDG) method that is based on a mixed formulation of the elliptic equation.

2.1 Preliminaries

2.1.1 Vector notation

The gradient of a scalar function $v : \mathbb{R}^d \to \mathbb{R}$ is a vector and the divergence of a vector function $\boldsymbol{w} : \mathbb{R}^d \to \mathbb{R}^d$ is a scalar:

$$\nabla v = \left(\frac{\partial v}{\partial x_i}\right)_{1 \le i \le d}, \quad \nabla \cdot \boldsymbol{w} = \sum_{i=1}^{d} \frac{\partial w_i}{\partial x_i}.$$

The dot product between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is

$$\boldsymbol{u} \cdot \boldsymbol{v} = \sum_{i=1}^{d} u_i v_i.$$

2.1.2 Sobolev spaces

Throughout the book, Ω denotes a bounded polygonal domain in $\mathbb{R}^d$. The vector space $L^2(\Omega)$ is the space of square-integrable functions:

$$L^2(\Omega) = \left\{ v \text{ measurable } : \int_\Omega v^2 < \infty \right\}.$$

Without going into too many details, we can say that the measure considered here is the Lebesgue measure and that the elements of $L^2(\Omega)$ are actually classes of functions: two functions v_1 and v_2 belong to the same class if and only if they differ on a set of measure

zero. We say that $v_1 = v_2$ almost everywhere (a.e. for short). The reader can refer to [97] for an introduction to Lebesgue measure.

Definition 2.1. *Let V be a vector space. A symmetric bilinear form $a : V \times V \to \mathbb{R}$ is an inner product if $a(v, v) \geq 0$ for all $v \in V$ and $a(v, v) = 0$ if and only if $v = 0$. The space V is a normed space for the norm $\|\cdot\|_V = (a(\cdot,\cdot))^{1/2}$. Furthermore, the space V equipped with an inner product is a Hilbert space if it is complete, i.e., if every Cauchy sequence is convergent. A sequence $(v_n)_n$ is said to be a Cauchy sequence if for all $\delta > 0$ there is a natural integer n_0 such that for all $n, m > n_0$, we have $\|v_n - v_m\|_V \leq \delta$. The dual space of V, denoted by V', is the space of continuous linear mappings from V to $\mathbb{R}$.*

The space $L^2(\Omega)$ is a Hilbert space with respect to the following inner product and norm:

$$(u, v)_\Omega = \int_\Omega uv, \quad \|v\|_{L^2(\Omega)} = \left(\int_\Omega v^2\right)^{1/2}.$$

We extend naturally these definitions to vector functions $\boldsymbol{u} = (u_i)_{1\leq i\leq d}$ and $\boldsymbol{v} = (v_i)_{1\leq i\leq d}$:

$$(\boldsymbol{u}, \boldsymbol{v})_\Omega = \int_\Omega \boldsymbol{u}\cdot\boldsymbol{v}, \quad \|\boldsymbol{v}\|_{L^2(\Omega)} = \left(\sum_{i=1}^d \|v_i\|^2_{L^2(\Omega)}\right)^{1/2}.$$

The space $L^\infty(\Omega)$ is the space of bounded functions:

$$L^\infty(\Omega) = \{v : \|v\|_{L^\infty(\Omega)} < \infty\}$$

with

$$\|v\|_{L^\infty(\Omega)} = \text{ess sup}\{|v(\boldsymbol{x})| : \boldsymbol{x} \in \Omega\}.$$

Definition 2.2. *The support of a continuous function v defined on $\mathbb{R}^d$ is the closure of the set of points at which the function is not equal to zero. If it is bounded and included in the interior of the domain Ω, then v is said to have compact support in Ω.*

Let $\mathcal{D}(\Omega)$ denote the space of $\mathcal{C}^\infty$ functions with compact support in Ω. The dual space $\mathcal{D}'(\Omega)$ is called the space of distributions. For any multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ and $|\alpha| = \sum_{i=1}^d \alpha_i$, the distributional derivative $D^\alpha v \in \mathcal{D}'(\Omega)$ is defined by

$$\forall \phi \in \mathcal{D}(\Omega), \quad D^\alpha v(\phi) = (-1)^{|\alpha|} \int_\Omega v(x) \frac{\partial^{|\alpha|}\phi}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}.$$

For instance, we have

$$\forall \phi \in \mathcal{D}(\Omega), \quad \frac{\partial v}{\partial x_1}(\phi) = -\int_\Omega v\frac{\partial \phi}{\partial x_1}.$$

We introduce the Sobolev space

$$H^1(\Omega) = \left\{v \in L^2(\Omega) : \frac{\partial v}{\partial x_i} \in L^2(\Omega), i = 1, \ldots, d\right\}.$$

It can be shown that if v belongs to $L^2(\Omega)$, then v can be identified with a distribution, still denoted by v, in the following sense:

$$\forall \phi \in \mathcal{D}(\Omega), \quad v(\phi) = \int_\Omega v\phi.$$

Therefore, for $v \in H^1(\Omega)$, we can write

$$\forall \phi \in \mathcal{D}(\Omega), \quad \frac{\partial v}{\partial x_i}(\phi) = \int_\Omega \frac{\partial v}{\partial x_i}\phi = -\int_\Omega v\frac{\partial \phi}{\partial x_i}.$$

We will write for short

$$H^1(\Omega) = \{v \in L^2(\Omega) : \nabla v \in (L^2(\Omega))^d\}.$$

Similarly, we introduce $H^s(\Omega)$ for integer s:

$$H^s(\Omega) = \{v \in L^2(\Omega) : \forall 0 \le |\alpha| \le s, D^\alpha v \in L^2(\Omega)\}.$$

In particular, in two dimensions, we have

$$H^2(\Omega) = \left\{ v \in H^1(\Omega) : \frac{\partial^2 v}{\partial x_1^2}, \frac{\partial^2 v}{\partial x_1 \partial x_2}, \frac{\partial^2 v}{\partial x_2^2} \in L^2(\Omega) \right\},$$

and we write for short

$$H^2(\Omega) = \{v \in L^2(\Omega) : \nabla^2 v \in (L^2(\Omega))^{d\times d}\}.$$

For $v \in H^s(\Omega)$, we can write for $|\alpha| \le s$:

$$\forall \phi \in \mathcal{D}(\Omega), \quad D^\alpha v(\phi) = \int_\Omega D^\alpha v\, \phi = (-1)^{|\alpha|} \int_\Omega v \frac{\partial^{|\alpha|}\phi}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

If v is smooth enough, we recover the usual derivatives:

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

The Sobolev norm associated with $H^s(\Omega)$ is

$$\|v\|_{H^s(\Omega)} = \left(\sum_{0\le|\alpha|\le s} \|D^\alpha v\|^2_{L^2(\Omega)} \right)^{1/2}.$$

The Sobolev seminorm associated with $H^s(\Omega)$ is

$$|v|_{H^s(\Omega)} = \|\nabla^s v\|_{L^2(\Omega)} = \left(\sum_{|\alpha|=s} \|D^\alpha v\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Let us now define the Sobolev spaces with fractional indices. The space $H^{s+1/2}(\Omega)$ with s integer is obtained by interpolating between the spaces $H^s(\Omega)$ and $H^{s+1}(\Omega)$. The

K-interpolation method [14] is used: Given $v \in H^s(\Omega)$, we define the following splitting:

$$v = v_1 + v_2,$$

where $v_1 \in H^s(\Omega)$ and $v_2 \in H^{s+1}(\Omega)$. Then, for a given real number t, we define the kernel

$$K(v,t) = \left(\inf_{v_1+v_2=v} (\|v_1\|^2_{H^s(\Omega)} + t^2 \|v_2\|^2_{H^{s+1}(\Omega)}) \right)^{1/2}.$$

Definition 2.3. *A space V equipped with the norm $||\cdot||_V$ is said to be the completion of a subset W if, for any element $v \in V$ and any $\delta > 0$, there exists $w \in W$ such that*

$$||v - w||_V \leq \delta.$$

The space $H^{s+1/2}(\Omega)$ is then defined as the completion of all functions in $H^{s+1}(\Omega)$ with respect to the following norm:

$$\|v\|_{H^{s+1/2}(\Omega)} = \left(\int_0^\infty t^{-2} K^2(v,t) dt \right)^{1/2}.$$

Then, we have the properties

$$H^{s+1}(\Omega) \subset H^{s+1/2}(\Omega) \subset H^s(\Omega),$$
$$\forall v \in H^{s+1}(\Omega), \quad \|v\|_{H^{s+1/2}(\Omega)} \leq C(\Omega) \|v\|^{1/2}_{H^s(\Omega)} \|v\|^{1/2}_{H^{s+1}(\Omega)},$$

where $C(\Omega)$ is a positive constant that depends on the domain Ω.

An important result is the imbedding theorem that relates the Sobolev spaces to the standard spaces of $\mathcal{C}^r(\Omega)$ functions.

Theorem 2.4. *For $\Omega \subset \mathbb{R}^d$, we have*

$$H^s(\Omega) \subset \mathcal{C}^r(\Omega) \quad \text{if} \quad \frac{1}{2} < \frac{s-r}{d}.$$

To be more precise, the theorem says that under certain conditions depending on s and d, if $v \in H^s(\Omega)$, then there is a continuous representative in the equivalence class of v. The conditions are given below:

$$H^s(\Omega) \subset \mathcal{C}^0(\Omega) \quad \text{if} \quad \begin{cases} s > \frac{1}{2} & \text{for} \quad d = 1, \\ s > 1 & \text{for} \quad d = 2, \\ s > \frac{3}{2} & \text{for} \quad d = 3. \end{cases}$$

2.1.3 Trace theorems

Using distributional derivatives, we can formulate partial differential equations in the distributional sense. The notion of traces [81] is used to define the restriction of a Sobolev function along the boundary of the domain. This is important for properly defining boundary conditions.

Theorem 2.5. *Let Ω be a bounded domain with polygonal boundary $\partial\Omega$ and outward normal vector $\boldsymbol{n}$. There exist trace operators $\gamma_0 : H^s(\Omega) \to H^{s-1/2}(\partial\Omega)$ for $s > 1/2$ and*

$\gamma_1 : H^s(\Omega) \to H^{s-3/2}(\partial\Omega)$ for $s > 3/2$ that are extensions of the boundary values and boundary normal derivatives, respectively. The operators γ_j are surjective. Furthermore, if $v \in \mathcal{C}^1(\bar{\Omega})$, then

$$\gamma_0 v = v|_{\partial\Omega}, \quad \gamma_1 v = \nabla v \cdot \boldsymbol{n}|_{\partial\Omega}.$$

As a consequence, if $v \in H^1(\Omega)$, then its trace $\gamma_0 v$ belongs to $H^{1/2}(\partial\Omega)$, the interpolated space between $L^2(\partial\Omega)$ and $H^1(\partial\Omega)$. In that case, $\gamma_1 v$ may not be defined.

The subspace of $H^s(\Omega)$, $s > 1/2$, consisting of functions whose traces vanish on the boundary is denoted by

$$H_0^s(\Omega) = \{v \in H^s(\Omega) : \gamma_0 v = 0 \text{ on } \partial\Omega\}.$$

We recall some important trace inequalities that are frequently used in the analysis of the DG methods. Let E be a bounded polygonal domain with diameter h_E:

$$h_E = \sup_{\boldsymbol{x},\boldsymbol{y}\in E} \|\boldsymbol{x} - \boldsymbol{y}\|,$$

where $\|\boldsymbol{x}\|$ is the Euclidean norm ($\|\boldsymbol{x}\| = (\boldsymbol{x} \cdot \boldsymbol{x})^{1/2}$). Let $|E|$ denote the length of E in one dimension (1D), the area of E in two dimensions (2D), and the volume of E in three dimensions (3D). Similarly, we will use the length or area $|e|$ for an edge or a face of E. Then, there is a constant C independent of h_E and v such that for any $v \in H^s(E)$

$$s \geq 1\, \forall e \subset \partial E,\ \|\gamma_0 v\|_{L^2(e)} \leq C|e|^{1/2}|E|^{-1/2}(\|v\|_{L^2(E)} + h_E\|\nabla v\|_{L^2(E)}), \tag{2.1}$$

$$s \geq 2\, \forall e \subset \partial E,\ \|\gamma_1 v\|_{L^2(e)} \leq C|e|^{1/2}|E|^{-1/2}(\|\nabla v\|_{L^2(E)} + h_E\|\nabla^2 v\|_{L^2(E)}). \tag{2.2}$$

In the rest of the text, we will abuse the notation and replace the traces $\gamma_0 v$ and $\gamma_1 v$ by v and $\nabla v \cdot \boldsymbol{n}$, respectively.

Note that if v is a polynomial, we can take advantage of equivalence of norms in finite-dimensional spaces. Denote by $\mathbb{P}_k(E)$ the space of polynomials of degree less than or equal to k:

$$\mathbb{P}_k(E) = \text{span}\{x_1^{i_1} x_2^{i_2} \cdots x_d^{i_d} : i_1 + i_2 + \cdots + i_d \leq k, \boldsymbol{x} \in E\}.$$

The trace inequalities now become

$$\forall v \in \mathbb{P}_k(E),\ \forall e \subset \partial E,\ \|v\|_{L^2(e)} \leq \tilde{C}_t|e|^{1/2}|E|^{-1/2}\|v\|_{L^2(E)}, \tag{2.3}$$

$$\forall v \in \mathbb{P}_k(E),\ \forall e \subset \partial E,\ \|v\|_{L^2(e)} \leq C_t h_E^{-1/2}\|v\|_{L^2(E)}, \tag{2.4}$$

$$\forall v \in \mathbb{P}_k(E),\ \forall e \subset \partial E,\ \|\nabla v \cdot \boldsymbol{n}\|_{L^2(e)} \leq \tilde{C}_t|e|^{1/2}|E|^{-1/2}\|\nabla v\|_{L^2(E)}, \tag{2.5}$$

$$\forall v \in \mathbb{P}_k(E),\ \forall e \subset \partial E,\ \|\nabla v \cdot \boldsymbol{n}\|_{L^2(e)} \leq C_t h_E^{-1/2}\|\nabla v\|_{L^2(E)}. \tag{2.6}$$

Here, the constants $\tilde{C}_t$, C_t are independent of h_E, v but depend on the polynomial degree k. In the case where E is an interval, a triangle, or a tetrahedron, one can obtain an exact expression for the constant C_t as a function of the polynomial degree [108]:

$$d = 1 \quad \forall v \in \mathbb{P}_k(E), \quad \forall t \in \partial E, \quad |v(t)| \leq \frac{k+1}{\sqrt{|E|}}\|v\|_{L^2(E)}, \tag{2.7}$$

$$d = 2 \quad \forall v \in \mathbb{P}_k(E), \qquad \|v\|_{L^2(e)} \leq \sqrt{\frac{(k+1)(k+2)}{2}\frac{|e|}{|E|}}\|v\|_{L^2(E)}, \tag{2.8}$$

$$d = 3 \quad \forall v \in \mathbb{P}_k(E), \qquad \|v\|_{L^2(e)} \leq \sqrt{\frac{(k+1)(k+3)}{3}\frac{|e|}{|E|}}\|v\|_{L^2(E)}. \tag{2.9}$$

2.1.4 Approximation properties

In this section, we state approximation results in the space of polynomials of degree k (see [9, 96]).

Theorem 2.6. *Let E be a triangle or parallelogram in 2D or a tetrahedron or hexahedron in 3D. Let $v \in H^s(E)$ for $s \geq 1$. Let $k \geq 0$ be an integer. There exist a constant C independent of v and h_E and a function $\tilde{v} \in \mathbb{P}_k(E)$ such that*

$$\forall 0 \leq q \leq s, \quad \|v - \tilde{v}\|_{H^q(E)} \leq C h_E^{\min(k+1,s)-q}|v|_{H^s(E)}. \tag{2.10}$$

As a consequence, if Ω is subdivided into triangles or tetrahedra, one can construct a global approximation $\tilde{v}$ that is continuous over the domain Ω and satisfies the same approximation result (2.10). If Ω is subdivided into parallelograms or hexahedra, the same result holds if the space $\mathbb{P}_k(E)$ is replaced by the space $\mathbb{Q}_k(E)$, namely the space of polynomials of degree less than or equal to k in each space direction.

The next result yields an approximation that conserves the average of the normal flux on each edge.

Theorem 2.7. *Let E be a triangle or parallelogram in 2D or a tetrahedron in 3D. Denote by $\boldsymbol{n}_E$ the outward normal to E. Let $v \in H^s(E)$ for $s \geq 2$. Let $\boldsymbol{K}$ be a symmetric positive definite matrix with constant entries. There exists an approximation $\tilde{v} \in \mathbb{P}_k(E)$ of v satisfying*

$$\int_e \boldsymbol{K}\nabla(\tilde{v} - v) \cdot \boldsymbol{n}_E = 0 \quad \forall e \in \partial E$$

and the optimal error bounds

$$\forall i = 0, 1, 2, \quad \|\nabla^i(\tilde{v} - v)\|_{L^2(E)} \leq C h_E^{\min(k+1,s)-i}|v|_{H^s(E)}, \tag{2.11}$$

where C is independent of h_E.

If the matrix $\boldsymbol{K}$ is a function of space, the previous result is still valid for small enough h_E. The proof of this theorem for a triangle or a tetrahedron is given in Appendix C.

2.1.5 Green's theorem

Given E a bounded domain and $\boldsymbol{n}_E$ the outward normal vector to ∂E, we have for all $v \in H^2(E)$ and $w \in H^1(E)$

$$-\int_E w\Delta v = \int_E \nabla v \cdot \nabla w - \int_{\partial E} \nabla v \cdot \boldsymbol{n}_E w, \tag{2.12}$$

where $\Delta w = \nabla \cdot \nabla w = \sum_{i=1}^{d} \frac{\partial^2 w}{\partial x_i^2}$. A more generalized Green's theorem is

$$-\int_E w \nabla \cdot \boldsymbol{F} \nabla v = \int_E \boldsymbol{F} \nabla v \cdot \nabla w - \int_{\partial E} \boldsymbol{F} \nabla v \cdot \boldsymbol{n}_E w, \tag{2.13}$$

where $\boldsymbol{F}$ is a matrix-valued function.

2.1.6 Cauchy–Schwarz's and Young's inequalities

The following two inequalities are used at several places in this text.

Cauchy–Schwarz's inequality:

$$\forall f, g \in L^2(\Omega), \quad |(f, g)_\Omega| \leq \|f\|_{L^2(\Omega)} \|g\|_{L^2(\Omega)}. \tag{2.14}$$

Young's inequality:

$$\forall \epsilon > 0, \quad \forall a, b \in \mathbb{R}, \quad ab \leq \frac{\epsilon}{2} a^2 + \frac{1}{2\epsilon} b^2. \tag{2.15}$$

2.2 Model problem

Let Ω be a polygonal domain in $\mathbb{R}^d, d = 2$ or 3. The sides of the boundary $\partial\Omega$ of the domain are grouped into two disjoint sets Γ_{D} and Γ_{N}. Let $\boldsymbol{n}$ be the unit normal vector to the boundary exterior to Ω. For f given in $L^2(\Omega)$, g_{D} given in $H^{\frac{1}{2}}(\Gamma_{\mathrm{D}})$, and g_{N} given in $L^2(\Gamma_{\mathrm{N}})$, we consider the following elliptic problem:

$$-\nabla \cdot (\boldsymbol{K} \nabla p) + \alpha p = f \quad \text{in} \quad \Omega, \tag{2.16}$$

$$p = g_{\mathrm{D}} \quad \text{on} \quad \Gamma_{\mathrm{D}}, \tag{2.17}$$

$$\boldsymbol{K} \nabla p \cdot \boldsymbol{n} = g_{\mathrm{N}} \quad \text{on} \quad \Gamma_{\mathrm{N}}. \tag{2.18}$$

The coefficient $\boldsymbol{K}$ is a matrix-valued function $\boldsymbol{K} = (k_{ij})_{1 \leq i,j \leq d}$ that is symmetric ($k_{ij} = k_{ji}$) positive definite and bounded below and above uniformly; i.e., there exist two positive constants K_0 and K_1 such that

$$\forall \boldsymbol{x} \in \mathbb{R}^d, \quad K_0 \boldsymbol{x} \cdot \boldsymbol{x} \leq \boldsymbol{K} \boldsymbol{x} \cdot \boldsymbol{x} \leq K_1 \boldsymbol{x} \cdot \boldsymbol{x}. \tag{2.19}$$

The other coefficient α is a nonnegative scalar function. The second equation (2.17) is called a Dirichlet boundary condition. The value of the solution is prescribed on Γ_{D}. The third equation (2.18) is called a Neumann boundary condition. The normal derivative or flux is prescribed on Γ_{N}.

The problem (2.16)–(2.18) has a solution $p \in \mathcal{C}^2(\bar{\Omega})$, called strong solution under additional smoothness on the data f, g_{D}, g_{N}, $\boldsymbol{K}$, and α. The equations are then satisfied pointwisely. With the definition of weak derivatives, we can rewrite the partial differential equation into a weak form and define a weak solution.

2.2.1 Weak solution

For simplicity, assume that $\partial\Omega = \Gamma_{\mathrm{D}}$. From the trace theorem, there is an extension of $g_{\mathrm{D}} \in H^{1/2}(\partial\Omega)$ in Ω. Let $p_{\mathrm{D}} \in H^1(\Omega)$ be the extension:

$$p_{\mathrm{D}} = g_{\mathrm{D}} \quad \text{on} \quad \partial\Omega.$$

The variational formulation (or weak formulation) of problem (2.16)–(2.18) is as follows: Find $p = p_{\mathrm{D}} + w$ with $w \in H_0^1(\Omega)$ such that

$$\forall v \in H_0^1(\Omega), \quad \int_\Omega (\boldsymbol{K}\nabla w \cdot \nabla v + \alpha w v) = \int_\Omega f v - \int_\Omega (\boldsymbol{K}\nabla p_{\mathrm{D}} \cdot \nabla v + \alpha p_{\mathrm{D}} v). \tag{2.20}$$

The solution p is called the weak solution to problem (2.16)–(2.18). Existence and uniqueness of w is a consequence of the Lax–Milgram theorem given below [79].

Theorem 2.8. *Let V be a real Hilbert space. Let $a : V \times V \to \mathbb{R}$ be a bilinear form that is*
(i) *continuous: $|a(u, v)| \le C_1 \|u\|_V \|v\|_V$,*
(ii) *coercive: $C_2 \|u\|_V^2 \le a(u, u)$, with positive constants C_1 and C_2.*
Let $L : V \to \mathbb{R}$ be a continuous linear functional. Then, there exists a unique $u \in V$ satisfying

$$\forall v \in V, \quad a(u, v) = L(v).$$

Moreover, the solution u is bounded by the data

$$\|u\|_V \le \frac{1}{C_2} \|L\|.$$

If $\partial\Omega = \Gamma_{\mathrm{N}}$ and $\alpha = 0$, the weak solution is unique up to an additive constant, provided the compatibility condition $\int_\Omega f + \int_{\partial\Omega} g_{\mathrm{N}} = 0$ is satisfied. Indeed, this condition is obtained by integrating (2.16) over Ω and by using Green's theorem.

2.2.2 Numerical solution

There are several methods available for solving problem (2.16)–(2.18). We mention here two basic ones: finite difference method and finite element method.

The finite difference method approximates the partial derivatives by finite differences. Let the domain be subdivided into uniform squares with vertices $A_{ij}(x_i, y_j)$ for $1 \le i, j \le M$. This grid is characterized by the length of the side of a square denoted by h (see Fig. 2.1). We have

$$\frac{\partial^2 p}{\partial x^2}(A_{ij}) \approx \frac{p(x_{i-1}, y_j) - 2p(x_{i,j}) + p(x_{i+1}, y_j)}{h^2},$$
$$\frac{\partial^2 p}{\partial y^2}(A_{ij}) \approx \frac{p(x_i, y_{j-1}) - 2p(x_{i,j}) + p(x_i, y_{j+1})}{h^2}.$$

The finite difference solution is a set of values P_{ij} approximating $p(x_i, y_j)$. For instance, the finite difference method applied to the Poisson equation $-\Delta p = f$ is

$$\forall i, j, \quad -\frac{P_{i-1,j} - 2P_{ij} + P_{i+1,j}}{h^2} - \frac{P_{i,j-1} - 2P_{i,j} + P_{i,j+1}}{h^2} = f(x_i, y_j).$$

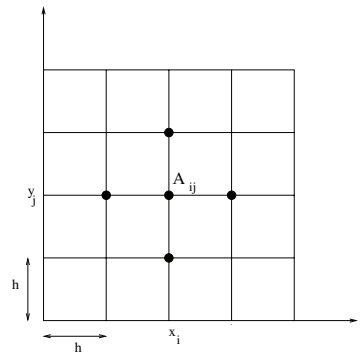

Figure 2.1. *Finite difference grid.*

After taking account of the boundary conditions, we obtain a linear system with unknowns P_{ij}. This method is easy to implement. However, the accuracy is limited, as the method is of low order and the method is not well suited to complicated geometries.

The finite element method uses the variational formulation of the partial differential equation. Let Ω be partitioned into elements (for instance triangles or rectangles in 2D) that form a mesh. Let X_h be the finite-dimensional subspace of $H_0^1(\Omega)$, consisting of continuous piecewise polynomials of degree k on each element. Based on (2.20), the finite element method is to find $P_h = \tilde{p}_{\mathrm{D}} + W_h$ with $W_h \in X_h$ such that

$$\forall v \in X_h, \quad \int_\Omega (\boldsymbol{K}\nabla W_h \cdot \nabla v + \alpha W_h v) = \int_\Omega f v - \int_\Omega (\boldsymbol{K}\nabla \tilde{p}_{\mathrm{D}} \cdot \nabla v + \alpha \widetilde{p_{\mathrm{D}}} v). \tag{2.21}$$

The function $\tilde{p}_{\mathrm{D}} \in X_h$ is an interpolant of the extension p_{D}. Finite element methods were first introduced by engineers in the 1950s. The mathematical theory was developed in the late 1960s for steady-state problems. We refer the reader to [28, 17] for a general treatment of the theory. Compared to the finite difference methods, finite element methods offer several attractive features: their accuracy depends on the polynomial degree k; they can handle complicated geometries by the use of unstructured grids. However, these methods are not locally mass conservative (see Section 2.7.3), which means that in nonlinear reactive transport problems, finite difference methods still prevailed. Another issue is the rather complicated use of local mesh refinement.

DG methods also use a variational formulation of the problem. In that sense, DG and finite element methods share many properties, and we can abuse the terminology by saying that the DG method is a particular type of finite element method. In addition to the high order of accuracy and the use of unstructured meshes, DG methods are locally mass conservative, and they easily handle local mesh refinement. A more detailed comparison of the finite element method with DG is given in Section 2.12.

2.3 Broken Sobolev spaces

Broken Sobolev spaces are natural spaces to work with the DG methods. These spaces depend strongly on the partition of the domain. Let Ω be a polygonal domain subdivided into

elements E, where E is a triangle or a quadrilateral in 2D, or a tetrahedron or hexahedron in 3D. For simplicity, we assume that the intersection of two elements is either empty, a vertex, an edge, or a face. Such a mesh is called a conforming mesh. The resulting subdivision (or mesh) is denoted by $\mathcal{E}_h$, and h is the maximum element diameter. We also assume that the subdivision is regular [28]. This means that if h_E denotes the diameter of E and ρ_E denotes the maximum diameter of a ball inscribed in E, there is a constant $\rho > 0$ such that

$$\forall E \in \mathcal{E}_h, \quad \frac{h_E}{\rho_E} \le \rho.$$

We introduce the broken Sobolev space for any real number s,

$$H^s(\mathcal{E}_h) = \{v \in L^2(\Omega) : \forall E \in \mathcal{E}_h, v|_E \in H^s(E)\},$$

equipped with the broken Sobolev norm:

$$|||v|||_{H^s(\mathcal{E}_h)} = \left(\sum_{E \in \mathcal{E}_h} \|v\|^2_{H^s(E)} \right)^{1/2}.$$

In particular, we will use the broken gradient seminorm:

$$|||\nabla v|||_{H^0(\mathcal{E}_h)} = \left(\sum_{E \in \mathcal{E}_h} \|\nabla v\|^2_{L^2(E)} \right)^{1/2}.$$

Clearly, we have

$$H^s(\Omega) \subset H^s(\mathcal{E}_h) \quad \text{and} \quad H^{s+1}(\mathcal{E}_h) \subset H^s(\mathcal{E}_h).$$

In Sections 3.1.4, 5.1.2, and 7.1.1, the classical Poincaré inequality, Korn's inequality, and a Sobolev imbedding are generalized for the broken Sobolev space.

2.3.1 Jumps and averages

We denote by Γ_h the set of interior edges (or faces) of the subdivision $\mathcal{E}_h$. With each edge (or face) e, we associate a unit normal vector $\boldsymbol{n}_e$. If e is on the boundary $\partial\Omega$, then $\boldsymbol{n}_e$ is taken to be the unit outward vector normal to $\partial\Omega$.

If v belongs to $H^1(\mathcal{E}_h)$, the trace of v along any side of one element E is well defined. If two elements E_1^e and E_2^e are neighbors and share one common side e, there are two traces of v along e. We can add or subtract those values, and we obtain an average and a jump for v. We assume that the normal vector $\boldsymbol{n}_e$ is oriented from E_1^e to E_2^e:

$$\{v\} = \frac{1}{2}(v|_{E_1^e}) + \frac{1}{2}(v|_{E_2^e}), \quad [v] = (v|_{E_1^e}) - (w|_{E_2^e}) \quad \forall e = \partial E_1^e \cap \partial E_2^e.$$

As in the one-dimensional case, by convention, we extend the definition of jump and average to sides that belong to the boundary $\partial\Omega$:

$$\{v\} = [v] = (v|_{E_1^e}) \quad \forall e = \partial E_1^e \cap \partial\Omega.$$

2.4 Variational formulation

In what follows, we assume that $s > 3/2$. We introduce two bilinear forms $J_0^{\sigma_0,\beta_0}$, $J_1^{\sigma_1,\beta_1}$: $H^s(\mathcal{E}_h) \times H^s(\mathcal{E}_h) \to \mathbb{R}$ that penalize the jump of the function values and the jump of the normal derivatives values:

$$J_0^{\sigma_0,\beta_0}(v, w) = \sum_{e\in\Gamma_h\cup\Gamma_D} \frac{\sigma_e^0}{|e|^{\beta_0}} \int_e [v][w],$$

$$J_1^{\sigma_1,\beta_1}(v, w) = \sum_{e\in\Gamma_h} \frac{\sigma_e^1}{|e|^{\beta_1}} \int_e [\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e][\boldsymbol{K}\nabla w \cdot \boldsymbol{n}_e].$$

The parameters σ_e^0 and σ_e^1 are called penalty parameters. They are nonnegative real numbers. The powers β_0 and β_1 are positive numbers that depend on the dimension d. All parameters will be specified later. We recall that the notation $|e|$ simply means the length of e in 2D and the area of e in 3D. We clearly have

$$\forall e \subset \partial E, \quad |e| \le h_E^{d-1} \le h^{d-1}. \tag{2.22}$$

We now define the DG bilinear forms $a_\epsilon : H^s(\mathcal{E}_h) \times H^s(\mathcal{E}_h) \to \mathbb{R}$:

$$\begin{aligned} a_\epsilon(v, w) = &\sum_{E\in\mathcal{E}_h} \int_E \boldsymbol{K}\nabla v \cdot \nabla w + \int_\Omega \alpha v w \\ &- \sum_{e\in\Gamma_h\cup\Gamma_D} \int_e \{\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e\}[w] + \epsilon \sum_{e\in\Gamma_h\cup\Gamma_D} \int_e \{\boldsymbol{K}\nabla w \cdot \boldsymbol{n}_e\}[v] \\ &+ J_0^{\sigma_0,\beta_0}(v, w) + J_1^{\sigma_1,\beta_1}(v, w). \end{aligned} \tag{2.23}$$

The bilinear form a_ϵ contains another parameter ϵ that may take the value -1, 0, or 1. As in Section 1.2, we have the following symmetry property: a_ϵ is symmetric if $\epsilon = -1$ and it is nonsymmetric otherwise.

We also define the following linear form:

$$L(v) = \int_\Omega f v + \epsilon \sum_{e\in\Gamma_D} \int_e \left(\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e + \frac{\sigma_e^0}{|e|^{\beta_0}} v \right) g_D + \sum_{e\in\Gamma_N} \int_e v g_N.$$

Cauchy–Schwarz's inequality and trace inequalities imply that all integral terms in the forms defined above make sense if the functions belong to $H^s(\mathcal{E}_h)$ for any $s > 3/2$.

The general DG variational formulation of problem (2.16)–(2.18) is as follows: Find p in $H^s(\mathcal{E}_h)$, $s > 3/2$, such that

$$\forall v \in H^s(\mathcal{E}_h), \quad a_\epsilon(p, v) = L(v). \tag{2.24}$$

Remark: We note that the problem (2.24) is independent of the choice of the normal $\boldsymbol{n}_e$. Indeed, let e be one edge (or face) shared by two elements E_i and E_j. Let $\boldsymbol{n}_{ij}$ be the unit normal vector pointing from E_i to E_j. If $\boldsymbol{n}_e$ coincides with $\boldsymbol{n}_{ij}$, we have

$$\{\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e\}[w] = \{\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_{ij}\}(w|_{E_i} - w|_{E_j}).$$

If $\boldsymbol{n}_e$ has the opposite direction to $\boldsymbol{n}_{ij}$, the jump $[w]$ has a different sign and

$$\{\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e\}[w] = \{\boldsymbol{K}\nabla v \cdot (-\boldsymbol{n}_{ij})\}(w|_{E_j} - w|_{E_i}),$$

which gives the same expression as above.

2.4.1 Consistency

The next proposition establishes the equivalence between the model problem and the variational formulation.

Proposition 2.9. *Let $s > 3/2$. Assume that the weak solution p of problem* (2.16)–(2.18) *belongs to $H^s(\mathcal{E}_h)$; then p satisfies the variational problem* (2.24)*. Conversely, if $p \in H^1(\Omega) \cap H^s(\mathcal{E}_h)$ satisfies* (2.24)*, then p is the solution of problem* (2.16)–(2.18)*.*

Proof. First, we prove that if the solution p of (2.16)–(2.18) belongs to $H^s(\Omega)$, then it also solves (2.24). For this, let v be an element in $H^s(\mathcal{E}_h)$. We multiply (2.16) by v, integrate on one element E, and use Green's theorem (2.13):

$$\int_E (\boldsymbol{K}\nabla p \cdot \nabla v + \alpha p v) - \int_{\partial E} \boldsymbol{K}\nabla p \cdot \boldsymbol{n}_E v = \int_E f v.$$

We recall that $\boldsymbol{n}_E$ is the outward normal to E. We sum over all elements, switch to the normal vectors $\boldsymbol{n}_e$, and observe that

$$\sum_{E\in\mathcal{E}_h} \int_{\partial E} \boldsymbol{K}\nabla p \cdot \boldsymbol{n}_E v = \sum_{e\in\Gamma_h} \int_e [\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e v] + \sum_{e\in\partial\Omega} \int_e \boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e v. \tag{2.25}$$

By regularity of the solution p, we have

$$\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e = \{\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e\} \quad \text{a.e.}$$

Therefore, we obtain the resulting equation

$$\sum_{E\in\mathcal{E}_h} \int_E (\boldsymbol{K}\nabla p \cdot \nabla v + \alpha p v) - \sum_{e\in\Gamma_h} \int_e \{\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e\}[v] - \int_{\partial\Omega} (\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e) v = \int_\Omega f v.$$

Using the Neumann boundary condition (2.18), we get

$$\begin{aligned}\sum_{E\in\mathcal{E}_h} \int_E (\boldsymbol{K}\nabla p \cdot \nabla v + \alpha p v) &- \sum_{e\in\Gamma_h} \int_e \{\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e\}[v] \\ &- \sum_{e\in\Gamma_D} \int_e (\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e) v = \int_\Omega f v + \sum_{e\in\Gamma_N} \int_e g_N v.\end{aligned}$$

We add $\epsilon \sum_{e\in\Gamma_D} \int_e (\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e) p$ and $\sum_{e\in\Gamma_D} \frac{\sigma_e^0}{|e|^{\beta_0}} \int_e p v$ to both sides and use the Dirichlet boundary condition (2.17):

$$\sum_{E\in\mathcal{E}_h} \int_E (\boldsymbol{K}\nabla p \cdot \nabla v + \alpha p v) - \sum_{e\in\Gamma_h} \int_e \{\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e\}[v] - \sum_{e\in\Gamma_D} \int_e (\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e) v$$

$$+\epsilon \sum_{e\in\Gamma_{\mathrm{D}}} \int_e (\boldsymbol{K}\nabla v\cdot \boldsymbol{n}_e)p + \sum_{e\in\Gamma_{\mathrm{D}}} \frac{\sigma_e^0}{|e|^{\beta_0}} \int_e pv = \int_\Omega fv + \epsilon \sum_{e\in\Gamma_{\mathrm{D}}} \int_e (\boldsymbol{K}\nabla v\cdot \boldsymbol{n}_e) g_{\mathrm{D}}$$
$$+ \sum_{e\in\Gamma_{\mathrm{N}}} \int_e g_{\mathrm{N}} v + \sum_{e\in\Gamma_{\mathrm{D}}} \frac{\sigma_e^0}{|e|^{\beta_0}} \int_e g_{\mathrm{D}} v.$$

Finally, we note that the jumps $[p] = [\boldsymbol{K}\nabla p\cdot \boldsymbol{n}_e]$ are zero a.e. on the interior edges (or faces). Then, we clearly have (2.24).

Conversely, take $v \in \mathcal{D}(E)$. Then (2.24) reduces to

$$\sum_{E\in\mathcal{E}_h} \int_E \boldsymbol{K}\nabla p\cdot\nabla v + \int_\Omega \alpha p v = \int_\Omega f v,$$

which immediately yields in the distributional sense, for all $E \in \mathcal{E}_h$,

$$-\nabla\cdot \boldsymbol{K}\nabla p + \alpha p = f \quad \text{in} \quad E. \tag{2.26}$$

Next, let e be an interior edge (or face) and let E_e^1 and E_e^2 be the two elements adjacent to e. Take $v \in H_0^2(E_e^1 \cup E_e^2)$ and extend it by zero over the rest of the domain. On one hand, if we multiply (2.26) by v and use Green's theorem (2.13), we have

$$\int_{E_e^1\cup E_e^2} \boldsymbol{K}\nabla p\cdot\nabla v + \int_{E_e^1\cup E_e^2} \alpha p v - \int_e [\boldsymbol{K}\nabla p\cdot \boldsymbol{n}_e] v = \int_{E_e^1\cup E_e^2} f v. \tag{2.27}$$

On the other hand, since $[v] = 0$, (2.24) reduces to

$$\int_{E_e^1\cup E_e^2} \boldsymbol{K}\nabla p \nabla v + \int_{E_e^1\cup E_e^2} \alpha p v = \int_{E_e^1\cup E_e^2} f v.$$

Hence, we have

$$\forall v \in H_0^2(E_e^1 \cup E_e^2), \qquad \int_e [\boldsymbol{K}\nabla p\cdot \boldsymbol{n}_e] v = 0.$$

This implies that $[\boldsymbol{K}\nabla p\cdot \boldsymbol{n}_e]|_e = 0$ in $L^2(e)$. Since this holds for all e, it implies that $\nabla\cdot \boldsymbol{K}\nabla p \in L^2(\Omega)$, and hence we have globally

$$-\nabla\cdot \boldsymbol{K}\nabla p + \alpha p = f \quad \text{in} \quad \Omega. \tag{2.28}$$

To recover the Dirichlet boundary conditions, we multiply (2.28) by a function v in $H^2(\Omega)\cap H_0^1(\Omega)$, apply Green's theorem (2.13), and compare with (2.24):

$$-\sum_{e\in\Gamma_{\mathrm{D}}} \int_e (\boldsymbol{K}\nabla v\cdot \boldsymbol{n}_e)(p - g_{\mathrm{D}}) = 0.$$

This being true for all $v \in H^2(\Omega) \cap H_0^1(\Omega)$, we have $p = g_{\mathrm{D}}$ on Γ_{D}. Finally, choosing $v \in H^2(\Omega)$, $v|_{\Gamma_{\mathrm{D}}} = 0$, we find

$$-\sum_{e\in\Gamma_{\mathrm{N}}} \int_e (\boldsymbol{K}\nabla p\cdot \boldsymbol{n}_e) v = -\sum_{e\in\Gamma_{\mathrm{N}}} \int_e g v,$$

and this gives the other boundary condition. We clearly have (2.18). □

2.5 Finite element spaces

We will consider finite-dimensional subspaces of the broken Sobolev space $H^s(\mathcal{E}_h)$ for $s > 3/2$. Let k be a positive integer. The finite element subspace is taken to be

$$\mathcal{D}_k(\mathcal{E}_h) = \{v \in L^2(\Omega) : \forall E \in \mathcal{E}_h, v|_E \in \mathbb{P}_k(E)\}, \tag{2.29}$$

where $\mathbb{P}_k(E)$ denotes the space of polynomials of total degree less than or equal to k. We will refer to the functions in $\mathcal{D}_k(\mathcal{E}_h)$ as test functions. We note that the test functions are discontinuous along the edges (or faces) of the mesh.

As is done in the classical finite element method, each mesh element E (also called physical element) is mapped to a reference element $\hat{E}$, and all computations are done on the reference element. The following section introduces triangular, quadrilateral, and tetrahedral reference elements.

2.5.1 Reference elements versus physical elements

When implementing the DG method, one has to compute integrals over volumes (such as triangles or quadrilaterals in 2D, tetrahedra or hexahedra in 3D) and faces (such as edges in 2D, triangles or quadrilaterals in 3D). It would be too costly to compute the integrals over each physical element in the mesh. A more economical and effective approach is to use a change of variables to obtain an integral on a fixed element, called the reference element [28, 101].

Reference triangular element: It consists of a triangle $\hat{E}$ with vertices $\hat{A}_1(0, 0)$, $\hat{A}_2(1, 0)$, and $\hat{A}_3(0, 1)$ (see Fig. 2.2). For a given physical element E, there is an affine map F_E from the reference element onto E. If E has vertices $A_i(x_i, y_i)$ for $i = 1, 2, 3$, then the map F_E is defined by

$$F_E\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad x = \sum_{i=1}^{3} x_i \hat{\phi}_i(\hat{x}, \hat{y}), \quad y = \sum_{i=1}^{3} y_i \hat{\phi}_i(\hat{x}, \hat{y}),$$

where

$$\begin{aligned} \hat{\phi}_1(\hat{x}, \hat{y}) &= 1 - \hat{x} - \hat{y}, \\ \hat{\phi}_2(\hat{x}, \hat{y}) &= \hat{x}, \\ \hat{\phi}_3(\hat{x}, \hat{y}) &= \hat{y}. \end{aligned}$$

We can rewrite the mapping

$$\begin{pmatrix} x \\ y \end{pmatrix} = F_E\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \boldsymbol{B}_E \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + \mathbf{b}_E, \tag{2.30}$$

where $\boldsymbol{B}_E$ is a 2×2 matrix and $\mathbf{b}_E$ a vector. It is easy to show that

$$\boldsymbol{B}_E = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \quad \mathbf{b}_E = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

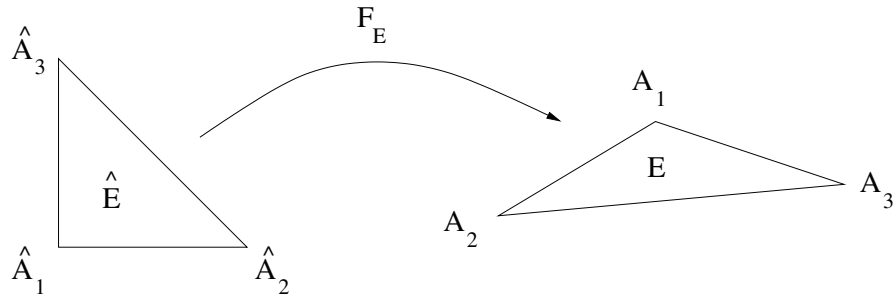

Figure 2.2. *Reference triangular element $\hat{E}$ and physical element E.*

The determinant of $\boldsymbol{B}_E$ appears in the computation of the integrals. If $|E|$ denotes the area of E, then we have

$$\det(\boldsymbol{B}_E) = 2|E|. \tag{2.31}$$

Thus $\boldsymbol{B}_E$ is invertible and the matrix norm (induced by the Euclidean norm) of $\boldsymbol{B}_E$ and $\boldsymbol{B}_E^{-1}$ is bounded as follows:

$$\|\boldsymbol{B}_E\| \equiv \sup_{(\hat{x},\hat{y})\in\hat{E}} \frac{\|\boldsymbol{B}_E\binom{\hat{x}}{\hat{y}}\|}{\|\binom{\hat{x}}{\hat{y}}\|} \le \frac{h_E}{\hat{\rho}}, \quad \|\boldsymbol{B}_E^{-1}\| \le \frac{\hat{h}}{\rho_E}.$$

Here, $\hat{h}$ denotes the diameter of $\hat{E}$ and $\hat{\rho}$ denotes the diameter of the largest circle inscribed in $\hat{E}$. Similarly, ρ_E denotes the diameter of the largest circle inscribed in E.

The mapping F_E corresponds to a change of variable. We denote

$$\hat{v} = v \circ F_E.$$

In other words, $\hat{v}(\hat{x}, \hat{y}) = v(x, y)$. We also denote by $\hat{\nabla}\hat{v}$ the gradient of $\hat{v}$ with respect to $\hat{x}$ and $\hat{y}$:

$$\hat{\nabla}\hat{v} = \begin{pmatrix} \frac{\partial \hat{v}}{\partial \hat{x}} \\ \frac{\partial \hat{v}}{\partial \hat{y}} \end{pmatrix}.$$

We can prove that

$$\hat{\nabla}\hat{v} = \boldsymbol{B}_E^T \nabla v \circ F_E, \tag{2.32}$$

where $\boldsymbol{B}_E^T$ is the transpose of the matrix $\boldsymbol{B}_E$ (i.e., $(\boldsymbol{B}_E^T)_{ij} = (\boldsymbol{B}_E)_{ji}$).

Reference quadrilateral element: It consists of the square $\hat{E}$ with vertices $\hat{A}_1(-1, -1)$, $\hat{A}_2(1, -1)$, $\hat{A}_3(1, 1)$, and $\hat{A}_4(-1, 1)$ (see Fig. 2.3). If E has vertices $A_i(x_i, y_i)$ for $i = 1, \ldots, 4$, the transformation map $F_E : \hat{E} \to E$ is defined by

$$F_E\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad x = \sum_{i=1}^{4} x_i \hat{\phi}_i(\hat{x}, \hat{y}), \quad y = \sum_{i=1}^{4} y_i \hat{\phi}_i(\hat{x}, \hat{y}), \tag{2.33}$$

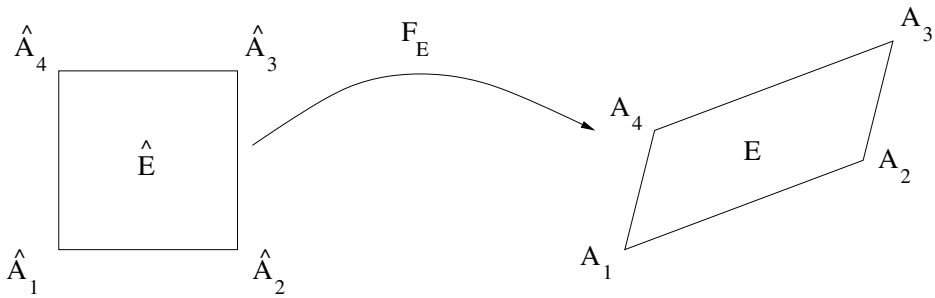

Figure 2.3. *Reference quadrilateral element $\hat{E}$ and physical element E.*

where

$$\hat{\phi}_1(\hat{x}, \hat{y}) = \frac{1}{4}(1 - \hat{x})(1 - \hat{y}),$$
$$\hat{\phi}_2(\hat{x}, \hat{y}) = \frac{1}{4}(1 + \hat{x})(1 - \hat{y}),$$
$$\hat{\phi}_3(\hat{x}, \hat{y}) = \frac{1}{4}(1 + \hat{x})(1 + \hat{y}),$$
$$\hat{\phi}_4(\hat{x}, \hat{y}) = \frac{1}{4}(1 - \hat{x})(1 + \hat{y}).$$

The mapping F_E is affine if the physical element E is a parallelogram. In the general case, we define $\boldsymbol{B}_E$ to be the Jacobian matrix of F_E:

$$\boldsymbol{B}_E = \begin{pmatrix} \frac{\partial x}{\partial \hat{x}} & \frac{\partial x}{\partial \hat{y}} \\ \frac{\partial y}{\partial \hat{x}} & \frac{\partial y}{\partial \hat{y}} \end{pmatrix}.$$

It is sufficient to have the determinant of $\boldsymbol{B}_E$ nonvanishing in order to have an invertible map F_E. This condition is satisfied if E is convex.

Reference tetrahedral element: It consists of the tetrahedron $\hat{E}$ with vertices $\hat{A}_1(0, 0, 0)$, $\hat{A}_2(1, 0, 0)$, $\hat{A}_3(0, 1, 0)$, and $\hat{A}_4(0, 0, 1)$. There is an affine map $F_E : \hat{E} \to E$, defined from the coordinates of the vertices $A_i(x_i, y_i)$:

$$F_E \begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

$$x = \sum_{i=1}^{4} x_i \hat{\phi}_i(\hat{x}, \hat{y}, \hat{z}), \quad y = \sum_{i=1}^{4} y_i \hat{\phi}_i(\hat{x}, \hat{y}, \hat{z}), \quad z = \sum_{i=1}^{4} z_i \hat{\phi}_i(\hat{x}, \hat{y}, \hat{z}),$$

where

$$\hat{\phi}_1(\hat{x}, \hat{y}, \hat{z}) = 1 - \hat{x} - \hat{y} - \hat{z},$$
$$\hat{\phi}_2(\hat{x}, \hat{y}, \hat{z}) = \hat{x},$$

$$\hat{\phi}_3(\hat{x}, \hat{y}, \hat{z}) = \hat{y},$$
$$\hat{\phi}_4(\hat{x}, \hat{y}, \hat{z}) = \hat{z}.$$

All properties for the reference triangle are valid for the reference tetrahedron.

Remark on choice of finite element spaces: We recall that the DG finite element space $\mathcal{D}_k(\mathcal{E}_h)$ is the space of discontinuous polynomials defined on the physical elements and not on the reference element. In practice, in the case of triangles, parallelograms in 2D, and tetrahedra or parallelepipeds, we could and *should* choose instead

$$\tilde{\mathcal{D}}_k(\mathcal{E}_h) = \{v \in L^2(\Omega) : \forall E \in \mathcal{E}_h,\ v \circ F_E \in \mathbb{P}_k(\hat{E})\}.$$

On such elements, the approximation results for $\mathcal{D}_k(\mathcal{E}_h)$ and $\tilde{\mathcal{D}}_k(\mathcal{E}_h)$ are the same (see Section 2.1.4). However, in the case of general quadrilaterals, the space $\mathbb{P}_k(\hat{E})$ does not have optimal approximation properties (see [2]), whereas the space $\mathbb{P}_k(E)$ has optimal approximation properties (see [58]).

Therefore, for general quadrilateral meshes, we can either choose $\mathbb{P}_k(E)$ and do the computations on the physical elements, or we can choose to increase the discrete space and use the space $\mathbb{Q}_k(\hat{E})$, where $\mathbb{Q}_k$ denotes the space of polynomials of degree less than k in each space direction. The space $\mathbb{Q}_k$ is a tensor product space, and its dimension is strictly greater than the dimension of $\mathbb{P}_k$ for $k \geq 1$. Therefore, the computational costs increase.

2.5.2 Basis functions

Because of the lack of continuity constraints between mesh elements for the test functions, the basis functions of $\mathcal{D}_k(\mathcal{E}_h)$ have a support contained in one element. We write

$$\mathcal{D}_k(\mathcal{E}_h) = \operatorname{span}\{\phi_i^E : 1 \leq i \leq N_{\text{loc}},\ E \in \mathcal{E}_h\}$$

with

$$\phi_i^E(\boldsymbol{x}) = \begin{cases} \hat{\phi}_i \circ F_E(\boldsymbol{x}), & \boldsymbol{x} \in E, \\ 0, & \boldsymbol{x} \notin E. \end{cases} \tag{2.34}$$

The local basis functions $(\hat{\phi}_i)_{1 \leq i \leq N_{\text{loc}}}$ are defined on the reference element. We propose to simply use the monomial functions. For instance, in 2D, we have

$$\hat{\phi}_i(\hat{x}, \hat{y}) = \hat{x}^I \hat{y}^J, \quad I + J = i, \quad 0 \leq i \leq k.$$

This yields the local dimension

$$N_{\text{loc}} = \frac{(k+1)(k+2)}{2}.$$

For instance, we have the following:

- Piecewise linears:

$$\hat{\phi}_0(\hat{x}, \hat{y}) = 1, \quad \hat{\phi}_1(\hat{x}, \hat{y}) = \hat{x}, \quad \hat{\phi}_2(\hat{x}, \hat{y}) = \hat{y}.$$

- Piecewise quadratics:

$$\hat{\phi}_0(\hat{x},\hat{y}) = 1, \quad \hat{\phi}_1(\hat{x},\hat{y}) = \hat{x}, \quad \hat{\phi}_2(\hat{x},\hat{y}) = \hat{y},$$
$$\hat{\phi}_3(\hat{x},\hat{y}) = \hat{x}^2, \quad \hat{\phi}_4(\hat{x},\hat{y}) = \hat{x}\hat{y}, \quad \hat{\phi}_5(\hat{x},\hat{y}) = \hat{y}^2.$$

Similarly, in 3D, we define

$$\hat{\phi}_i(\hat{x},\hat{y},\hat{z}) = \hat{x}^I\hat{y}^J\hat{z}^K, \quad I+J+K=i, \quad 0 \le i \le k.$$

This yields the local dimension

$$N_{\text{loc}} = \frac{(k+1)(k+2)(k+3)}{6}.$$

The flexibility of DG methods allows us to easily change basis functions. For instance, we could use Legendre polynomials or some other polynomials satisfying a desired orthogonality property.

2.5.3 Numerical quadrature

One-dimensional case: An integral over a segment is computed by first mapping the physical edge to the segment $(-1, 1)$, which is the reference element in 1D. Then, the integral is approximated by using a numerical quadrature rule on the interval $(-1, 1)$ such as the Gauss quadrature rule (1.11) defined in Section 1.4.2 and in Appendix A.

Two-dimensional case: The integral of a function $\hat{v}$ defined on the reference element $\hat{E}$ can be computed by using a quadrature rule [44]:

$$\int_{\hat{E}} \hat{v} \approx \sum_{j=1}^{Q_{\text{D}}} w_j \hat{v}(s_{x,j}, s_{y,j}).$$

Appendix A contains the sets of weights w_j and nodes $(s_{x,j}, s_{y,j}) \in \hat{E}$ for different values of Q_{D}. For instance, Table 2.1 gives a rule with 6 quadrature points that is exact for polynomials of total degree less than 4. Since DG methods easily allow for high order approximation, it is important to have high order quadrature rules.

Let E be a triangle or a tetrahedron. The mapping $F_E : \hat{E} \to E$ is affine, and we have

$$\int_E v = \int_{\hat{E}} v \circ F_E \det(\boldsymbol{B}_E) = 2|E| \int_{\hat{E}} \hat{v}.$$

This integral is then approximated by

$$\int_E v \approx 2|E| \sum_{j=1}^{Q_{\text{D}}} w_j \hat{v}(s_{x,j}, s_{y,j}).$$

Table 2.1. *Weights and points for quadrature rule on reference triangle.*

w_j	$s_{x,j}$	$s_{y,j}$
0.11169079483901	0.445948490915965	0.445948490915965
0.11169079483901	0.108103018168070	0.445948490915965
0.11169079483901	0.445948490915965	0.108103018168070
0.05497587182766	0.091576213509771	0.091576213509771
0.05497587182766	0.816847572980459	0.091576213509771
0.05497587182766	0.091576213509771	0.816847572980459

If the integrand involves a vector function $\boldsymbol{w}$ and the gradient of v, we have

$$\begin{aligned}\int_E \nabla v \cdot \boldsymbol{w} &= 2|E| \int_{\hat{E}} (\boldsymbol{B}_E^T)^{-1} \hat{\nabla} \hat{v} \cdot \hat{\boldsymbol{w}} \\ &\approx 2|E| \sum_{j=1}^{Q_{\mathrm{D}}} w_j (\boldsymbol{B}_E^T)^{-1} \hat{\nabla} \hat{v}(s_{x,j}, s_{y,j}) \cdot \hat{\boldsymbol{w}}(s_{x,j}, s_{y,j}).\end{aligned}$$

Similarly, if the integrand involves the gradient of both v and w, we have

$$\int_E \nabla v \cdot \nabla w \approx 2|E| \sum_{j=1}^{Q_{\mathrm{D}}} w_j (\boldsymbol{B}_E^T)^{-1} \hat{\nabla} \hat{v}(s_{x,j}, s_{y,j}) \cdot (\boldsymbol{B}_E^T)^{-1} \hat{\nabla} \hat{w}(s_{x,j}, s_{y,j}).$$

2.6 DG scheme

The general DG finite element method is as follows: Find P_h in $\mathcal{D}_k(\mathcal{E}_h)$ such that

$$\forall\, v \in \mathcal{D}_k(\mathcal{E}_h), \quad a_\epsilon(P_h, v) = L(v). \tag{2.35}$$

The same terminology defined for the one-dimensional case (see Section 1.2) applies here.

- If $\epsilon = -1$, the method is called symmetric interior penalty Galerkin (SIPG). We will see that this method converges if the penalty parameter σ_e^0 is large enough.
- If $\epsilon = +1$, the method is called nonsymmetric interior penalty Galerkin (NIPG). We will see that this method converges for any nonnegative values of the penalty parameter σ_e^0. This class of methods also encompasses the case where $\sigma_e^0 = 0$, which has appeared in the literature as the OBB method [84].
- If $\epsilon = 0$, the method is called incomplete interior penalty Galerkin (IIPG). We will see that this method converges under the same condition as for the SIPG; namely the penalty parameter σ_e^0 should be large enough.
- The $J_1^{\sigma_1,\beta_1}$ term is an extra stabilization term. The analysis of the method is independent of this term, and, from now on, we will assume for simplicity that $\sigma_e^1 = 0$ for all e.

2.7 Properties

2.7.1 Coercivity of bilinear forms

Definition 2.10. *A bilinear form a defined on a normed linear space V with norm $\|\cdot\|_V$ is coercive if there is a positive constant κ such that*

$$\forall v \in V, \quad \kappa \|v\|_V^2 \leq a(v, v).$$

For the DG bilinear form, we have

$$\begin{aligned} a_\epsilon(v, v) = &\sum_{E \in \mathcal{E}_h} \int_E \boldsymbol{K}(\nabla v)^2 + \int_\Omega \alpha v^2 \\ &+ (\epsilon - 1) \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e\} [v] + J_0^{\sigma_0, \beta_0}(v, v). \end{aligned}$$

Define the *energy* norm on $\mathcal{D}_k(\mathcal{E}_h)$:

$$\|v\|_{\mathcal{E}} = \left(\sum_{E \in \mathcal{E}_h} \int_E \boldsymbol{K} \nabla v \cdot \nabla v + \int_\Omega \alpha v^2 + J_0^{\sigma_0, \beta_0}(v, v) \right)^{1/2}. \tag{2.36}$$

It is easy to check that it is indeed a norm if $\sigma_0^e > 0$ for all e. We remark that we immediately have the coercivity property satisfied for $\epsilon = 1$. The coercivity constant is $\kappa = 1$. Indeed,

$$\forall v \in \mathcal{D}_k(\mathcal{E}_h), \quad \|v\|_{\mathcal{E}}^2 = a_\epsilon(v, v).$$

In the case where $\epsilon = -1$ or $\epsilon = 0$, we obtain using Cauchy–Schwarz's inequality an upper bound of the term $\sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e\} [v]$:

$$\begin{aligned} \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e\}[v] &\leq \sum_{e \in \Gamma_h \cup \Gamma_D} \|\{\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e\}\|_{L^2(e)} \|[v]\|_{L^2(e)} \\ &\leq \sum_{e \in \Gamma_h \cup \Gamma_D} \|\{\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e\}\|_{L^2(e)} \left(\frac{1}{|e|^{\beta_0}} \right)^{1/2 - 1/2} \|[v]\|_{L^2(e)}. \end{aligned}$$

Next, we consider the average of the fluxes for an interior edge e shared by the elements E_1^e and E_2^e:

$$\|\{\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e\}\|_{L^2(e)} \leq \frac{1}{2} \|(\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e)|_{E_1^e}\|_{L^2(e)} + \frac{1}{2} \|(\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e)|_{E_2^e}\|_{L^2(e)}.$$

Using the property (2.19) of $\boldsymbol{K}$ and the trace inequality (2.6), we have

$$\begin{aligned} \|\{\boldsymbol{K} \nabla v \cdot \boldsymbol{n}_e\}\|_{L^2(e)} &\leq \frac{K_1}{2} \|(\nabla v \cdot \boldsymbol{n}_e)|_{E_1^e}\|_{L^2(e)} + \frac{K_1}{2} \|(\nabla v \cdot \boldsymbol{n}_e)|_{E_2^e}\|_{L^2(e)} \\ &\leq \frac{C_t K_1}{2} h_{E_1^e}^{-1/2} \|\nabla v\|_{L^2(E_1^e)} + \frac{C_t K_1}{2} h_{E_2^e}^{-1/2} \|\nabla v\|_{L^2(E_2^e)}. \end{aligned}$$

So we have using (2.22)

$$\begin{aligned}\int_e \{\boldsymbol{K}\nabla v\cdot \boldsymbol{n}_e\}[v] &\le \frac{C_t K_1}{2}|e|^{\beta_0/2}\Big(h_{E_1^e}^{-1/2}\|\nabla v\|_{L^2(E_1^e)}\\ &\quad + h_{E_2^e}^{-1/2}\|\nabla v\|_{L^2(E_2^e)}\Big)\left(\frac{1}{|e|^{\beta_0}}\right)^{1/2}\|[v]\|_{L^2(e)}\\ &\le \frac{C_t K_1}{2}\left(h_{E_1^e}^{\frac{\beta_0}{2}(d-1)-\frac{1}{2}} + h_{E_2^e}^{\frac{\beta_0}{2}(d-1)-\frac{1}{2}}\right)\Big(\|\nabla v\|^2_{L^2(E_1^e)}\\ &\quad + \|\nabla v\|^2_{L^2(E_2^e)}\Big)^{1/2}\left(\frac{1}{|e|^{\beta_0}}\right)^{1/2}\|[v]\|_{L^2(e)}\\ &\le C_t K_1\Big(\|\nabla v\|^2_{L^2(E_1^e)} + \|\nabla v\|^2_{L^2(E_2^e)}\Big)^{1/2}\left(\frac{1}{|e|^{\beta_0}}\right)^{1/2}\|[v]\|_{L^2(e)}\end{aligned}$$

if β_0 satisfies the condition $\beta_0(d-1)\ge 1$ and if we assume, without loss of generality, that $h\le 1$. A similar bound is obtained if e is a boundary edge. Let n_0 denote the maximum number of neighbors an element can have, i.e., for a conforming mesh, $n_0 = 3$ for a triangle and $n_0 = 4$ for a quadrilateral or tetrahedron:

$$\begin{aligned}\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\int_e \{\boldsymbol{K}\nabla v\cdot \boldsymbol{n}_e\}[v] &\le C_t K_1\left(\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{1}{|e|^{\beta_0}}\|[v]\|^2_{L^2(e)}\right)^{1/2}\\ &\quad\times\left(\sum_{e\in\Gamma_h}\|\nabla v\|^2_{L^2(E_1^e)} + \|\nabla v\|^2_{L^2(E_2^e)} + \sum_{e\in\Gamma_D}\|\nabla v\|^2_{0,E_1^e}\right)\\ &\le C_t K_1\sqrt{n_0}\left(\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{1}{|e|^{\beta_0}}\|[v]\|^2_{L^2(e)}\right)^{1/2}\left(\sum_{E\in\mathcal{E}_h}\|\nabla v\|^2_{L^2(E)}\right)^{1/2}.\end{aligned}$$

Using Young's inequality, we have for $\delta > 0$

$$\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\int_e \{\boldsymbol{K}\nabla v\cdot \boldsymbol{n}_e\}[v] \le \frac{\delta}{2}\sum_{E\in\mathcal{E}_h}\|\boldsymbol{K}^{1/2}\nabla v\|^2_{L^2(E)} + \frac{C_t^2K_1^2n_0}{2\delta K_0}\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{1}{|e|^{\beta_0}}\|[v]\|^2_{L^2(e)}.$$

Thus, we obtain a lower bound for $a_\epsilon(v,v)$:

$$a_\epsilon(v,v) \ge \left(1-\frac{\delta}{2}|1-\epsilon|\right)\sum_{E\in\mathcal{E}_h}\|\boldsymbol{K}^{1/2}\nabla v\|^2_{L^2(E)} + \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{\sigma_e^0 - \frac{C_t^2K_1^2n_0}{2\delta K_0}|1-\epsilon|}{|e|^{\beta_0}}\|[v]\|^2_{L^2(e)}.$$

Choosing, for instance, $\delta = 1$ if $\epsilon = 0$ and $\delta = 1/2$ if $\epsilon = -1$ and choosing σ_e^0 large enough (for example, $\sigma_e^0 \ge (C_t^2K_1^2n_0/K_0)$ if $\epsilon = 0$ and $\sigma_e^0 \ge (2C_t^2K_1^2n_0/K_0)$ if $\epsilon = -1$), then we have the coercivity result with $\kappa = 1/2$:

$$a_\epsilon(v,v) \ge \kappa\|v\|^2_{\mathcal{E}}. \tag{2.37}$$

Summarizing the results above, we have

- a_{+1} is coercive;
- a_{-1} and a_0 are coercive if $\beta_0(d-1) \geq 1$ and if σ_e^0 is bounded below by a constant σ_e^* that depends only on K_0, K_1, and the constant in the trace inequality (2.6).

Remark: As expected, the threshold value for the penalty parameter is twice as large for the SIPG method as for the IIPG method. A more precise value of σ_e^* can be obtained if one uses the trace inequalities (2.7)–(2.9) rather than (2.6). For instance, on a triangular mesh, for a given triangle E, if θ^E denotes the smallest angle in E, if K_0^E, K_1^E denote the lower and upper bound of $\boldsymbol{K}$ on E, and if k^E denotes the polynomial degree of the approximation on E, the limiting value of the penalty depends on the local quantities θ^E, K_0^E, K_1^E, and k^E as follows:

$$\forall e \in \Gamma_h, \quad \sigma_e^* = \frac{3(K_1^{E_e^1})^2}{2K_0^{E_e^1}}(k^{E_e^1})(k^{E_e^1}+1)|e|^{\beta_0-1}\cot\theta^{E_e^1} + \frac{3(K_1^{(E_e^2)})^2}{2K_0^{E_e^2}}(k^{E_e^2})(k^{E_e^2}+1)|e|^{\beta_0-1}\cot\theta^{E_e^2}, \tag{2.38}$$

$$\forall e \in \Gamma_D, \quad \sigma_e^* = \frac{6(K_1^{E_e^1})^2}{K_0^{E_e^1}}(k^{E_e^1})(k^{E_e^1}+1)\cot\theta^{E_e^1}|e|^{\beta_0-1}. \tag{2.39}$$

Similarly, in the three-dimensional case, with a tetrahedral mesh, the limiting value depends also on local quantities such as the dihedral angle θ^E in the tetrahedron E that yields the smallest value for $\sin\theta$ over all dihedral angles θ of E:

$$\forall e \in \Gamma_h, \quad \sigma_e^* = \frac{3}{2}\frac{(K_1^{E_e^1})^2}{K_0^{E_e^1}}k^{E_e^1}(k^{E_e^1}+2)h\,|e|^{\beta_0-1}\cot\theta_{E_e^1} + \frac{3}{2}\frac{(K_1^{E_e^2})^2}{K_0^{E_e^2}}k^{E_e^2}(k^{E_e^2}+2)h\,|e|^{\beta_0-1}\cot\theta_{E_e^2}, \tag{2.40}$$

$$\forall e \in \Gamma_D, \quad \sigma_e^* = 6\frac{(K_1^{E_e^1})^2}{K_0^{E_e^1}}k^{E_e^1}(k^{E_e^1}+2)h\,|e|^{\beta_0-1}\cot\theta_{E_e^1}. \tag{2.41}$$

If $\sigma_e \geq \sigma_e^*$, then the SIPG and IIPG methods are stable and convergent. The proof of these results can be found in [50].

2.7.2 Continuity of bilinear form

Definition 2.11. *A bilinear form a defined on a linear space V equipped with norm $\|\cdot\|_V$ is continuous if there is a positive constant M such that*

$$\forall v, w \in V, \quad a(v,w) \leq M\|v\|_V\|w\|_V.$$

If $\sigma_e^0 > 0$ for all e, then one can show that the bilinear form a_ϵ is continuous on $\mathcal{D}_k(\mathcal{E}_h)$ equipped with the energy norm $\|\cdot\|_\mathcal{E}$:

$$\forall v \in \mathcal{D}_k(\mathcal{E}_h), \quad a_\epsilon(v, w) \leq M\|v\|_\mathcal{E}\|w\|_\mathcal{E}.$$

However, the bilinear form is not continuous in general on the broken space $H^2(\mathcal{E}_h)$ with respect to the energy norm.

2.7.3 Local mass conservation

One interesting property that naturally comes with the primal DG methods is the conservation of mass on each mesh element. Because of the lack of continuity constraints between the elements, we can choose a test function $v \in \mathcal{D}_k(\mathcal{E}_h)$ that is equal to a different constant on each element. If we fix an element E that belongs to the interior of the domain and if we choose v equal to the constant 1 on E and the constant 0 elsewhere, the method (2.35) reduces to

$$\int_E \alpha P_h - \sum_{e\in\partial E}\int_e \{\boldsymbol{K}\nabla P_h \cdot \boldsymbol{n}_e\}[v] + \sum_{e\in\partial E}\frac{\sigma_e^0}{|e|^{\beta_0}}\int_e [P_h][v] = \int_E f.$$

This is equivalent to

$$\int_E (\alpha P_h - f) + \sum_{e\in\partial E}\frac{\sigma_e^0}{|e|^{\beta_0}}\int_e (P_h|_E - P_h|_{\mathcal{N}(e;E)}) = \int_{\partial E}\{\boldsymbol{K}\nabla P_h \cdot \boldsymbol{n}_E\},$$

where $\mathcal{N}(e; E)$ denotes the element in $\mathcal{E}_h$ that is a neighbor of E through the edge e. Thus, we have obtained a balance equation valid on the element E. A similar equation can be derived if the element E shares at least one face with the boundary of the domain. If we assume that the quantity P_h represents a mass density, then the term $\int_E(\alpha P_h - f)$ corresponds to the mass that is created or destroyed inside E and the term $\int_{\partial E}\{\boldsymbol{K}\nabla P_h \cdot \boldsymbol{n}_E\}$ corresponds to the flux of mass passing through the boundary ∂E. The additional term involving the penalty parameter is a pure numerical mass that is zero if the penalty value is zero. In general, this artificial mass can be exactly computed and can be subtracted if needed.

Local mass conservation is important in particular in coupled flow and transport problems arising in porous media. For instance, Darcy flow can be characterized with the elliptic problem (2.16) with $\alpha = 0$, and the flow velocity $\boldsymbol{u} = -\boldsymbol{K}\nabla p$ is approximated by $\mathbf{U}_h = -\boldsymbol{K}\nabla P_h$. Then, the reactive transport of a chemical species of concentration c can be modeled by the following partial differential equation:

$$\frac{\partial c}{\partial t} - \nabla\cdot(D\nabla c - \boldsymbol{u}c) = r(c).$$

In this case, if the penalty is zero, local mass conservation means

$$\int_{\partial E}\{\mathbf{U}_h\}\cdot\boldsymbol{n}_E = \int_E f.$$

If the numerical approximation of the velocity is not locally conservative, the numerical solution of the transport equation becomes unstable after a few time steps. Chapter 4 describes the transport problem in more detail.

2.7.4 Existence and uniqueness of DG solution

Lemma 2.12. *Assume that* (i), (ii), *or* (iii) *holds true:*

(i) *in the NIPG case,* $k \geq 1$ *and either* $\alpha > 0$ *or* $\sigma_e^0 > 0$ *for all* e*;*

(ii) *in the SIPG or IIPG case,* $k \geq 1$ *and* σ_e^0 *is bounded below by a large constant for all* e*;*

(iii) *in the NIPG case,* $k \geq 2$ *and* $\sigma_e^0 = 0$ *for all* e *and* $\alpha = 0$.

Then, the DG solution P_h *exists and is unique.*

Proof. Since (2.35) is a linear problem in finite dimension, existence is equivalent to uniqueness. We assume that there are two solutions P_h^1 and P_h^2. The difference $w_h = P_h^1 - P_h^2$ satisfies

$$a_\epsilon(w_h, w_h) = 0.$$

By the coercivity result (2.37), we have

$$\|w_h\|_{\mathcal{E}} = 0.$$

Clearly in both cases (i) and (ii), this implies that $w_h = 0$ since $\|\cdot\|_{\mathcal{E}}$ is a norm. The case (iii) is not as easy. Indeed, we can conclude only that w_h is piecewise constant on each element $E \in \mathcal{E}_h$. In order to prove that w_h is globally constant in Ω, we need to construct a test function v on a given element E such that the quantity $\int_e \boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e$ is given on one edge (or face) of E and vanishes on the other edges (or faces). If $\boldsymbol{K}$ is constant in each E, one can construct such a test function on a triangle, parallelogram, or tetrahedron (see Lemma C.1). If $\boldsymbol{K}$ is not constant in each E, one needs to assume in addition that h is small enough. ∎

2.8 Error analysis

In this section, we assume that the exact solution p belongs to $H^s(\mathcal{E}_h)$ for some $s > 3/2$, and we prove that the DG solution converges to the exact solution. We will first derive a priori error estimates in the energy norm.

2.8.1 Error estimates in the energy norm

By the triangle inequality, we have

$$\|p - P_h\|_{\mathcal{E}} \leq \|p - \tilde{p}\|_{\mathcal{E}} + \|P_h - \tilde{p}\|_{\mathcal{E}}$$

for a function $\tilde{p} \in \mathcal{D}_k(\mathcal{E}_h)$ that approximates the exact solution p as in Theorem 2.6. Then, it suffices to bound $\|P_h - \tilde{p}\|_{\mathcal{E}}$. By consistency (see Section 2.4.1), the error satisfies the *orthogonality* equation

$$\forall v \in \mathcal{D}_k(\mathcal{E}_h), \quad a_\epsilon(P_h - p, v) = 0. \tag{2.42}$$

Denoting $\chi = P_h - \tilde{p}$ and adding and subtracting $\tilde{p}$ in each term yields

$$a_\epsilon(\chi, v) = a_\epsilon(p - \tilde{p}, v).$$

Choosing the test function $v = \chi$ and using the coercivity result (2.37) gives

$$\begin{aligned}\kappa\|\chi\|_{\mathcal{E}}^2 \le \Big| \sum_{E\in\mathcal{E}_h}\int_E (\boldsymbol{K}\nabla(p-\tilde{p})\nabla\chi + \alpha(p-\tilde{p})\chi) - \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\int_e \{\boldsymbol{K}\nabla(p-\tilde{p})\cdot\boldsymbol{n}_e\}[\chi] \\ +\epsilon\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\int_e \{\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e\}[p-\tilde{p}] + J_0^{\sigma_0,\beta_0}(p-\tilde{p},\chi)\Big| \\ \le |T_1+\cdots+T_4|.\end{aligned}$$

Using the bound (2.19), Cauchy–Schwarz's inequality, and Young's inequality, we have

$$\begin{aligned}|T_1| &\le K_1^{1/2}\left(\sum_E \|\boldsymbol{K}^{1/2}\nabla\chi\|_{L^2(E)}^2\right)^{1/2}\left(\sum_E \|\nabla(p-\tilde{p})\|_{L^2(E)}^2\right)^{1/2} \\ &\quad +\|\alpha\|_{L^\infty(\Omega)}^{1/2}\|\alpha^{\frac{1}{2}}\chi\|_{L^2(\Omega)}\|p-\tilde{p}\|_{L^2(\Omega)} \\ &\le \frac{3}{2\kappa}(K_1+\|\alpha\|_{L^\infty(\Omega)})|||p-\tilde{p}|||_{H^1(\mathcal{E}_h)}^2 + \frac{\kappa}{6}\|\chi\|_{\mathcal{E}}^2.\end{aligned}$$

Let C denote a generic constant independent of h that takes different values at different places. From the approximation result (2.10), we obtain

$$T_1 \le Ch^{2\min(k+1,s)-2}|||p|||_{H^s(\mathcal{E}_h)}^2 + \frac{\kappa}{6}\|\chi\|_{\mathcal{E}}^2.$$

Let us now bound T_3: this term disappears if the method is IIPG ($\epsilon = 0$) or if $\tilde{p}$ is chosen to be a continuous interpolant (such as the classical Lagrange interpolant) and either $|\Gamma_{\mathrm{D}}| = 0$ or g_{D} is a polynomial of degree k (hence, one can choose $\tilde{p} = g_{\mathrm{D}}$ on Γ_{D}). However, in the general case (for example, if $\tilde{p}$ is not continuous), we can still control this term by using trace inequalities and approximation results. First, we have by Cauchy–Schwarz's inequality

$$|T_3| \le \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \|\{\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e\}\|_{L^2(e)}\|[p-\tilde{p}]\|_{L^2(e)}.$$

Now if the edge (or face) is interior, $e = \partial E_e^1 \cap \partial E_e^2$, we can apply the trace inequality (2.1) for each neighboring element:

$$\begin{aligned}\|[p-\tilde{p}]\|_{L^2(e)} &\le \|(p-\tilde{p})|_{E_e^1}\|_{L^2(e)} + \|(p-\tilde{p})|_{E_e^2}\|_{L^2(e)} \\ &\le C|e|^{1/2}|E_e^1|^{-1/2}(\|p-\tilde{p}\|_{L^2(E_e^1)} + h_{E_e^1}\|\nabla(p-\tilde{p})\|_{L^2(E_e^1)}) \\ &\quad +C|e|^{1/2}|E_e^2|^{-1/2}(\|p-\tilde{p}\|_{L^2(E_e^2)} + h_{E_e^2}\|\nabla(p-\tilde{p})\|_{L^2(E_e^2)}).\end{aligned}$$

Using the trace inequality (2.5) in finite-dimensional spaces, we have

$$\begin{aligned}\|\{\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e\}\|_{L^2(e)} &\le \frac{1}{2}\|(\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e)|_{E_e^1}\|_{L^2(e)} + \frac{1}{2}\|(\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e)|_{E_e^2}\|_{L^2(e)} \\ &\le \frac{K_1}{2}\tilde{C}_t|e|^{1/2}|E_e^1|^{-1/2}||\nabla\chi\|_{L^2(E_e^1)} + \frac{K_1}{2}\tilde{C}_t|e|^{1/2}|E_e^2|^{-1/2}||\nabla\chi\|_{L^2(E_e^2)}.\end{aligned}$$

Combining the two bounds above, we obtain

$$\begin{aligned}
\forall e \in \Gamma_h, \quad & \|\{\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e\}\|_{L^2(e)}\|[p-\tilde{p}]\|_{L^2(e)} \\
&\le C|e||E_e^1|^{-1}(\|p-\tilde{p}\|_{L^2(E_e^1)} + h_{E_e^1}\|\nabla(p-\tilde{p})\|_{L^2(E_e^1)})||\nabla\chi||_{L^2(E_e^1)} \\
&+C(|e||E_e^2|^{-1/2})|E_e^1|^{-1/2}(\|p-\tilde{p}\|_{L^2(E_e^1)} + h_{E_e^1}\|\nabla(p-\tilde{p})\|_{L^2(E_e^1)})||\nabla\chi||_{L^2(E_e^2)} \\
&+C|e||E_e^2|^{-1}(\|p-\tilde{p}\|_{L^2(E_e^2)} + h_{E_e^2}\|\nabla(p-\tilde{p})\|_{L^2(E_e^2)})||\nabla\chi||_{L^2(E_e^2)} \\
&+C(|e||E_e^1|^{-1/2})|E_e^2|^{-1/2}(\|p-\tilde{p}\|_{L^2(E_e^2)} + h_{E_e^2}\|\nabla(p-\tilde{p})\|_{L^2(E_e^2)})||\nabla\chi||_{L^2(E_e^1)}.
\end{aligned}$$

Using the approximation results (2.10) and the fact that for $i = 1, 2$, the product $|e||E_e^i|^{-1/2}$ is bounded by a constant C in 2D and bounded by $Ch_{E_e^i}^{1/2}$ in 3D, we have

$$\begin{aligned}
\forall e \in \Gamma_h, \quad & \|\{\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e\}\|_{L^2(e)}\|[p-\tilde{p}]\|_{L^2(e)} \\
&\le Ch^{\min(k+1,s)-1}(|p|_{H^s(E_e^1)} + |p|_{H^s(E_e^2)})(\|\nabla\chi\|_{L^2(E_e^1)} + \|\nabla\chi\|_{L^2(E_e^2)}).
\end{aligned}$$

Assume now that the edge (or face) e is on the Dirichlet boundary Γ_{D} and belongs to the element E_e^1. Following a similar argument as above, we have

$$\|\boldsymbol{K}\nabla\chi\cdot\boldsymbol{n}_e\|_{L^2(e)}\|p-\tilde{p}\|_{L^2(e)} \le Ch^{\min(k+1,s)-1}|p|_{H^s(E_e^1)}\|\nabla\chi\|_{L^2(E_e^1)}.$$

Therefore, the term T_3 is bounded by

$$T_3 \le Ch^{2\min(k+1,s)-2}|||p|||^2_{H^s(\mathcal{E}_h)} + \frac{\kappa}{6}\|\chi\|^2_{\mathcal{E}}.$$

The term T_4 is zero if $\sigma_e^0 = 0$ for all e or if $\tilde{p}$ is continuous and either $|\Gamma_{\mathrm{D}}| = 0$ or g_{D} is a continuous piecewise polynomial of degree k. Otherwise, using the fact that $|e| \le h^{d-1}$, the term T_4 is simply bounded using Cauchy–Schwarz's and Young's inequalities:

$$\begin{aligned}
|T_4| &\le \frac{3}{2\kappa}J_0^{\sigma_0,\beta_0}(p-\tilde{p}, p-\tilde{p}) + \frac{\kappa}{6}J_0^{\sigma_0,\beta_0}(\chi,\chi) \\
&\le \frac{\kappa}{6}\|\chi\|^2_{\mathcal{E}} + Ch^{2\min(k+1,s)-1-\beta_0(d-1)}|||p|||^2_{H^s(\mathcal{E}_h)}.
\end{aligned}$$

Thus, T_4 is optimal if the condition $\beta_0(d-1) \le 1$ is satisfied. Under the assumptions given above, we obtain

$$\frac{\kappa}{2}\|\chi\|^2_{\mathcal{E}} \le Ch^{2\min(k+1,s)-2}|||p|||^2_{H^s(\mathcal{E}_h)} + |T_2|.$$

In order to conclude, it remains to bound the term T_2. On one hand, this term is relatively easy to bound if all penalty values are nonzero. On the other hand, if some penalty values are zero, the bound of T_2 requires an additional property on the approximation $\tilde{p}$ and a restriction of the polynomial degree $k \ge 2$. Thus, we distinguish two cases. First, let us assume that $\sigma_e^0 > 0$ for all e; then we can write

$$\left|\int_e \{\boldsymbol{K}\nabla(p-\tilde{p})\cdot\boldsymbol{n}_e\}[\chi]\right| \le \left(\frac{|e|^{\beta_0}}{\sigma_e^0}\right)^{\frac{1}{2}}\|\{\boldsymbol{K}\nabla(p-\tilde{p})\cdot\boldsymbol{n}_e\}\|_{L^2(e)}\left(\frac{\sigma_e^0}{|e|^{\beta_0}}\right)^{\frac{1}{2}}\|[\chi]\|_{L^2(e)},$$

$$\left|\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\{\boldsymbol{K}\nabla(p-\tilde{p})\cdot\boldsymbol{n}_e\}[\chi]\right| \le \frac{\kappa}{6}J_0^{\sigma_0,\beta_0}(\chi,\chi) + C\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{|e|^{\beta_0}}{\sigma_e^0}\|\{\boldsymbol{K}\nabla(p-\tilde{p})\cdot\boldsymbol{n}_e\}\|^2_{L^2(e)}.$$

Using the trace inequality (2.2) and the approximation result (2.10), we have

$$|T_2| \le \frac{\kappa}{6}\|\chi\|_{\mathcal{E}}^2 + Ch^{2\min(k+1,s)-3+\beta_0(d-1)}|||p|||_{H^s(\mathcal{E}_h)}^2.$$

Thus, if $\beta_0(d-1) \ge 1$, we have

$$|T_2| \le \frac{\kappa}{6}\|\chi\|_{\mathcal{E}}^2 + Ch^{2\min(k+1,s)-2}|||p|||_{H^s(\mathcal{E}_h)}^2, \tag{2.43}$$

and the final error estimate is

$$\frac{\kappa}{3}\|\chi\|_{\mathcal{E}}^2 \le Ch^{2\min(k+1,s)-2}|||p|||_{H^s(\mathcal{E}_h)}^2. \tag{2.44}$$

In the second case, let us assume that $\sigma_e^0 = 0$ for some e. Then, for each element E, we use the approximation $\tilde{p} \in \mathbb{P}_k(E)$ defined in Theorem 2.7. Since this approximation is defined locally on each E, we have

$$\forall E \in \mathcal{E}_h, \quad \forall e \in \partial E, \int_e \{\boldsymbol{K}\nabla(p-\tilde{p})\cdot \boldsymbol{n}_e\} = 0.$$

We then rewrite the term T_2 for any real number c_e:

$$\begin{aligned} T_2 &= \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \int_e \{\boldsymbol{K}\nabla(p-\tilde{p})\cdot \boldsymbol{n}_e\}([\chi]-c_e) \\ &\le \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \|\{\boldsymbol{K}\nabla(p-\tilde{p})\cdot \boldsymbol{n}_e\}\|_{L^2(e)}\|[\chi]-c_e\|_{L^2(e)}. \end{aligned}$$

If e is an interior edge (or face) and is shared by E_e^1 and E_e^2, we choose

$$c_e = c_1 - c_2, \quad c_i = \frac{1}{|E_e^i|}\int_{E_e^i} \chi, \quad i = 1, 2,$$

and we observe that

$$[\chi] - c_e = \chi|_{E_e^1} - \chi|_{E_e^2} - (c_1 - c_2) = (\chi|_{E_e^1} - c_1) - (\chi|_{E_e^2} - c_2).$$

Thus, we have by the trace inequality (2.1):

$$\begin{aligned} \|[\chi]-c_e\|_{L^2(e)} &\le \|\chi|_{E_1} - c_1\|_{L^2(e)} + \|\chi|_{E_2} - c_2\|_{L^2(e)} \\ &\le Ch_{E_e^1}^{-1/2}(\|\chi - c_1\|_{L^2(E_e^1)} + h_{E_e^1}\|\nabla\chi\|_{L^2(E_e^1)}) \\ &\quad + Ch_{E_e^2}^{-1/2}(\|\chi - c_2\|_{L^2(E_e^2)} + h_{E_e^2}^{1/2}\|\nabla\chi\|_{L^2(E_e^2)}). \end{aligned}$$

Next, by definition of the constant c_i, we have

$$\int_{E_e^i} (\chi|_{E_e^i} - c_i) = 0, \quad i = 1, 2.$$

Thus, we have

$$\|[\chi] - c_e\|_{L^2(e)} \le Ch_{E_e^1}^{1/2}\|\nabla\chi\|_{L^2(E_e^1)} + Ch_{E_e^1}^{1/2}\|\nabla\chi\|_{L^2(E_e^2)}.$$

Indeed, we have used the following result: If a function ϕ belongs to $H^1(E)$ such that $\int_E \phi = 0$, then there is a constant C independent of h_E such that

$$\|\phi\|_{L^2(E)} \le Ch_E\|\nabla\phi\|_{L^2(E)}.$$

Note that if the face e belongs to the boundary $\Gamma_h \cap \partial E_e^1$, then we choose $c_e = \frac{1}{|E_e^1|}\int_{E_e^1}\chi$, and we obtain similarly

$$\|\chi - c_e\|_{L^2(e)} \le Ch_{E_e^1}^{1/2}\|\nabla\chi\|_{L^2(E_e^1)}.$$

The other factor in the term T_2 is bounded using trace inequality (2.2) and approximation result (2.10):

$$\forall e \in \Gamma_h, \quad \|\{\boldsymbol{K}\nabla(p-\tilde{p})\cdot\boldsymbol{n}_e\}\|_{L^2(e)} \le Ch^{\min(k+1,s)-3/2}(|p|_{H^s(E_e^1)} + |p|_{H^s(E_e^2)}),$$
$$\forall e \in \Gamma_{\mathrm{D}}, \quad \|\boldsymbol{K}\nabla(p-\tilde{p})\cdot\boldsymbol{n}_e\|_{L^2(e)} \le Ch^{\min(k+1,s)-3/2}|p|_{H^s(E_e^1)}.$$

Combining the bounds above gives an inequality identical to (2.43), and thus the bound (2.44) is obtained. We saw that the derivation of the error estimates requires a constraint on the power β_0, under a certain condition. Before summarizing the results, we state that condition.

Condition A: The approximation $\tilde{p}$ of the exact solution p can be chosen to be continuous. In addition, either the Dirichlet data g_{D} is a continuous piecewise polynomial of degree k, or the whole boundary is a Neumann boundary ($\partial\Omega = \Gamma_{\mathrm{N}}$).

Theorem 2.13. *Assume that the exact solution to* (2.16)–(2.18) *belongs to* $H^s(\mathcal{E}_h)$ *for* $s > 3/2$. *Assume also that the penalty parameter* σ_e^0 *is large enough for the SIPG and IIPG methods and that* $k \ge 2$ *for the NIPG method with zero penalty. Then, there is a constant* C *independent of* h *such that the following optimal a priori error estimate holds:*

$$\|p - P_h\|_{\mathcal{E}} \le Ch^{\min(k+1,s)-1}|||p|||_{H^s(\mathcal{E}_h)}.$$

This estimate is valid if Condition A holds true and if $\beta_0 \ge (d-1)^{-1}$. *Otherwise, if Condition A fails, this estimate is valid if* $\beta_0 = (d-1)^{-1}$.

2.8.2 Error estimates in the L^2 norm

Next, we prove an error estimate in the L^2 norm. We will apply the Aubin–Nitsche lift technique used in the analysis of the classical finite element method to the DG method. This technique works well if the scheme is symmetric. This is the case for the SIPG method. We will see below that optimal estimates cannot be derived for IIPG and NIPG. For simplicity, we assume that the entire boundary is a Dirichlet boundary, i.e., $\partial\Omega = \Gamma_{\mathrm{D}}$. We assume that the domain is convex and that the solution to the dual problem

$$-\nabla\cdot(\boldsymbol{K}\nabla\phi) + \alpha\phi = p - P_h \quad \text{in} \quad \Omega,$$
$$\phi = 0 \quad \text{on} \quad \partial\Omega$$

belongs to $H^2(\Omega)$ with continuous dependence on $p - P_h$:

$$\|\phi\|_{H^2(\Omega)} \leq C\|p - P_h\|_{L^2(\Omega)}. \tag{2.45}$$

Then, we have

$$\|P_h - p\|^2_{L^2(\Omega)} = \int_\Omega (P_h - p)^2 = \int_\Omega (-\nabla \cdot (\boldsymbol{K}\nabla\phi) + \alpha\phi)(P_h - p).$$

Denoting $\theta = P_h - p$ and integrating by parts on each element yields

$$\|\theta\|^2_{L^2(\Omega)} = \sum_{E\in\mathcal{E}_h} \int_E (\boldsymbol{K}\nabla\phi \cdot \nabla\theta + \alpha\phi\theta) - \sum_{E\in\mathcal{E}_h} \int_{\partial E} (\boldsymbol{K}\nabla\phi \cdot \boldsymbol{n}_E)\theta.$$

The last term can be rewritten as in (2.25). Since $\phi \in H^2(\Omega)$, we have

$$\|\theta\|^2_{L^2(\Omega)} = \sum_{E\in\mathcal{E}_h} \int_E (\boldsymbol{K}\nabla\phi\nabla\theta + \alpha\phi\theta) - \sum_{e\in\Gamma_h\cup\partial\Omega} \int_e \{\boldsymbol{K}\nabla\phi \cdot \boldsymbol{n}_e\}[\theta].$$

We now subtract the orthogonality equation (2.42) from the equation above:

$$\begin{aligned}
\forall v \in \mathcal{D}_k(\mathcal{E}_h), \quad \|\theta\|^2_{L^2(\Omega)} &= \sum_{E\in\mathcal{E}_h} \int_E (\boldsymbol{K}\nabla(\phi - v)\nabla\theta + \alpha(\phi - v)\theta) \\
&\quad -\epsilon \sum_{e\in\Gamma_h\cup\partial\Omega} \int_e \{\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e\}[\theta] - \sum_{e\in\Gamma_h\cup\partial\Omega} \int_e \{\boldsymbol{K}\nabla\phi \cdot \boldsymbol{n}_e\}[\theta] \\
&\quad + \sum_{e\in\Gamma_h\cup\partial\Omega} \int_e \{\boldsymbol{K}\nabla\theta \cdot \boldsymbol{n}_e\}[v] - J_0^{\sigma_0,\beta_0}(\theta, v) \\
&= A_1 + \cdots + A_5.
\end{aligned} \tag{2.46}$$

We choose $v = \tilde{\phi}$, a continuous interpolant of ϕ of degree k. We assume that such an interpolant exists. In that case, we note that $\tilde{\phi} = \phi = 0$ on the boundary $\partial\Omega$. The last two terms on the right-hand side of (2.46), namely A_4 and A_5, vanish. The first term is easily bounded using Cauchy–Schwarz's inequality and the approximation result (2.10):

$$A_1 \leq Ch\|\phi\|_{H^2(\Omega)}\|\theta\|_{\mathcal{E}}.$$

Therefore, we obtain

$$\|\theta\|^2_{L^2(\Omega)} \leq Ch\|\phi\|_{H^2(\Omega)}\|\theta\|_{\mathcal{E}} + S,$$

where

$$S = \left| \sum_{e\in\Gamma_h\cup\partial\Omega} \int_e \{\boldsymbol{K}\nabla(\phi + \epsilon\tilde{\phi}) \cdot \boldsymbol{n}_e\}[\theta] \right|.$$

If the method employed is the SIPG method, the term $(\phi + \epsilon\tilde{\phi}) = (\phi - \tilde{\phi})$ is the approximation error. A bound of S can be derived by taking advantage of the penalty parameter:

$$\begin{aligned} S &\leq \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \left(\frac{|e|^{\beta_0}}{\sigma_e^0}\right)^{\frac{1}{2}-\frac{1}{2}} \|\{\boldsymbol{K}\nabla(\phi-\tilde{\phi})\cdot\boldsymbol{n}_e\}\|_{L^2(e)}\|[\theta]\|_{L^2(e)} \\ &\leq J_0^{\sigma_0,\beta_0}(\theta,\theta)^{\frac{1}{2}} \left(\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \frac{|e|^{\beta_0}}{\sigma_e^0}\|\{\boldsymbol{K}\nabla(\phi-\tilde{\phi})\cdot\boldsymbol{n}_e\}\|^2_{L^2(e)}\right)^{\frac{1}{2}} \\ &\leq Ch^{\frac{\beta_0}{2}(d-1)+\frac{1}{2}}\|\phi\|_{H^2(\Omega)}\|\theta\|_{\mathcal{E}}. \end{aligned}$$

Therefore, using the bound (2.45), we obtain

$$\|\theta\|^2_{L^2(\Omega)} \leq C(h + h^{\frac{\beta_0}{2}(d-1)+\frac{1}{2}})\|\theta\|_{L^2(\Omega)}\|\theta\|_{\mathcal{E}}.$$

With Theorem 2.13 and under the condition $\beta_0(d-1) \geq 1$, this implies

$$\|\theta\|_{L^2(\Omega)} \leq Ch^{\min(k+1,s)}|\!|\!|p|\!|\!|_{H^s(\mathcal{E}_h)}. \tag{2.47}$$

If the method employed is the IIPG method or the NIPG method with nonzero penalty, one can recover an additional power of h with the term S if a stricter constraint is imposed on the parameter β_0. Indeed, using Cauchy–Schwarz's inequality and the trace inequality (2.2), we obtain if $\epsilon = 0$ or $\epsilon = 1$

$$\begin{aligned} S &\leq 2J_0^{\sigma_0,\beta_0}(\theta,\theta)^{\frac{1}{2}} \left(\sum_{e\in\Gamma_h\cup\partial\Omega} \frac{|e|^{\beta_0}}{\sigma_e^0}\|\{\boldsymbol{K}\nabla\phi\cdot\boldsymbol{n}_e\}\|^2_{L^2(e)}\right)^{\frac{1}{2}} \\ &\leq Ch^{\frac{\beta_0}{2}(d-1)-\frac{1}{2}}\|\phi\|_{H^2(\Omega)}\|\theta\|_{\mathcal{E}}. \end{aligned}$$

Therefore, under the assumptions of Theorem 2.13 and if $\beta_0(d-1) \geq 3$, we obtain the optimal error estimate (2.47). One can check that a similar result holds true in the NIPG case with nonzero penalty. We say that the DG method is *superpenalized* if $\beta_0 > (d-1)^{-1}$.

The only case that we did not consider is the case of the NIPG method with $\sigma_e^0 = 0$ for all e. One can prove a suboptimal error estimate, namely

$$\|\theta\|_{L^2(\Omega)} = \mathcal{O}(h^{\min(k+1,s)-1}).$$

We summarize the results below.

Theorem 2.14. *Assume that Theorem* 2.13 *holds. There is a constant C independent of h such that*

$$\|p - P_h\|_{L^2(\Omega)} \leq Ch^{\min(k+1,s)}|\!|\!|p|\!|\!|_{H^s(\mathcal{E}_h)}.$$

This estimate is valid for the SIPG method unconditionally and for the NIPG and IIPG methods under Condition A and the superpenalization $\beta_0 \geq 3(d-1)^{-1}$. If Condition A is not satisfied, then the numerical error for both the NIPG and IIPG methods satisfies the following suboptimal error estimate:

$$\|p - P_h\|_{L^2(\Omega)} \leq Ch^{\min(k+1,s)-1}|\!|\!|p|\!|\!|_{H^s(\mathcal{E}_h)}.$$

Remark: In the standard penalization $\beta_0 = (d-1)^{-1}$, we know how to prove suboptimal error estimates for both the IIPG and NIPG methods. It has been observed numerically on uniform meshes that convergence rates are optimal if the polynomial degree is odd and suboptimal if the polynomial degree is even (see Section 2.10). This is an interesting question that remains to be theoretically solved. For general meshes one can construct an example for which the numerical rates are suboptimal even if the polynomial degree is odd.

Remark: In this section, we have considered convergence of the h-version of the DG method. The polynomial degree is kept fixed, and the mesh is successively refined. In the hp-version, both mesh size and polynomial degrees can be changed and error estimates can be derived (see, for instance, [96, 72]). They are suboptimal with respect to the polynomial degree.

Remark: In the case of meshes containing quadrilaterals in 2D and hexahedra in 3D, Condition A can be satisfied if we use the space of piecewise polynomials of degree k in each direction given by

$$\tilde{\mathcal{D}}_k(\mathcal{E}_h) = \{v \in L^2(\Omega) : \forall E \in \mathcal{E}_h,\, v|_E \in \mathbb{Q}_k(E)\}.$$

Indeed, one can construct a continuous interpolant of p in the space $\tilde{\mathcal{D}}_k(\mathcal{E}_h)$. The benefit of using $\tilde{\mathcal{D}}_k(\mathcal{E}_h)$ rather than $\mathcal{D}_k(\mathcal{E}_h)$ is that optimal L^2 error estimates are obtained if superpenalization is used. The drawback is that the method is more expensive as the total number of degrees of freedom increases. Therefore, if one does not want superpenalization (it is known that increasing the power β_0 worsens the condition number of the global matrix), then one should use $\mathcal{D}_k(\mathcal{E}_h)$ for all meshes.

2.9 Implementing the DG method

There is more than one way to write a DG code. Our preferred choice is to use a parent-child data structure. This allows for an easy implementation of local mesh refinement and derefinement. In this section, we first present the data structure and then discuss the construction of the local and global matrices. For simplicity, we will assume that $\boldsymbol{K}$ is piecewise constant.

2.9.1 Data structure

A parent-child data structure uses a list of elements, faces, and vertices. It is understood that for two-dimensional problems, an edge is called a face. We assume that a given element has M_F faces and that each interior face belongs to two elements. If an element is refined, it has at most M_C children. We also denote by M_V the number of vertices of one face. Attributes of the elements and faces are given in Table 2.2. Those attributes contain the information that is being stored. One can choose to either store more information or recompute information when needed. There is a delicate balance between the amount of storage and the amount of computation that will yield a minimum simulation time. In the programming language C, we can take advantage of the structure data type to store the attributes. For instance, for a triangular mesh, we give below the definition of the structures `element`, `face`, and `vertex` and arrays of those particular data types.

Table 2.2. *Attributes of elements and faces for the data structure.*

Object	Attributes	Definition
element	face	array of M_F components: global number of faces
	parent	integer: global number of parents
	child	array of M_C components: global number of children
	degree	integer: polynomial degree
	reftype	integer: -1 for inactive (not refined) element 0 for active (refined) element
	soldofs	array of N_{loc} components: local degrees of freedom
face	vertex	array of M_V components: global number of vertices
	neighbor	array of 2 components: global number of elements sharing the face
	reftype	integer: -1 for inactive (not refined) face 0 for active (refined) face
	bctype	integer: 0 for interior face 1 for Dirichlet face 2 for Neumann face
vertex	coor	array: coordinates of the vertex

```
typedef struct {
    int    face[3];
    int    parent;
    int    child[4];
    int    degree;
    int    reftype;
    double *soldofs;
} element;

typedef struct {
    int    vertex[2];
    int    neighbor[2];
    int    reftype;
    int    bctype;
} face;

typedef struct {
    double coor[2];
} vertex;

element meshelt[100];
face meshface[300];
vertex meshvertex[300];
```

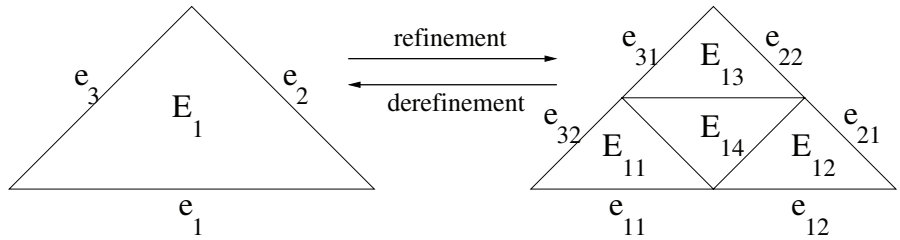

Figure 2.4. *Example of a refinement/derefinement strategy for a triangular element.*

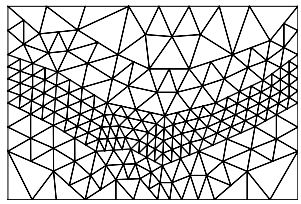

Figure 2.5. *Example of a nonconforming mesh.*

If the mesh uses quadrilateral elements, it suffices to increase the size of the attribute `face` to four entries. Fig. 2.4 shows an example of refinement/derefinement of a triangle. In this case, the element has $M_F = 3$ faces and $M_C = 4$ children, and each face has $M_V = 2$ vertices. The children of the element E_1 are the elements E_{11}, E_{12}, E_{13}, and E_{14}. New faces corresponding to the refinement of the faces e_1, e_2, e_3 are created. For example, the children of face e_1 are the faces e_{11} and e_{12}. Once the element and faces are refined, they become "inactive." The inverse process, also called derefinement, changes the "inactive" state of the parents to "active" and vice versa for the children. Note that the parents of the elements in the coarsest mesh do not exist and by default can be set to zero. With the DG method, it is possible to refine a few elements in the mesh as many times as possible without worrying about refining their neighbors. The resulting mesh is called nonconforming. An example of a nonconforming mesh is given in Fig. 2.5.

2.9.2 Local matrices and right-hand sides

There are two types of local matrices depending on the domain of integration. First, we compute the matrix $\boldsymbol{A}_E$ resulting from the volume integral over a fixed element E. We recall (see Section 2.5.2) that the local basis functions $\phi_{i,E}$ are obtained from mapping monomial functions $\hat{\phi}_i$ from the reference element $\hat{E}$ onto the element E:

$$\phi_{i,E} = \hat{\phi}_i \circ F_E^{-1}.$$

Then, we have

$$\forall 1 \le i, j \le N_{\text{loc}}, \quad (A_E)_{i,j} = \int_E (\boldsymbol{K}\nabla\phi_{i,E} \cdot \nabla\phi_{j,E} + \alpha\phi_{i,E}\phi_{j,E}).$$

Applying a change of variable with the mapping F_E (see (2.30) and (2.32)), we can compute the integral on the reference element:

$$(A_E)_{i,j} = 2|E| \int_{\hat{E}} (\boldsymbol{K}(\boldsymbol{B}_E^T)^{-1}\hat{\nabla}\hat{\phi}_i \cdot (\boldsymbol{B}_E^T)^{-1}\hat{\nabla}\hat{\phi}_j + (\alpha \circ F_E)\hat{\phi}_i\hat{\phi}_j).$$

The volume contributions to the local right-hand side $\boldsymbol{b}_E$ are

$$(\boldsymbol{b}_E)_i = \int_E f\phi_{i,E}.$$

We now compute the local matrices corresponding to the integrals over a fixed face e. If e is an interior face, let us denote by E_e^1 and E_e^2 the elements that share the face such that the normal vector $\boldsymbol{n}_e$ points from E_e^1 to E_e^2. The terms involving integrals on e in the bilinear form a_ϵ are recalled below:

$$T = -\int_e \{\boldsymbol{K}\nabla P_h \cdot \boldsymbol{n}_e\}[v] + \epsilon \int_e \{\boldsymbol{K}\nabla v \cdot \boldsymbol{n}_e\}[P_h] + \frac{\sigma_e^0}{|e|^{\beta_0}} \int_e [P_h][v].$$

Denoting by $P_{h,i}$ and v_i the restrictions of P_h and v to the element E_i and expanding the averages and jumps, we obtain

$$T = m_e^{11} + m_e^{22} + m_e^{12} + m_e^{21},$$

where the term m_e^{11} (resp., m_e^{22}) corresponds to the interactions of the local basis of the neighboring element E_e^1 (resp., E_e^2) with itself and the term m_e^{12} (resp., m_e^{21}) corresponds to the interactions of the local basis of the neighboring element E_e^1 (resp., E_e^2) with the element E_e^2 (resp., E_e^1). More precisely, we have the expressions

$$\begin{aligned}
m_e^{11} &= -\frac{1}{2}\int_e \boldsymbol{K}\nabla P_{h,1} \cdot \boldsymbol{n}_e v_1 + \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla v_1 \cdot \boldsymbol{n}_e P_{h,1} + \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e P_{h,1}v_1,\\
m_e^{22} &= \frac{1}{2}\int_e \boldsymbol{K}\nabla P_{h,2} \cdot \boldsymbol{n}_e v_2 - \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla v_2 \cdot \boldsymbol{n}_e P_{h,2} + \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e P_{h,2}v_2,\\
m_e^{12} &= -\frac{1}{2}\int_e \boldsymbol{K}\nabla P_{h,2} \cdot \boldsymbol{n}_e v_1 - \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla v_1 \cdot \boldsymbol{n}_e P_{h,2} - \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e P_{h,2}v_1,\\
m_e^{21} &= -\frac{1}{2}\int_e \boldsymbol{K}\nabla P_{h,1} \cdot \boldsymbol{n}_e v_2 + \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla v_2 \cdot \boldsymbol{n}_e P_{h,1} - \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e P_{h,1}v_2.
\end{aligned}$$

These four terms will yield four matrices of size $N_{\text{loc}} \times N_{\text{loc}}$, namely $\boldsymbol{M}_e^{11}$, $\boldsymbol{M}_e^{22}$, $\boldsymbol{M}_e^{12}$, $\boldsymbol{M}_e^{21}$, whose entries are defined below:

$$\begin{aligned}
(\boldsymbol{M}_e^{11})_{ij} &= -\frac{1}{2}\int_e \boldsymbol{K}\nabla\phi_{j,E_e^1} \cdot \boldsymbol{n}_e\phi_{i,E_e^1} + \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla\phi_{i,E_e^1} \cdot \boldsymbol{n}_e\phi_{j,E_e^1} + \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e \phi_{j,E_e^1}\phi_{i,E_e^1},\\
(\boldsymbol{M}_e^{22})_{ij} &= \frac{1}{2}\int_e \boldsymbol{K}\nabla\phi_{j,E_e^2} \cdot \boldsymbol{n}_e\phi_{i,E_e^2} - \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla\phi_{i,E_e^2} \cdot \boldsymbol{n}_e\phi_{j,E_e^2} + \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e \phi_{j,E_e^2}\phi_{i,E_e^2},
\end{aligned}$$

$$(\boldsymbol{M}_e^{12})_{ij} = -\frac{1}{2}\int_e \boldsymbol{K}\nabla\phi_{j,E_e^2}\cdot\boldsymbol{n}_e\phi_{i,E_e^1} - \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla\phi_{i,E_e^1}\cdot\boldsymbol{n}_e\phi_{j,E_e^2} - \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e \phi_{j,E_e^2}\phi_{i,E_e^1},$$
$$(\boldsymbol{M}_e^{21})_{ij} = -\frac{1}{2}\int_e \boldsymbol{K}\nabla\phi_{j,E_e^1}\cdot\boldsymbol{n}_e\phi_{i,E_e^2} + \frac{\epsilon}{2}\int_e \boldsymbol{K}\nabla\phi_{i,E_e^2}\cdot\boldsymbol{n}_e\phi_{j,E_e^1} - \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e \phi_{j,E_e^1}\phi_{i,E_e^2}.$$

Next, if e is a boundary face, let us also denote by E_e^1 the element to which it belongs. If a Dirichlet boundary condition is applied on e, the following local matrix $\boldsymbol{M}_e^{11}$ is created:

$$(\boldsymbol{M}_e^{11})_{ij} = -\int_e \boldsymbol{K}\nabla\phi_{j,E_e^1}\cdot\boldsymbol{n}_e\phi_{i,E_e^1} + \epsilon\int_e \boldsymbol{K}\nabla\phi_{i,E_e^1}\cdot\boldsymbol{n}_e\phi_{j,E_e^1} + \frac{\sigma_e^0}{|e|^{\beta_0}}\int_e \phi_{j,E_e^1}\phi_{i,E_e^1},$$

and the local right-hand side $\boldsymbol{b}_e$ is

$$(\boldsymbol{b}_e)_i = \epsilon\int_e \left(\boldsymbol{K}\nabla\phi_{i,E_e^1}\cdot\boldsymbol{n}_e + \frac{\sigma_e^0}{|e|^{\beta_0}}\phi_{i,E_e^1}\right) g_{\mathrm{D}}.$$

If the edge e is a Neumann boundary edge, no local matrix is created, but the following local right-hand side is defined:

$$(\boldsymbol{b}_e)_i = \int_e \phi_{i,E_e^1} g_{\mathrm{N}}.$$

As usual, all integrals on the physical face are transformed into integrals on the reference element in the space $\mathbb{R}^{d-1}$.

We now present the algorithm for computing the local matrices and the local right-hand sides.

Algorithm 2.1.
Computing local contributions from element E
initialize $\boldsymbol{A}_E = \boldsymbol{0}$
initialize the quadrature weights $\boldsymbol{w}$ and points $\boldsymbol{s}$
loop over quadrature points: for $k = 1$ to N_{G} do
    compute Jacobian matrix $\boldsymbol{B}_E$
    for $i = 1$ to N_{loc} do
        compute values of basis function $\phi_{i,E}(\boldsymbol{s}(k))$
        compute derivatives of basis functions $\nabla\phi_{i,E}(\boldsymbol{s}(k))$
    end
    compute global coordinates $\boldsymbol{x}$ of quadrature point $\boldsymbol{s}(k)$
    compute source function $f(\boldsymbol{x})$
    for $i = 1$ to N_{loc} do
        for $j = 1$ to N_{loc} do
            $\boldsymbol{A}_E(i,j) = \boldsymbol{A}_E(i,j) + \boldsymbol{w}(k)\det(\boldsymbol{B}_E)\nabla\phi_{i,E}(\boldsymbol{s}(k))\cdot\nabla\phi_{j,E}(\boldsymbol{s}(k))$
        end
        $\boldsymbol{b}_E(i) = \boldsymbol{b}_E(i) + \boldsymbol{w}(k)\det(\boldsymbol{b}_E)f(\boldsymbol{x})\phi_{i,E}(\boldsymbol{s}(k))$
    end
end

The next algorithm computes the local stiffness matrices obtained by the integration over one interior edge shared by two elements. We recall that the choice of the method is defined by the parameters ϵ and σ_e^0.

ALGORITHM 2.2.
Computing local contributions from edge e
initialize $\boldsymbol{M}_e^{11} = \boldsymbol{M}_e^{22} = \boldsymbol{M}_e^{12} = \boldsymbol{M}_e^{21} = \boldsymbol{0}$
initialize parameters ϵ and σ_e^0
initialize the quadrature weights $\boldsymbol{w}$ and points $\boldsymbol{s}$
compute edge length $|e|$
compute normal vector $\boldsymbol{n}_e$
get face neighbors E_e^1 and E_e^2
loop over quadrature points: for $k = 1$ to N_{G} do
    compute Jacobian matrices $\boldsymbol{M}_{E_e^1}$ and $\boldsymbol{M}_{E_e^2}$
    for $i = 1$ to N_{loc} do
        compute values of basis functions $\phi_{i,E_e^1}(\boldsymbol{s}(k))$ and $\phi_{i,E_e^2}(\boldsymbol{s}(k))$
        compute derivatives of basis functions $\nabla\phi_{i,E_e^1}(\boldsymbol{s}(k))$ and $\nabla\phi_{i,E_e^2}(\boldsymbol{s}(k))$
    end
    compute $\boldsymbol{M}_k^{11}$ contributions:
        for $i = 1$ to N_{loc} do
            for $j = 1$ to N_{loc} do
                $\boldsymbol{M}_e^{11}(i,j) = \boldsymbol{M}_e^{11}(i,j) - 0.5\boldsymbol{w}(k)|e|\phi_{i,E_e^1}(\boldsymbol{s}(k))(\nabla\phi_{j,E_e^1}(\boldsymbol{s}(k)) \cdot \boldsymbol{n}_e)$
                $\boldsymbol{M}_e^{11}(i,j) = \boldsymbol{M}_e^{11}(i,j) + 0.5\epsilon\boldsymbol{w}(k)|e|\phi_{j,E_e^1}(\boldsymbol{s}(k))(\nabla\phi_{i,E_e^1}(\boldsymbol{s}(k)) \cdot \boldsymbol{n}_e)$
                $\boldsymbol{M}_e^{11}(i,j) = \boldsymbol{M}_e^{11}(i,j) + \sigma_e^0\boldsymbol{w}(k)\phi_{i,E_e^1}(\boldsymbol{s}(k))\phi_{j,E_e^1}(\boldsymbol{s}(k))$
            end
        end
    compute $\boldsymbol{M}_k^{22}$ contributions:
        for $i = 1$ to N_{loc} do
            for $j = 1$ to N_{loc} do
                $\boldsymbol{M}_e^{22}(i,j) = \boldsymbol{M}_e^{22}(i,j) + 0.5\boldsymbol{w}(k)|e|\phi_{i,E_e^2}(\boldsymbol{s}(k))(\nabla\phi_{j,E_e^2}(\boldsymbol{s}(k)) \cdot \boldsymbol{n}_e)$
                $\boldsymbol{M}_e^{22}(i,j) = \boldsymbol{M}_e^{22}(i,j) - 0.5\epsilon\boldsymbol{w}(k)|e|\phi_{j,E_e^2}(\boldsymbol{s}(k))(\nabla\phi_{i,E_e^2}(\boldsymbol{s}(k)) \cdot \boldsymbol{n}_e)$
                $\boldsymbol{M}_e^{22}(i,j) = \boldsymbol{M}_e^{22}(i,j) + \sigma_e^0\boldsymbol{w}(k)\phi_{i,E_e^2}(\boldsymbol{s}(k))\phi_{j,E_e^2}(\boldsymbol{s}(k))$
            end
        end
    compute $\boldsymbol{M}_k^{12}$ contributions:
        for $i = 1$ to N_{loc} do
            for $j = 1$ to N_{loc} do
                $\boldsymbol{M}_e^{12}(i,j) = \boldsymbol{M}_e^{12}(i,j) - 0.5\boldsymbol{w}(k)|e|\phi_{i,E_e^1}(\boldsymbol{s}(k))(\nabla\phi_{j,E_e^2}(\boldsymbol{s}(k)) \cdot \boldsymbol{n}_e)$
                $\boldsymbol{M}_e^{12}(i,j) = \boldsymbol{M}_e^{12}(i,j) - 0.5\epsilon\boldsymbol{w}(k)|e|\phi_{j,E_e^2}(\boldsymbol{s}(k))(\nabla\phi_{i,E_e^1}(\boldsymbol{s}(k)) \cdot \boldsymbol{n}_e)$
                $\boldsymbol{M}_e^{12}(i,j) = \boldsymbol{M}_e^{12}(i,j) - \sigma_e^0\boldsymbol{w}(k)\phi_{i,E_e^1}(\boldsymbol{s}(k))\phi_{j,E_e^2}(\boldsymbol{s}(k))$
            end
        end
    compute $\boldsymbol{M}_k^{21}$ contributions:
        for $i = 1$ to N_{loc} do

```
        for j = 1 to N_loc do
            M_e^21(i, j) = M_e^21(i, j) + 0.5 w(k)|e| φ_{i,E_e^2}(s(k)) (∇φ_{j,E_e^1}(s(k)) · n_e)
            M_e^21(i, j) = M_e^21(i, j) + 0.5 ε w(k)|e| φ_{j,E_e^1}(s(k)) (∇φ_{i,E_e^2}(s(k)) · n_e)
            M_e^21(i, j) = M_e^21(i, j) − σ_e^0 w(k) φ_{i,E_e^1}(s(k)) φ_{j,E_e^1}(s(k))
        end
    end
end
```

The corresponding C routines are given in Appendix B.2.

2.9.3 Global matrix and right-hand side

Assembling of the global matrix $\boldsymbol{A}_{\text{global}}$ is done in two steps. First, the local matrices $\boldsymbol{A}_E$ are added to the block diagonal entries of $\boldsymbol{A}_{\text{global}}$. We can assume that the mesh elements are numbered from 1 to N_{el}. We denote the global right-hand side by $\boldsymbol{b}_{\text{global}}$. The local contributions $\boldsymbol{b}_E$ can be added to $\boldsymbol{b}_{\text{global}}$ in the same algorithm.

Algorithm 2.3.
Volume contributions

```
initialize k = 0
loop over the elements: for k = 1 to N_el do
    compute local volume matrix A_{E_k}
    compute local right-hand side b_{E_k}
    for i = 1 to N_el do
        ie = i + k
        b_global(ie) = b_global(ie) + b_{E_k}(i)
        for j = 1 to N_loc do
            je = j + k
            A_global(ie, je) = A_global(ie, je) + A_{E_k}(i, j)
        end
        k = k + N_loc
    end
end
```

Second, we assemble the local matrices $\boldsymbol{M}_e^{ij}$ for $1 \le i, j \le 2$. We can assume that the edges are numbered from 1 to N_{face}. The numbers of the neighboring elements of the face k are E_1^k and E_2^k.

Algorithm 2.4.
Face contributions

```
loop over the edges: for k = 1 to N_face do
    get face neighbors E_k^1 and E_k^2
    if face is an interior face do
    compute local matrices M_k^11, M_k^22, M_k^12, M_k^21
```

```
    assemble M_k^11 contributions:
        for i = 1 to N_loc do
            ie = i + (E_k^1 − 1)N_loc
            for j = 1 to N_loc do
                je = j + (E_k^1 − 1)N_loc
                A_global(ie, je) = A_global(ie, je) + M_k^11(i, j)
            end
        end
    assemble M_k^22 contributions:
        for i = 1 to N_loc do
            ie = i + (E_k^2 − 1)N_loc
            for j = 1 to N_loc do
                je = j + (E_k^2 − 1)N_loc
                A_global(ie, je) = A_global(ie, je) + M_k^22(i, j)
            end
        end
    assemble M_k^12 contributions:
        for i = 1 to N_loc do
            ie = i + (E_k^1 − 1)N_loc
            for j = 1 to N_loc do
                je = j + (E_k^2 − 1)N_loc
                A_global(ie, je) = A_global(ie, je) + M_k^12(i, j)
            end
        end
    assemble M_k^21 contributions:
        for i = 1 to N_loc do
            ie = i + (E_k^2 − 1)N_loc
            for j = 1 to N_loc do
                je = j + (E_k^1 − 1)N_loc
                A_global(ie, je) = A_global(ie, je) + M_k^11(i, j)
            end
        end
    end
    else if face is a boundary face do
    compute local matrix M_k^11
    assemble M_k^11 contributions:
        for i = 1 to N_loc do
            ie = i + (E_k^1 − 1)N_loc
            for j = 1 to N_loc do
                je = j + (E_k^1 − 1)N_loc
                A_global(ie, je) = A_global(ie, je) + M_k^11(i, j)
            end
        end
    end
end
```

The corresponding C routines are given in Appendix B.2.

2.10 Numerical experiments

We solve on the unit square the model problem (2.16)–(2.18) with $\boldsymbol{K} = \mathbf{I}$, $\alpha = 0$, and Dirichlet boundary condition ($\Gamma_{\mathrm{D}} = \partial\Omega$). We present numerical convergence rates for both smooth and unsmooth exact solutions and for all primal DG methods. We vary the polynomial degree from $k = 1, 2, 3$. We denote the numerical error by

$$e_h = p - P_h.$$

We compute the seminorm $|||\nabla e_h|||_{H^0(\mathcal{E}_h)}$, which is bounded above by the energy norm, and the L^2 norm $\|e_h\|_{L^2(\Omega)}$. The penalty parameter σ_e^0 is equal to a constant σ for all interior edges. For the boundary edges, the penalty parameter is equal to 2σ for both IIPG and SIPG and equal to σ for NIPG.

2.10.1 Smooth solution

Let the exact solution be

$$\forall (x, y) \in (0, 1)^2, \quad p(x, y) = e^{-x-y^2}.$$

Table 2.3 contains the numerical errors $|||\nabla e_h|||_{H^0(\mathcal{E}_h)}$ and $\|e_h\|_{L^2(\Omega)}$ obtained on a fine triangular mesh. Convergence rates are computed as in (1.13). We choose $\beta_0 = 1$. The rates correspond to the theoretical rates: they are all optimal in the gradient broken norm $|||\nabla e_h|||_{H^0(\mathcal{E}_h)} = \mathcal{O}(h^k)$. The L^2 rates are optimal for the SIPG method: $\|e_h\|_{L^2(\Omega)} = \mathcal{O}(h^{k+1})$. For NIPG or IIPG, they are suboptimal if the polynomial degree is even.

Next, we use superpenalization and choose $\beta_0 = 3$. We consider a different smooth solution such that its Dirichlet value is a polynomial of degree k. The exact solution is given by

$$\forall (x, y) \in (0, 1)^2, \quad p(x, y) = x(x-1)y(y-1)e^{-x^2-y^2}.$$

Table 2.3. *Numerical errors and convergence rates for smooth function without superpenalization.*

Method	k	σ	$\|\|\|\nabla e_h\|\|\|_{H^0(\mathcal{E}_h)}$	Rate	$\|e_h\|_{L^2(\Omega)}$	Rate
NIPG	1	1	8.4846×10^{-3}	1.0123	8.9099×10^{-5}	2.0083
	2	1	7.6614×10^{-5}	2.0011	1.8632×10^{-6}	2.0186
	3	1	4.1740×10^{-7}	3.0157	3.3112×10^{-9}	4.0153
NIPG	2	0	8.3851×10^{-5}	2.0035	1.7316×10^{-6}	2.0307
	3	0	4.9857×10^{-7}	3.0103	3.8794×10^{-9}	4.0036
SIPG	1	6	8.9986×10^{-3}	1.0007	3.9981×10^{-5}	1.9717
	2	18	7.3139×10^{-5}	2.0009	1.5827×10^{-7}	2.9942
	3	36	3.8845×10^{-7}	3.0044	1.4746×10^{-9}	3.9879
IIPG	1	6	8.9885×10^{-3}	0.9996	3.2571×10^{-5}	1.9994
	2	18	7.1979×10^{-5}	2.0014	2.7825×10^{-7}	2.4695
	3	36	3.8427×10^{-7}	3.0023	1.5009×10^{-9}	3.9921

Table 2.4. *Numerical errors and convergence rates for smooth function with superpenalization.*

Method	k	σ	$\lVert\lvert\nabla e_h\rvert\rVert_{H^0(\mathcal{E}_h)}$	Rate	$\lVert e_h\rVert_{L^2(\Omega)}$	Rate
NIPG	1	1	5.1010×10^{-3}	0.9872	6.1576×10^{-5}	1.9537
	2	1	9.8300×10^{-5}	1.9707	3.7058×10^{-7}	3.1578
	3	1	8.5460×10^{-7}	2.9787	4.6797×10^{-9}	4.0106
IIPG	1	6	5.1107×10^{-3}	0.9959	6.2081×10^{-5}	1.9893
	2	18	9.8839×10^{-5}	1.9951	3.5405×10^{-7}	3.0000
	3	36	8.6042×10^{-7}	3.0135	4.6953×10^{-9}	4.0230

Therefore, Condition A is satisfied. Table 2.4 shows the numerical errors and convergence rates for NIPG and IIPG. The rates are optimal for the L^2 norm, as predicted by the theory.

2.10.2 Singular solution

We consider a solution $p \in H^{1+\delta}(\Omega)$ with $0 < \delta < 1$. Consider a domain $\Omega = (-1, 1)^2$ subdivided into four subdomains Ω_i such that $\Omega_1 = (0, 1)^2$, $\Omega_2 = (-1, 0) \times (0, 1)$, $\Omega_3 = (-1, 0)^2$, and $\Omega_4 = (0, 1) \times (-1, 0)$. We solve (2.16)–(2.18) with $\alpha = 0$, $f = 0$, and $\Gamma_{\mathrm{D}} = \partial\Omega$. The coefficient matrix $\boldsymbol{K}$ is equal to $K_i \boldsymbol{I}$ on each subdomain Ω_i. We assume that $K_1 = K_3 = 5$ and $K_2 = K_4 = 1$. The exact solution in polar coordinates is

$$p(r, \theta) = r^{\delta}(a_i \sin(\delta\theta) + b_i \cos(\delta\theta)) \quad \text{in} \quad \Omega_i$$

with coefficients given up to nine accurate digits:

$$\begin{aligned}
a_1 &= 0.4472135955,\\
a_2 &= -0.7453559925,\\
a_3 &= -0.9441175905,\\
a_4 &= -2.401702643,\\
b_1 &= 1,\\
b_2 &= 2.333333333,\\
b_3 &= 0.55555555555,\\
b_4 &= -0.4814814814,\\
\delta &= 0.5354409456.
\end{aligned}$$

The exact solution is singular at the origin in the sense that its gradient is not defined at the point $(0, 0)$. We compute the DG solution on a sequence of uniformly refined rectangular meshes. The relative error in the L^2 norm, defined as $\frac{\lVert p - P_h\rVert_{L^2(\Omega)}}{\lVert p\rVert_{L^2(\Omega)}}$, is plotted

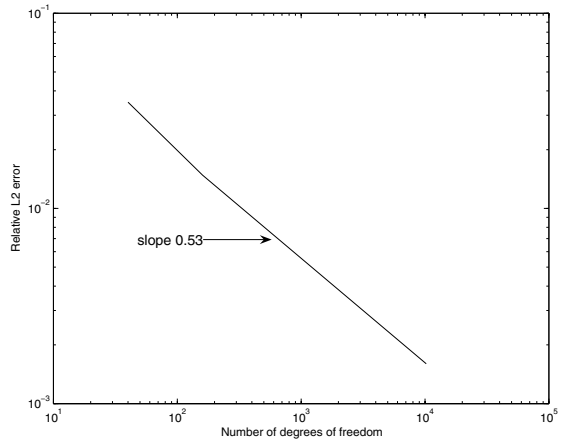

Figure 2.6. *Relative error in the L^2 norm versus the number of degrees of freedom.*

against the number of degrees of freedom in Fig. 2.6. We use the NIPG method without penalty and polynomials of degree three. We see that the convergence rate is independent of the polynomial order. This is expected, as the solution has poor regularity. Indeed, since p belongs to $H^{1+\alpha}(\Omega)$, the convergence rate in the L^2 norm is $\mathcal{O}(h^{2\alpha})$, or equivalently $\mathcal{O}(N^{\alpha})$, where N is the total number of degrees of freedom. In order to recover the rate obtained with the polynomial degrees, we need to locally refine the mesh around the origin.

2.10.3 Condition number

We fix the polynomial degree $k = 2$ and compute the condition number $\|\boldsymbol{A}\| \, \|\boldsymbol{A}^{-1}\|$ of the global matrix for the DG method with and without superpenalization. Fig. 2.7 shows that if no superpenalization is used, $\beta = 1$, then the condition number is $\mathcal{O}(h^2)$, whereas it is $\mathcal{O}(h^4)$ if $\beta = 3$. The method used here is NIPG with $\sigma_e^0 = 1$. Similar results are observed with SIPG and IIPG methods.

2.11 The local discontinuous Galerkin method

The local discontinuous Galerkin (LDG) method was introduced by Cockburn and Shu [36] and is based on the work by Bassi and Rebay [12]. We present the method for the model problem (2.16)–(2.18) with $\boldsymbol{K} = \mathbf{I}$ and $\alpha = 0$. Because this method solves for two unknowns, namely the solution and its gradient, it can be called a *dual* DG method or a *mixed* DG method.

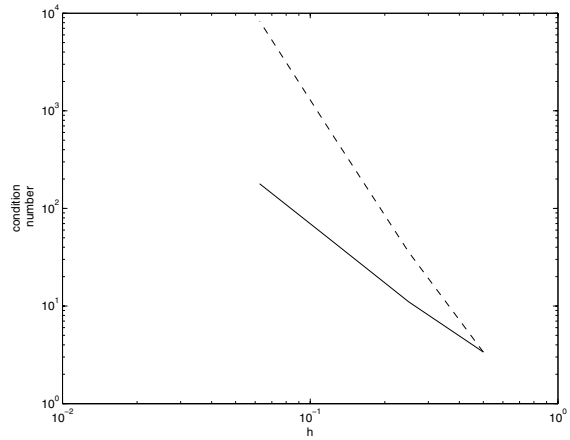

Figure 2.7. *Condition number versus mesh size for NIPG* 1: $\beta = 1$ *(solid line) and* $\beta = 3$ *(dashed line).*

2.11.1 Definition of the mixed DG method

Let us rewrite the model problem into a mixed form by introducing an auxiliary variable $\boldsymbol{u}$ for the gradient of the solution:

$$\boldsymbol{u} = \nabla p \quad \text{in} \quad \Omega, \tag{2.48}$$

$$-\nabla \cdot \boldsymbol{u} = f \quad \text{in} \quad \Omega. \tag{2.49}$$

The Dirichlet and Neumann boundary conditions are rewritten as

$$p = g_{\mathrm{D}} \quad \text{on} \quad \Gamma_{\mathrm{D}}, \tag{2.50}$$

$$\boldsymbol{u} \cdot \boldsymbol{n} = \boldsymbol{g} \cdot \boldsymbol{n} \quad \text{on} \quad \Gamma_{\mathrm{N}}. \tag{2.51}$$

Let $\mathcal{E}_h$ be a subdivision of Ω and let Γ_h be the set of interior edges (or faces). Let $\boldsymbol{v} \in H^1(\mathcal{E}_h)^d$ and let $q \in H^1(\mathcal{E}_h)$. We multiply (2.48) and (2.49) by $\boldsymbol{v}$ and q, integrate over one element $E \in \mathcal{E}_h$, and use Green's theorem (2.13):

$$\int_E \boldsymbol{u} \cdot \boldsymbol{v} = -\int_E p \nabla \cdot \boldsymbol{v} + \int_{\partial E} p \boldsymbol{v} \cdot \boldsymbol{n}_E,$$

$$\int_E \boldsymbol{u} \cdot \nabla q - \int_{\partial E} \boldsymbol{u} \cdot \boldsymbol{n}_E q = \int_E f q.$$

We look for a solution pair $(\boldsymbol{U}_h, P_h)$ that belongs to a finite-dimensional space $M_h^d \times M_h$, to be specified later, that satisfies for all $E \in \mathcal{E}_h$

$$\forall \boldsymbol{v} \in M_h^d, \quad \int_E \boldsymbol{U}_h \cdot \boldsymbol{v} + \int_E P_h \nabla \cdot \boldsymbol{v} = \int_{\partial E} \Phi(P_h) \boldsymbol{v} \cdot \boldsymbol{n}_E, \tag{2.52}$$

$$\forall q \in M_h, \quad \int_E \boldsymbol{U}_h \cdot \nabla q = \int_E f q + \int_{\partial E} \Psi(\boldsymbol{U}_h) \cdot \boldsymbol{n}_E q, \tag{2.53}$$

where $\Phi(P_h)$ and $\Psi(\boldsymbol{U}_h)$ are called numerical fluxes and they are defined below. Given two real numbers δ_1, δ_2 and a vector $\boldsymbol{\delta}_3 \in \mathbb{R}^d$, we define

$$\begin{aligned}
\forall e \in \Gamma_h, \quad & \Psi(\boldsymbol{U}_h)|_e = \{\boldsymbol{U}_h\} - (\delta_1[P_h])\boldsymbol{n}_e - ([\boldsymbol{U}_h]\cdot\boldsymbol{n}_e)\boldsymbol{\delta}_3,\\
\forall e \in \Gamma_h, \quad & \Phi(P_h)|_e = \{P_h\} + \boldsymbol{\delta}_3\cdot\boldsymbol{n}_e[P_h] - \delta_2[\boldsymbol{U}_h]\cdot\boldsymbol{n}_e,\\
\forall e \in \Gamma_{\mathrm{D}}, \quad & \Psi(\boldsymbol{U}_h)|_e = \boldsymbol{U}_h - \delta_1(P_h - g_{\mathrm{D}})\boldsymbol{n}_e,\\
\forall e \in \Gamma_{\mathrm{D}}, \quad & \Phi(P_h)|_e = g_{\mathrm{D}},\\
\forall e \in \Gamma_{\mathrm{N}}, \quad & \Psi(\boldsymbol{U}_h)|_e = \boldsymbol{g},\\
\forall e \in \Gamma_{\mathrm{N}}, \quad & \Phi(P_h)|_e = P_h - \delta_2(\boldsymbol{U}_h - \boldsymbol{g})\cdot\boldsymbol{n}.
\end{aligned}$$

We note that the scheme is consistent because of the regularity of the exact solution; the numerical fluxes are equal to the exact fluxes. More precisely, we have

$$\forall e, \quad \Phi(p)|_e = p|_e, \quad \Psi(\boldsymbol{u})|_e = \boldsymbol{u}|_e.$$

By summing (2.52) and (2.53) over all the elements, we obtain

$$\begin{aligned}
\int_\Omega \boldsymbol{U}_h\cdot\boldsymbol{v} + \sum_{E\in\mathcal{E}_h}\int_E P_h\nabla\cdot\boldsymbol{v} &= \sum_{E\in\mathcal{E}_h}\int_{\partial E}\Phi(P_h)\boldsymbol{v}\cdot\boldsymbol{n}_E\\
= \sum_{e\in\Gamma_h}\int_e(\{P_h\} + \boldsymbol{\delta}_3\cdot\boldsymbol{n}_e[P_h] - \delta_2[\boldsymbol{U}_h]\cdot\boldsymbol{n}_e)[\boldsymbol{v}]\cdot\boldsymbol{n}_e &+ \int_{\Gamma_{\mathrm{D}}} g_{\mathrm{D}}\boldsymbol{v}\cdot\boldsymbol{n}\\
&+ \int_{\Gamma_{\mathrm{N}}}(P_h - \delta_2(\boldsymbol{U}_h - \boldsymbol{g})\cdot\boldsymbol{n})\boldsymbol{v}\cdot\boldsymbol{n}
\end{aligned}$$

and

$$\begin{aligned}
\sum_{E\in\mathcal{E}_h}\int_E \boldsymbol{U}_h\cdot\nabla q &= \int_\Omega fq + \sum_{E\in\mathcal{E}_h}\int_{\partial E}\Psi(\boldsymbol{U}_h)\cdot\boldsymbol{n}_E q\\
= \int_\Omega fq + \sum_{e\in\Gamma_h}\int_e(\{\boldsymbol{U}_h\} - (\delta_1[P_h])\boldsymbol{n}_e - ([\boldsymbol{U}_h]\cdot\boldsymbol{n}_e)\boldsymbol{\delta}_3)\cdot\boldsymbol{n}_e[q]&\\
+ \int_{\Gamma_{\mathrm{D}}}(\boldsymbol{U}_h - \delta_1(P_h - g_{\mathrm{D}})\boldsymbol{n}_e)\cdot\boldsymbol{n}_e q + \int_{\Gamma_{\mathrm{N}}}\boldsymbol{g}\cdot\boldsymbol{n}q&\\
= \int_\Omega fq + \sum_{e\in\Gamma_h}\int_e(\{\boldsymbol{U}_h\}\cdot\boldsymbol{n}_e - (\delta_1[P_h]) - ([\boldsymbol{U}_h]\cdot\boldsymbol{n}_e)\boldsymbol{\delta}_3\cdot\boldsymbol{n}_e)[q]&\\
+ \int_{\Gamma_{\mathrm{D}}}(\boldsymbol{U}_h\cdot\boldsymbol{n}_e - \delta_1(P_h - g_{\mathrm{D}}))q + \int_{\Gamma_{\mathrm{N}}}(\boldsymbol{g}\cdot\boldsymbol{n})q.&
\end{aligned}$$

Let us define the following bilinear forms:

$$\begin{aligned}
a_{\mathrm{ldg}}(\boldsymbol{U}_h, \boldsymbol{v}) &= \int_\Omega \boldsymbol{U}_h\cdot\boldsymbol{v} + \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{N}}}\int_e \delta_2[\boldsymbol{U}_h]\cdot\boldsymbol{n}_e[\boldsymbol{v}]\cdot\boldsymbol{n}_e,\\
b_{\mathrm{ldg}}(P_h, \boldsymbol{v}) &= \sum_{E\in\mathcal{E}_h}\int_E P_h\nabla\cdot\boldsymbol{v} - \sum_{e\in\Gamma_h}\int_e(\{P_h\} + \boldsymbol{\delta}_3\cdot\boldsymbol{n}_e[P_h])[\boldsymbol{v}]\cdot\boldsymbol{n}_e - \int_{\Gamma_{\mathrm{N}}} P_h\boldsymbol{v}\cdot\boldsymbol{n},\\
J_{\mathrm{ldg}}(P_h, q) &= \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\int_e \delta_1[P_h][q].
\end{aligned}$$

We remark, by using Green's theorem (2.13), that the form b can be rewritten as

$$\begin{aligned} b_{\mathrm{ldg}}(P_h, \boldsymbol{v}) &= -\sum_{E\in\mathcal{E}_h}\int_E \nabla P_h\cdot\boldsymbol{v} + \sum_{e\in\Gamma_h}\int_e [P_h\boldsymbol{v}\cdot\boldsymbol{n}_e] + \int_{\Gamma_{\mathrm{D}}} P_h\boldsymbol{v}\cdot\boldsymbol{n} \\ &\quad -\sum_{e\in\Gamma_h}\int_e (\{P_h\} + \boldsymbol{\delta}_3\cdot\boldsymbol{n}_e[P_h])[\boldsymbol{v}]\cdot\boldsymbol{n}_e \\ &= -\sum_{E\in\mathcal{E}_h}\int_E \nabla P_h\cdot\boldsymbol{v} + \sum_{e\in\Gamma_h}\int_e [P_h]\{\boldsymbol{v}\cdot\boldsymbol{n}_e\} \\ &\quad + \int_{\Gamma_{\mathrm{D}}} P_h\boldsymbol{v}\cdot\boldsymbol{n} - \sum_{e\in\Gamma_h}\int_e \boldsymbol{\delta}_3\cdot\boldsymbol{n}_e[P_h][\boldsymbol{v}]\cdot\boldsymbol{n}_e; \end{aligned}$$

equivalently,

$$b_{\mathrm{ldg}}(P_h, \boldsymbol{v}) = -\sum_{E\in\mathcal{E}_h}\int_E \nabla P_h\cdot\boldsymbol{v} + \sum_{e\in\Gamma_h}\int_e [P_h](\{\boldsymbol{v}\cdot\boldsymbol{n}_e\} - \boldsymbol{\delta}_3\cdot\boldsymbol{n}_e[\boldsymbol{v}]\cdot\boldsymbol{n}_e) + \int_{\Gamma_{\mathrm{D}}} P_h\boldsymbol{v}\cdot\boldsymbol{n}. \quad (2.54)$$

The discrete space $M_h \subset H^1(\mathcal{E}_h)$ is chosen so that the following two conditions hold:

(i) $\{q \in L^2(\Omega) : \forall E \quad q|_E \in \mathbb{P}_k(E)\} \subset M_h$,

(ii) $\forall E \in \mathcal{E}_h, \forall q \in M_h, (\int_E \nabla q\cdot\boldsymbol{v} = 0 \quad \forall \boldsymbol{v} \in M_h^d) \Longrightarrow \nabla q = 0$.

The global formulation of the general LDG scheme is as follows: Find $\boldsymbol{U}_h \in M_h^d$ and $P_h \in M_h$ such that

$$\forall \boldsymbol{v} \in M_h^d, \quad a_{\mathrm{ldg}}(\boldsymbol{U}_h, \boldsymbol{v}) + b_{\mathrm{ldg}}(P_h, \boldsymbol{v}) = \int_{\Gamma_{\mathrm{D}}} g_{\mathrm{D}}\boldsymbol{v}\cdot\boldsymbol{n} + \int_{\Gamma_{\mathrm{N}}} \delta_2(\boldsymbol{g}\cdot\boldsymbol{n})\boldsymbol{v}\cdot\boldsymbol{n}, \quad (2.55)$$

$$\forall q \in M_h, \quad -b_{\mathrm{ldg}}(q, \boldsymbol{U}_h) + J_{\mathrm{ldg}}(P_h, q) = \int_\Omega f v + \int_{\Gamma_{\mathrm{D}}} \delta_1 g_{\mathrm{D}} q + \int_{\Gamma_{\mathrm{N}}} (\boldsymbol{g}\cdot\boldsymbol{n})q. \quad (2.56)$$

2.11.2 Existence and uniqueness of the solution

Lemma 2.15. *Assume that $\delta_1 > 0$ and $\delta_2 \geq 0$; then there exists a unique solution to the scheme* (2.55)–(2.56).

Proof. Assume that $g_{\mathrm{D}} = f = 0$ and $\boldsymbol{g} = \boldsymbol{0}$. Then, choosing $\boldsymbol{v} = \boldsymbol{U}_h$ in the first equation and $q = P_h$ in the second, we have

$$\begin{aligned} a_{\mathrm{ldg}}(\boldsymbol{U}_h, \boldsymbol{U}_h) + b_{\mathrm{ldg}}(P_h, \boldsymbol{U}_h) &= 0, \\ -b_{\mathrm{ldg}}(P_h, \boldsymbol{U}_h) + J_{\mathrm{ldg}}(P_h, P_h) &= 0. \end{aligned}$$

By adding the two equations above, we obtain

$$a_{\mathrm{ldg}}(\boldsymbol{U}_h, \boldsymbol{U}_h) + J_{\mathrm{ldg}}(P_h, P_h) = 0.$$

Equivalently,

$$\int_\Omega \boldsymbol{U}_h^2 + \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{N}}} \int_e \delta_2([\boldsymbol{U}_h]\cdot \boldsymbol{n}_e)^2 + \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \int_e \delta_1[P_h]^2 = 0.$$

Thus, if $\delta_1 > 0$ and $\delta_2 \geq 0$, we immediately have $\boldsymbol{U}_h = \boldsymbol{0}$, and we have

$$\forall e \in \Gamma_h \cup \Gamma_{\mathrm{D}}, \quad [P_h] = 0.$$

This implies that P_h is continuous across the domain. Then, (2.55) becomes

$$\forall \boldsymbol{v} \in M_h^d, \quad b_{\mathrm{ldg}}(P_h, \boldsymbol{v}) = 0.$$

Using the second form (2.54) of b_{ldg}, we have

$$\forall \boldsymbol{v} \in M_h^d, \quad \sum_{E\in\mathcal{E}_h} \int_E \nabla P_h \cdot \boldsymbol{v} = 0,$$

which implies that $\nabla P_h = 0$ because of the definition of M_h. Since P_h is continuous and zero on Γ_{D}, this implies that $P_h = 0$. $\square$

2.11.3 A priori error estimates

Assume that $p \in H^{s+2}(\Omega)$ with $s \geq 0$ and that the mesh consists of elements that are affine equivalent to a particular reference element. Define the parameters

$$\delta_1 = \sigma_1 h^{\beta_1}, \quad \sigma_1 > 0,$$
$$\delta_2 = \sigma_2 h^{\beta_2}, \quad \sigma_2 \geq 0,$$

with $-1 \leq \beta_1 \leq 0 \leq \beta_2 \leq 1$. Then, for $s \geq 0$ and $k \geq 1$, we have the following error estimates:

$$\|p - P_h\|_{L^2(\Omega)} \leq C h^{\min(s+\frac{1}{2}(1+m),k+\frac{1}{2}(1-M))+\frac{1}{2}(1+m)} \|p\|_{H^{s+2}(\Omega)},$$
$$\|\boldsymbol{u} - \boldsymbol{U}_h\|_{L^2(\Omega)} \leq C h^{\min(s+\frac{1}{2}(1+m),k+\frac{1}{2}(1-M))} \|p\|_{H^{s+2}(\Omega)}.$$

If $k = 0$, we have

$$\|p - P_h\|_{L^2(\Omega)} \leq C h^{1-M} \|p\|_{H^{s+2}(\Omega)},$$
$$\|\boldsymbol{u} - \boldsymbol{U}_h\|_{L^2(\Omega)} \leq C h^{\frac{1-M}{2}} \|p\|_{H^{s+2}(\Omega)},$$

where

$$M = \max(-\beta_1, \beta_2), \quad m = \min(-\beta_1, \beta_2) \quad \text{if} \quad \sigma_2 > 0,$$
$$M = \max(-\beta_1, 1), \quad m = \min(-\beta_1, 1) \quad \text{if} \quad \sigma_2 = 0.$$

The convergence rates for $k \geq 1$ and $k = 0$ are given in Tables 2.5 and 2.6, respectively. Thus, for $k \geq 1$, we do not have optimal convergence rates for both errors. Assuming s is large enough, the optimal convergence rate for $\|p - P_h\|_{L^2(\Omega)}$ is obtained for cases where

Table 2.5. *Convergence rates of LDG method for piecewise polynomial approximation of degree greater than or equal to one.*

δ_1	δ_2	$\|\boldsymbol{u} - \boldsymbol{U}_h\|_{L^2(\Omega)}$	$\|p - P_h\|_{L^2(\Omega)}$
1	0	$\min(s+1/2, k)$	$\min(s+1/2, k)+1/2$
1	h	$\min(s+1/2, k)$	$\min(s+1/2, k)+1/2$
h^{-1}	0	$\min(s+1, k)$	$\min(s+1, k)+1$
h^{-1}	h	$\min(s+1, k)$	$\min(s+1, k)+1$
1	1	$\min(s, k)+1/2$	$\min(s, k)+1$
h^{-1}	1	$\min(s+1/2, k)$	$\min(s+1/2, k)+1/2$

Table 2.6. *Convergence rates of LDG method for piecewise constant approximation.*

δ_1	δ_2	$\|\boldsymbol{u} - \boldsymbol{U}_h\|_{L^2(\Omega)}$	$\|p - P_h\|_{L^2(\Omega)}$
1	1	1/2	1
h^{-1}	0	0	0
h^{-1}	1	0	0
h^{-1}	h	0	0
1	0	0	0
1	h	0	0

(δ_1, δ_2) belongs to $\{(h^{-1}, 0), (h^{-1}, h), (1, 1)\}$. For the error $\|\boldsymbol{u} - \boldsymbol{U}_h\|_{L^2(\Omega)}$, the best rate is $\mathcal{O}(h^{k+1/2})$. In the case where $k = 0$, the method converges in the case where $\delta_1 = \mathcal{O}(1)$ and $\delta_2 = \mathcal{O}(1)$.

2.12 DG versus classical finite element method

In this section, we denote the finite element method by CG (continuous Galerkin), and we present a comparison of CG versus DG from a practical point of view. The CG method was briefly introduced in Section 2.2.2. We recall that the CG solution is a continuous piecewise polynomial, whereas the DG solution is a discontinuous piecewise polynomial.

(i) **Age of the method:** The CG method has been around for more than 60 years, and hundreds of books have been written on many aspects of the method. The primal DG methods have only recently gained an interest from the scientific community. In many cases, one can apply the techniques developed for CG to solve problems related to DG. Still, many questions remain unanswered.

(ii) **Size of problem:** For DG, the total number of degrees of freedom is proportional to the number of elements in the mesh. The constant of proportionality is a function of the polynomial degree. For CG, the degrees of freedom depend on the number of vertices and possibly the number of vertices and elements in the mesh. For instance, consider a structured mesh of 5×5 rectangular elements. The degrees of freedom for a DG approximation of degree $1, 2, 3, 4$ are $75, 150, 250, 375$, respectively, whereas the degrees of freedom for a CG approximation of degree $1, 2, 3, 4$ are $36, 121, 256, 441$, respectively. Thus, on such small mesh, if $k \geq 3$, the CG method is more costly than

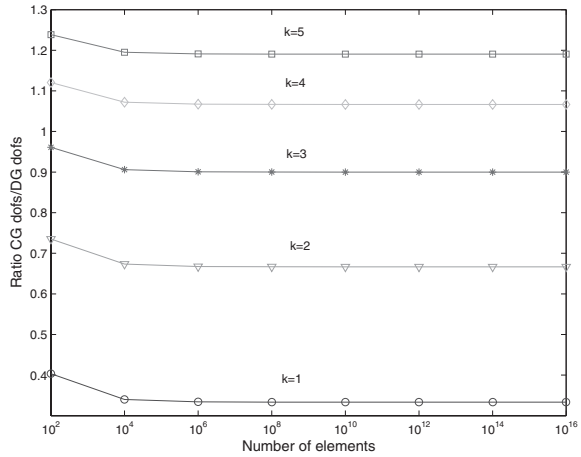

Figure 2.8. *Ratios of degrees of freedom for CG over DG with respect to the total number of degrees of freedom, computed on a uniform rectangular mesh.*

DG. The reason is that we have to use the space $\mathbb{Q}_k$ on rectangular elements for CG, but we can still use the space $\mathbb{P}_k$ on rectangular elements for DG. Fig. 2.8 gives the ratio of the total number of degrees of freedom for CG to the total number of degrees of freedom for DG on a uniform mesh of $N \times N$ rectangles. We vary N from 10 to 10^8. The CG method is less costly than DG if the polynomial degree is less than or equal to 3. The ratios tend to the limit values $1/3, 2/3, 9/10, 16/15, 15/21$ for the degrees $1, 2, 3, 4, 5$, respectively. On triangular meshes, the DG method is more costly than the CG method. For example, on a uniform mesh of $N \times N \times 2$ triangular elements, the ratios of the number of degrees of freedom for CG over DG tend to $1/6, 1/3, 9/20, 14/30, 25/42$ for the degrees $1, 2, 3, 4, 5$, respectively, as N tends to infinity. We see that this ratio increases as the order of polynomial increases.

(iii) **Meaning of degrees of freedom:** Many users of the finite element method compute only with piecewise polynomials of degree one. Because of the "chapeau" basis functions, the resulting CG degrees of freedom correspond to the values of the CG solution at the vertices of the mesh. This is a desirable property that can be exploited, for instance, in visualization routines. The degrees of freedom in the DG method do not have any meaning besides being coefficients in the expansion of the solution with respect to the basis functions. This means that in order to obtain the DG solution at a particular point, one has to compute the expansion, i.e., compute the basis functions and multiply them by the coefficients. At a given vertex, there are several values of the numerical solution. Note that we can also use the same local basis functions as in CG.

(iv) **Hanging nodes:** The name "hanging node" comes from the CG method for which mesh vertices correspond to degrees of freedom or nodes. We abuse the notation and call a hanging node any mesh vertex located on the interior of an edge (or face). Fig. 2.9 contains a mesh with 11 hanging nodes. This nonconforming mesh can be

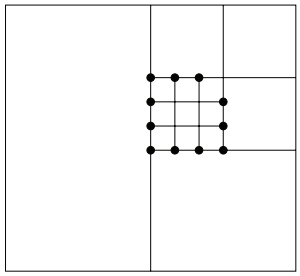

Figure 2.9. *Rectangular mesh with hanging nodes (black dots).*

used with the DG method of any order, but it cannot be used with the CG method. In general, one can have as many hanging nodes per face as one wishes for the DG method because there are no continuity constraints between the elements. In the case of the CG method, one can have at most one hanging node per edge, and special continuous basis functions have to be used.

(v) **Polynomial degree and basis functions:** It is relatively easy to change the degree of approximation of a DG solution using the same piece of software. Only the routine that computes the basis functions should be modified. The user (even beginners) can then easily perform hp-analysis of the method. Using the data structure described in Section 2.9.1, it is easy to write a DG code that uses different polynomial degrees for different mesh elements. This is an important benefit of using discontinuous approximations. For the CG method, things are less simple. In general, CG codes are first written for the piecewise linear approximations. The user then writes different codes for other polynomial degrees. As the degrees increase, the basis functions become more complicated and one has to keep track of the degrees of freedom. Some care and thought are required to obtain an hp software. In the CG method, basis functions are obtained by "pasting together" local basis functions whose support lie in one mesh element. These local basis functions can also be used to form the basis for the DG method. It suffices to extend those local basis functions by zero outside the mesh element. In practice, a simple choice of local basis functions for DG is the set of monomials.

(vi) **Accuracy:** Both methods converge as the mesh size decreases or as the polynomial degree increases. Error estimates in the energy norm are optimal. However, error estimates in the L^2 norm are optimal for the CG method, whereas they are optimal only in the symmetric version (SIPG) if no superpenalization is used. For a fixed mesh, it is irrelevant to compare the accuracy of DG with CG, as it is easy to come up with a problem that yields a better DG solution than CG and vice versa.

(vii) **Boundary condition:** Dirichlet boundary conditions are usually imposed weakly with the DG method, whereas they are imposed strongly with the CG method. But this is a matter of taste, and we can also impose the boundary conditions strongly with the DG method.

(viii) **Mass conservation:** As discussed in Section 2.7.3, the DG method satisfies a local mass balance. The CG method satisfies only a global mass balance over the whole computational domain. The property of mass conservation is crucial in flow and

transport problems, such as the ones arising in porous media. For other applications, the importance of the local mass conservation is questionable.

2.13 Bibliographical remarks

The introduction of penalty terms originates from Nitsche's work [83] in which Dirichlet boundary conditions are imposed weakly by means of the addition of a penalty term in the variational formulation rather than strongly in the space of test functions. Babuška [7] proposes another penalty method that enforces the Dirichlet boundary condition weakly. The idea of using discontinuous approximations and penalty parameters as a way to enforce interelement continuity was first introduced and analyzed by Wheeler [109] and Percell and Wheeler [85]. The method was generalized to nonlinear elliptic and parabolic problems by Arnold [1]. Similar ideas appear in the work of Baker for biharmonic problems [10]. More recently, the NIPG methods with zero penalty have been analyzed for one-dimensional problems by Babuška, Baumann, and Oden [8] and for two- and three-dimensional problems by Rivière, Wheeler, and Girault [96, 95]. The NIPG methods with nonzero penalty have been introduced by Rivière, Wheeler, and Girault [96, 95] and by Houston, Schwab, and Süli [72]: error estimates are obtained with respect to both the mesh size h and the polynomial degree k. The analysis of LDG methods can be found in the work of Castillo et al. [24], Perugia and Schötzau [86], and Dawson [40]. A unified framework for both primal and LDG methods is proposed by Arnold et al. [3, 4]. Other relevant works include [19, 50, 13, 23, 35].

Exercises

2.1. Define the set of locally integrable functions

$$L^1_{\text{loc}}(\Omega) = \{v : \forall K \text{compact} \subset \text{interior } \Omega : v|_K \in L^1(K)\}.$$

Show that if v is locally integrable, the mapping defined below is a distribution:

$$T_v(\phi) = \int_\Omega v\phi.$$

2.2. Show that if a function ϕ belongs to $H^1(E)$ such that $\int_E \phi = 0$, then there is a constant C independent of h_E such that

$$\|\phi\|_{L^2(E)} \leq Ch_E\|\nabla\phi\|_{L^2(E)}.$$

(Hint: use approximation results.)

2.3. Modify the assembling algorithm in the case of different polynomial degrees for different elements. (Hint: it might be useful to introduce an array containing the cumulative local degrees of freedom.)

2.4. Show that the mapping F_E defined by (2.33) is affine if E is a parallelogram.

2.5. Let Ω be the L-shaped domain given in Fig. 2.10. The domain is subdivided into 12 triangles. Element numbers and edge numbers are given in the left and middle

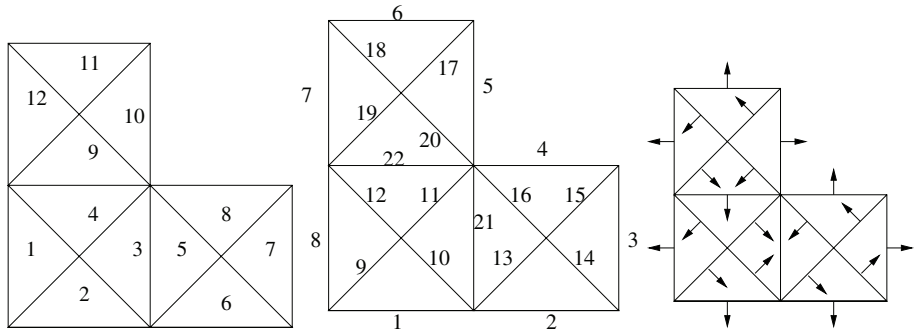

Figure 2.10. *Element numbers (left), edge numbers (middle), and normal directions (right).*

figures. The orientation of the unit normal vector $\boldsymbol{n}_e$ for each edge e is given in the right figure. Write the global matrix obtained in that case: the entries should be functions of the local matrices.

2.6. Prove Young's inequality (2.15) and Cauchy–Schwarz's inequality (2.14).

2.7. Show that the form a_ϵ is continuous on $(\mathcal{D}_k(\mathcal{E}_h))^2$ if $\sigma_e^0 > 0$; i.e, show that for all $v, w \in \mathcal{D}_k(\mathcal{E}_h)$

$$a_\epsilon(v, w) \leq M\|v\|_{\mathcal{E}}\|w\|_{\mathcal{E}}.$$