

Chapter 4

PDEs with Diffusion

The aim of this chapter is to investigate the approximation of scalar PDEs with diffusion. As a first step, we consider the *Poisson problem* with homogeneous Dirichlet boundary condition

$$-\Delta u = f \quad \text{in } \Omega, \tag{4.1a}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{4.1b}$$

and source term $f \in L^2(\Omega)$. The scalar-valued function u is termed the potential and the vector-valued function $-\nabla u$ the diffusive flux.

In Sect. 4.1, we briefly describe the continuous setting for the model problem (4.1). Then, in the following sections, we discuss three possible approaches to design a dG approximation for this problem. In Sect. 4.2, we present a heuristic derivation of a suitable discrete bilinear form loosely following the same path of ideas as in Chap. 2 hinging on consistency and discrete coercivity. There are, however, substantial differences: a specific term needs to be added to recover consistency, interface jumps as well as boundary values are penalized, and the penalty term scales as the reciprocal of the local meshsize so that discrete coercivity is expressed using a mesh-dependent norm. A further important difference is that we require to work at least with piecewise affine polynomials, thereby excluding, for the time being, methods of finite volume-type. This derivation yields the so-called *Symmetric Interior Penalty (SIP)* method of Arnold [14]. The error analysis follows fairly standard arguments and leads to optimal error estimates for smooth exact solutions. We also present a more recent analysis by the authors [132] in the case of low-regularity exact solutions. Then, using liftings of the interface jumps and boundary values, we introduce in Sect. 4.3 the important concept of discrete gradient. Applications include (1) a reformulation of the SIP bilinear form that plays a central role in Sect. 5.2 to analyze the convergence to minimal regularity solutions (as shown recently by the authors in [131]), and (2) an elementwise formulation of the discrete problem leading to a local conservation property in terms of numerical fluxes. Finally, the third approach is pursued in Sect. 4.4 where we consider mixed dG methods that approximate

both the potential and the diffusive flux. In such methods, local problems for the discrete potential and diffusive flux are formulated using numerical fluxes for both quantities, following the pioneering works of Bassi, Rebay, and coworkers [34, 35] and Cockburn and Shu [112]. This viewpoint has been adopted by Arnold, Brezzi, Cockburn, and Marini in [16] for a unified presentation of dG methods for the Poisson problem. For simplicity, we focus here on the SIP method and so-called *Local Discontinuous Galerkin (LDG)* methods [112]. In both methods, the discrete diffusive flux can be eliminated locally. We postpone the study of two-field dG methods to Sect. 7.3 in the more general context of Friedrichs' systems. Finally, we discuss hybrid mixed dG methods where additional degrees of freedom are introduced at interfaces, thereby allowing one to eliminate locally both the discrete potential and the discrete diffusive flux. This leads, in particular, to the so-called *Hybridized Discontinuous Galerkin (HDG)* methods introduced by Cockburn, Gopalakrishnan, and Lazarov [97]; see also Causin and Sacco [83] for a different approach based on a discontinuous Petrov–Galerkin formulation and Droniou and Eymard [135] for similar ideas in the context of hybrid mixed finite volume schemes.

The rest of this chapter is devoted to the study of diffusive PDEs that comprise additional terms with respect to the model problem (4.1). In Sect. 4.5, we extend the SIP method analyzed in Sect. 4.2 to heterogeneous (anisotropic) diffusion problems. The main ingredients are diffusion-dependent weighted averages to formulate the consistency and symmetry terms in the discrete bilinear form together with diffusion-dependent penalty parameters using the harmonic mean of the diffusion coefficient at each interface. In Sect. 4.6, we analyze heterogeneous diffusion-advection-reaction problems. We combine the ideas of Sect. 4.5 to handle the diffusion term with those developed in Sect. 2.3 for the upwind dG method to handle the advection-reaction term. The goal is a convergence analysis that covers both diffusion-dominated and advection-dominated regimes. The present analysis includes the case where the diffusion vanishes locally in some parts of the domain. Finally, in Sect. 4.7, we consider the heat equation as a prototype for time-dependent scalar PDEs with diffusion (that is, parabolic PDEs). The approximation is based on the SIP method for space discretization and an A-stable finite difference scheme in time; for simplicity, we focus on backward (or implicit) Euler and BDF2 schemes.

4.1 Pure Diffusion: The Continuous Setting

In this section, we present some basic facts concerning the model problem (4.1).

4.1.1 Weak Formulation and Well-Posedness

The weak formulation of (4.1) is classical:

$$\text{Find } u \in V \text{ s.t. } a(u, v) = \int_{\Omega} f v \text{ for all } v \in V, \quad (4.2)$$

with energy space $V = H_0^1(\Omega) := \{v \in H^1(\Omega) \mid v|_{\partial\Omega} = 0\}$ and bilinear form

$$a(u, v) := \int_{\Omega} \nabla u \cdot \nabla v. \quad (4.3)$$

Recalling the *Poincaré inequality* (see, e.g., Evans [153, p. 265] or Brézis [55, p. 174]) stating that there is C_{Ω} such that, for all $v \in H_0^1(\Omega)$,

$$\|v\|_{L^2(\Omega)} \leq C_{\Omega} \|\nabla v\|_{[L^2(\Omega)]^d}, \quad (4.4)$$

we infer that the bilinear form a is coercive on V . Therefore, owing to the Lax–Milgram Lemma, the weak problem (4.2) is well-posed.

4.1.2 Potential and Diffusive Flux

The PDE (4.1a) can be rewritten in *mixed form* as a system of first-order PDEs:

$$\sigma + \nabla u = 0 \quad \text{in } \Omega, \quad (4.5a)$$

$$\nabla \cdot \sigma = f \quad \text{in } \Omega. \quad (4.5b)$$

Definition 4.1 (Potential and diffusive flux). In the context of the mixed formulation (4.5), the scalar-valued function u is termed the *potential* and the vector-valued function $\sigma := -\nabla u$ is termed the *diffusive flux*.

The derivation of dG methods to approximate the model problems (4.1) on a given mesh \mathcal{T}_h hinges on the fact that the jumps of the potential and of the normal component of the diffusive flux vanish across interfaces. To allow for a more compact notation, we define boundary averages and jumps (cf. Definition 1.17 for interface averages and jumps).

Definition 4.2 (Boundary averages and jumps). For a smooth function v , for all $F \in \mathcal{F}_h^b$, and for a.e. $x \in F$, we define the *average and jump* of v as

$$\llbracket v \rrbracket_F(x) = \llbracket v \rrbracket_F(x) := v(x).$$

The subscript as well as the dependence on x are omitted unless necessary.

Since the potential u is in the energy space V , we infer that, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, letting $u_T := u|_T$, the trace $u_T|_F$ is in $L^2(F)$. Furthermore, the diffusive flux σ is in the space $H(\text{div}; \Omega)$ defined by (1.23). Traces on mesh faces of the normal component of functions in $H(\text{div}; \Omega)$ are discussed in Sect. 1.2.6. In particular, under the regularity assumption $u \in W^{2,1}(\Omega)$, there holds $\sigma \in [W^{1,1}(\Omega)]^d$, so that, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, letting $\sigma_T := \sigma|_T$ and $\sigma_{\partial T} := \sigma_T \cdot \mathbf{n}_T$ on ∂T , the trace $\sigma_{\partial T}|_F$ is in $L^1(F)$. This trace is in $L^2(F)$ under the stronger regularity assumption $u \in H^2(\Omega)$ (the assumption $u \in H^{3/2+\epsilon}(\Omega)$, $\epsilon > 0$, is actually sufficient).

We can now examine the jumps of the potential and of the normal component of the diffusive flux.

Lemma 4.3 (Jumps of potential and diffusive flux). *Assume $u \in V \cap W^{2,1}(\Omega)$. Then, there holds*

$$[[u]] = 0 \quad \forall F \in \mathcal{F}_h, \quad (4.6a)$$

$$[[\sigma]] \cdot \mathbf{n}_F = 0 \quad \forall F \in \mathcal{F}_h^i. \quad (4.6b)$$

Proof. Assertion (4.6a) results from Lemma 1.23 for interfaces and from Definition 4.2 for boundary faces. Assertion (4.6b) results from Lemma 1.24. \square

4.2 Symmetric Interior Penalty

Our goal is to approximate the solution of the model problem (4.2) using dG methods in the broken polynomial space $\mathbb{P}_d^k(\mathcal{T}_h)$ defined by (1.15). We set

$$V_h := \mathbb{P}_d^k(\mathcal{T}_h),$$

with polynomial degree $k \geq 1$ and \mathcal{T}_h belonging to an admissible mesh sequence. The focus of this section is on a specific dG method, the Symmetric Interior Penalty (SIP) method introduced by Arnold [14].

For simplicity, we enforce a somewhat strong regularity assumption on the exact solution. A weaker regularity assumption is made in Sect. 4.2.5.

Assumption 4.4 (Regularity of exact solution and space V_*). *We assume that the exact solution u is such that*

$$u \in V_* := V \cap H^2(\Omega).$$

*In the spirit of Sect. 1.3, we set $V_{*h} := V_* + V_h$.*

Without further knowledge on the exact solution u apart from the domain Ω and the datum $f \in L^2(\Omega)$, Assumption 4.4 can be asserted for instance if the domain Ω is convex; see Grisvard [177]. Assumption 4.4 differs from the concept of elliptic regularity (cf. Definition 4.24 below) since Assumption 4.4 only concerns the exact solution u .

4.2.1 Heuristic Derivation

To derive a suitable discrete bilinear form, we loosely follow the same path of ideas as in Chap. 2 aiming at a discrete bilinear form that satisfies the consistency requirement (1.32) and enjoys discrete coercivity. Moreover, we add a (consistent) term to recover, at the discrete level, the symmetry of the continuous problem.

4.2.1.1 Consistency

We begin localizing gradients to mesh elements in the exact bilinear form a , that is, we set, for all $(v, w_h) \in V_{*h} \times V_h$,

$$a_h^{(0)}(v, w_h) := \int_{\Omega} \nabla_h v \cdot \nabla_h w_h = \sum_{T \in \mathcal{T}_h} \int_T \nabla v \cdot \nabla w_h.$$

To examine the consistency requirement (1.32), we integrate by parts on each mesh element. This leads to

$$a_h^{(0)}(v, w_h) = - \sum_{T \in \mathcal{T}_h} \int_T (\Delta v) w_h + \sum_{T \in \mathcal{T}_h} \int_{\partial T} (\nabla v \cdot \mathbf{n}_T) w_h.$$

The second term on the right-hand side can be reformulated as a sum over mesh faces in the form

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} (\nabla v \cdot \mathbf{n}_T) w_h = \sum_{F \in \mathcal{F}_h^i} \int_F [(\nabla_h v) w_h] \cdot \mathbf{n}_F + \sum_{F \in \mathcal{F}_h^b} \int_F (\nabla v \cdot \mathbf{n}_F) w_h,$$

since for all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$, $\mathbf{n}_F = \mathbf{n}_{T_1} = -\mathbf{n}_{T_2}$. Moreover,

$$[(\nabla_h v) w_h] = \{\{\nabla_h v\}\}[w_h] + [\nabla_h v]\{\{w_h\}\},$$

since letting $a_i = (\nabla v)|_{T_i}$, $b_i = w_h|_{T_i}$, $i \in \{1, 2\}$, yields

$$\begin{aligned} [(\nabla_h v) w_h] &= a_1 b_1 - a_2 b_2 \\ &= \frac{1}{2}(a_1 + a_2)(b_1 - b_2) + (a_1 - a_2)\frac{1}{2}(b_1 + b_2) \\ &= \{\{\nabla_h v\}\}[w_h] + [\nabla_h v]\{\{w_h\}\}. \end{aligned}$$

As a result, and accounting for boundary faces using Definition 4.2, yields

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} (\nabla v \cdot \mathbf{n}_T) w_h = \sum_{F \in \mathcal{F}_h} \int_F \{\{\nabla_h v\}\} \cdot \mathbf{n}_F [w_h] + \sum_{F \in \mathcal{F}_h^i} \int_F [\nabla_h v] \cdot \mathbf{n}_F \{\{w_h\}\}.$$

Hence,

$$\begin{aligned} a_h^{(0)}(v, w_h) &= - \sum_{T \in \mathcal{T}_h} \int_T (\Delta v) w_h + \sum_{F \in \mathcal{F}_h} \int_F \{\{\nabla_h v\}\} \cdot \mathbf{n}_F [w_h] \\ &\quad + \sum_{F \in \mathcal{F}_h^i} \int_F [\nabla_h v] \cdot \mathbf{n}_F \{\{w_h\}\}. \end{aligned} \tag{4.7}$$

To check consistency, we set $v = u$ in (4.7). A consequence of (4.6b) is that, for all $w_h \in V_h$,

$$a_h^{(0)}(u, w_h) = \int_{\Omega} f w_h + \sum_{F \in \mathcal{F}_h} \int_F (\nabla u \cdot \mathbf{n}_F) [w_h].$$

In order to match the consistency requirement (1.32), we are prompted to modify $a_h^{(0)}$ as follows: For all $(v, w_h) \in V_{*h} \times V_h$,

$$a_h^{(1)}(v, w_h) := \int_{\Omega} \nabla_h v \cdot \nabla_h w_h - \sum_{F \in \mathcal{F}_h} \int_F \{\{\nabla_h v\}\} \cdot \mathbf{n}_F [w_h].$$

It is clear that $a_h^{(1)}$ is consistent in the sense of (1.32), i.e., for all $w_h \in V_h$,

$$a_h^{(1)}(u, w_h) = \int_{\Omega} f w_h.$$

4.2.1.2 Symmetry

A desirable property of the discrete bilinear form is to preserve the original symmetry of the exact bilinear form. Indeed, symmetry can simplify the solution of the resulting linear system and furthermore, it is a natural ingredient to derive optimal L^2 -norm error estimates (cf. Sect. 4.2.4); nonsymmetric variants are discussed in Sect. 5.3. In view of this remark, we set, for all $(v, w_h) \in V_{*h} \times V_h$,

$$a_h^{\text{cs}}(v, w_h) := \int_{\Omega} \nabla_h v \cdot \nabla_h w_h - \sum_{F \in \mathcal{F}_h} \int_F (\llbracket \nabla_h v \rrbracket \cdot \mathbf{n}_F \llbracket w_h \rrbracket + \llbracket v \rrbracket \llbracket \nabla_h w_h \rrbracket \cdot \mathbf{n}_F), \quad (4.8)$$

so that a_h^{cs} is symmetric on $V_h \times V_h$. The bilinear form a_h^{cs} remains consistent owing to (4.6a). The superscript in a_h^{cs} indicates the consistency and symmetry achieved so far. For future use, we record the following equivalent expression of a_h^{cs} resulting from (4.7),

$$\begin{aligned} a_h^{\text{cs}}(v, w_h) = & - \sum_{T \in \mathcal{T}_h} \int_T (\Delta v) w_h + \sum_{F \in \mathcal{F}_h^i} \int_F \llbracket \nabla_h v \rrbracket \cdot \mathbf{n}_F \llbracket w_h \rrbracket \\ & - \sum_{F \in \mathcal{F}_h} \int_F \llbracket v \rrbracket \llbracket \nabla_h w_h \rrbracket \cdot \mathbf{n}_F. \end{aligned} \quad (4.9)$$

4.2.1.3 Penalties on Interface Jumps and Boundary Values

The last requirement to match is discrete coercivity on the broken polynomial space V_h with respect to a suitable norm. The difficulty with the discrete bilinear form a_h^{cs} defined by (4.8) is that, for all $v_h \in V_h$,

$$a_h^{\text{cs}}(v_h, v_h) = \|\nabla_h v_h\|_{[L^2(\Omega)]^d}^2 - 2 \sum_{F \in \mathcal{F}_h} \int_F \llbracket \nabla_h v_h \rrbracket \cdot \mathbf{n}_F \llbracket v_h \rrbracket,$$

and the second term on the right-hand side has no a priori sign so that, without adding a further term, there is no hope for discrete coercivity (in some situations, discrete inf-sup stability can be achieved without penalty; cf. Remark 4.14). To achieve discrete coercivity, we add to a_h^{cs} a term penalizing interface and boundary jumps, namely we set, for all $(v, w_h) \in V_{*h} \times V_h$,

$$a_h^{\text{sip}}(v, w_h) := a_h^{\text{cs}}(v, w_h) + s_h(v, w_h), \quad (4.10)$$

with the stabilization bilinear form

$$s_h(v, w_h) := \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} \int_F \llbracket v \rrbracket \llbracket w_h \rrbracket, \quad (4.11)$$

where $\eta > 0$ is a user-dependent parameter and h_F a local length scale associated with the mesh face $F \in \mathcal{F}_h$. We observe that, owing to (4.6a), adding the bilinear form s_h to a_h^{cs} does not alter the consistency and symmetry achieved so far.

Moreover, Lemma 4.12 below shows that, provided the penalty parameter η is large enough, the discrete bilinear form a_h^{sip} enjoys discrete coercivity on V_h .

We now present a simple choice for the local length scale h_F .

Definition 4.5 (Local length scale h_F). For all $F \in \mathcal{F}_h$, in dimension $d \geq 2$, we set h_F to be equal to the diameter of the face F , while, in dimension 1, we set $h_F := \min(h_{T_1}, h_{T_2})$ if $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$ and $h_F := h_T$ if $F \in \mathcal{F}_h^b$ with $F = \partial T \cap \partial \Omega$. In all cases, for a mesh element $T \in \mathcal{T}_h$, h_T denotes its diameter (cf. Definition 1.13).

Remark 4.6 (Local length scale h_F). Other choices are possible for the local length scale h_F weighting the face penalties in the stabilization bilinear form s_h , e.g., the choice $h_F = \llbracket h \rrbracket := \frac{1}{2}(h_{T_1} + h_{T_2})$ for all $F \in \mathcal{F}_h^i$, or the choice $h_F = \frac{\llbracket |T|_d \rrbracket}{|F|_{d-1}}$ (that is, the mean value of the d -dimensional Hausdorff measures of the neighboring elements divided by the $(d-1)$ -dimensional Hausdorff measure of the face, recalling that for $d = 1$, $|F|_0 = 1$). Incidentally, we observe that modifying the choice for the local length scale impacts the value of the minimal threshold on the penalty parameter η for which discrete coercivity is achieved.

Combining (4.10) with (4.11) yields, for all $(v, w_h) \in V_{*h} \times V_h$,

$$\begin{aligned} a_h^{\text{sip}}(v, w_h) &= \int_{\Omega} \nabla_h v \cdot \nabla_h w_h - \sum_{F \in \mathcal{F}_h} \int_F (\llbracket \nabla_h v \rrbracket \cdot \mathbf{n}_F \llbracket w_h \rrbracket + \llbracket v \rrbracket \llbracket \nabla_h w_h \rrbracket \cdot \mathbf{n}_F) \\ &\quad + \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} \int_F \llbracket v \rrbracket \llbracket w_h \rrbracket, \end{aligned} \quad (4.12)$$

or, equivalently using (4.9),

$$\begin{aligned} a_h^{\text{sip}}(v, w_h) &= - \sum_{T \in \mathcal{T}_h} \int_T (\Delta v) w_h + \sum_{F \in \mathcal{F}_h^i} \int_F \llbracket \nabla_h v \rrbracket \cdot \mathbf{n}_F \llbracket w_h \rrbracket \\ &\quad - \sum_{F \in \mathcal{F}_h} \int_F \llbracket v \rrbracket \llbracket \nabla_h w_h \rrbracket \cdot \mathbf{n}_F + \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} \int_F \llbracket v \rrbracket \llbracket w_h \rrbracket. \end{aligned} \quad (4.13)$$

The idea of weakly enforcing boundary and jump conditions on the discrete solution using penalties can be traced back to the seventies, in particular the work of Nitsche [248, 249], Babuška [20], Babuška and Zlámal [24], Douglas and Dupont [134], Baker [25], and Wheeler [306]. The discrete bilinear form a_h^{sip} defined by (4.12) corresponds to the Symmetric Interior Penalty (SIP) method introduced by Arnold [14]; henceforth, a_h^{sip} is called the SIP bilinear form. In the present context, interior penalty means interior as well as boundary penalties.

Definition 4.7 (Consistency, symmetry, and penalty terms). The second, third, and fourth terms on the right-hand side of (4.12) are respectively called *consistency*, *symmetry*, and *penalty* terms.

4.2.1.4 The Discrete Problem

The discrete problem is

$$\text{Find } u_h \in V_h \text{ s.t. } a_h^{\text{sip}}(u_h, v_h) = \int_{\Omega} f v_h \text{ for all } v_h \in V_h. \quad (4.14)$$

Lemma 4.12 below states that provided the penalty parameter η is large enough, the SIP bilinear form is coercive on V_h . Thus, owing to the Lax–Milgram Lemma, the discrete problem (4.14) is well-posed. Moreover, a straightforward consequence of the above derivation is consistency.

Lemma 4.8 (Consistency). *Assume $u \in V_*$. Then, for all $v_h \in V_h$,*

$$a_h^{\text{sip}}(u, v_h) = \int_{\Omega} f v_h.$$

Remark 4.9 (Rough right-hand side). At the continuous level, the Poisson problem can be posed for a right-hand side $f \in V' = H^{-1}(\Omega)$, the dual space of the energy space $V = H_0^1(\Omega)$, leading to the weak formulation

$$a(u, v) = \langle f, v \rangle_{V', V} \quad \forall v \in V.$$

Since the discrete space V_h is nonconforming in V , it is not possible, at the discrete level, to take $\langle f, v_h \rangle_{V', V}$ as right-hand side in (4.14). One possibility is to use a smoothing operator $\mathcal{I}_h : V_h \rightarrow V_h \cap H_0^1(\Omega)$ and to consider the discrete problem

$$\text{Find } u_h \in V_h \text{ s.t. } a_h^{\text{sip}}(u_h, v_h) = \langle f, \mathcal{I}_h v_h \rangle_{V', V} \text{ for all } v_h \in V_h. \quad (4.15)$$

One example of smoothing operator is the averaging operator considered in Sect. 5.5.2. An important observation is that (4.15) is no longer consistent.

Remark 4.10 (Stencil). With an eye toward implementation, we identify the elementary stencil (cf. Definition 2.26) associated with the SIP bilinear form. For all $T \in \mathcal{T}_h$, the stencil of the volume contribution is just the element T , while the stencil associated with the consistency, symmetry, and penalty terms consists of T and its neighbors in the sense of faces. Figure 4.1 illustrates the stencil for a matching triangular mesh; cf. Sect. A.1.3 for further insight.

4.2.2 Other Boundary Conditions

The discrete problem (4.14), which was derived in the context of homogeneous Dirichlet boundary conditions, needs to be slightly modified when dealing with other boundary conditions. The modifications are designed so as to maintain consistency when the exact solution satisfies other boundary conditions. For instance, when (weakly) enforcing the nonhomogeneous Dirichlet boundary condition $u = g$ on $\partial\Omega$ with $g \in H^{1/2}(\partial\Omega)$, the discrete problem becomes

$$\text{Find } u_h \in V_h \text{ s.t. } a_h^{\text{sip}}(u_h, v_h) = l_h^D(g; v_h) \text{ for all } v_h \in V_h,$$

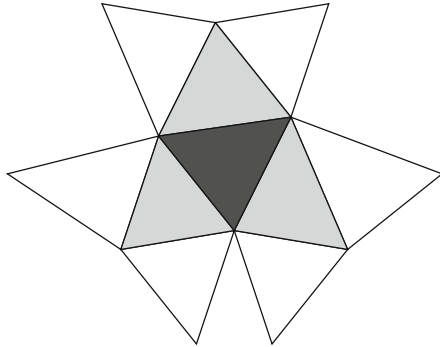


Fig. 4.1: Example of stencil of an element $T \in \mathcal{T}_h$ when \mathcal{T}_h is a matching triangular mesh; the mesh element T is highlighted in *dark grey*, and its three neighbors, which all belong to the stencil, are highlighted in *light grey*; the other triangles do not belong to the stencil

with a_h^{sip} still defined by (4.12) and the new right-hand side

$$l_h^D(g; v_h) := \int_{\Omega} f v_h - \int_{\partial\Omega} g \nabla_h v_h \cdot \mathbf{n} + \sum_{F \in \mathcal{F}_h^b} \frac{\eta}{h_F} \int_F g v_h.$$

As a result, for the exact solution $u \in V_*$, $a_h^{\text{sip}}(u, v_h) = l_h^D(g; v_h)$, for all $v_h \in V_h$. Furthermore, when (weakly) enforcing the Robin boundary condition $\gamma u + \nabla u \cdot \mathbf{n} = g$ on $\partial\Omega$ with $g \in L^2(\partial\Omega)$ and $\gamma \in L^\infty(\partial\Omega)$ such that γ is nonnegative a.e. on $\partial\Omega$, the discrete problem becomes

$$\text{Find } u_h \in V_h \text{ s.t. } a_h^R(u_h, v_h) = l_h^R(g; v_h) \text{ for all } v_h \in V_h,$$

where, for all $(v, w_h) \in V_{*h} \times V_h$,

$$\begin{aligned} a_h^R(v, w_h) := & \int_{\Omega} \nabla_h v \cdot \nabla_h w_h - \sum_{F \in \mathcal{F}_h^i} \int_{\Omega} (\{\nabla_h v\} \cdot \mathbf{n}_F \llbracket w_h \rrbracket + \llbracket v \rrbracket \{\nabla_h w_h\} \cdot \mathbf{n}_F) \\ & + \sum_{F \in \mathcal{F}_h^i} \frac{\eta}{h_F} \int_F \llbracket v \rrbracket \llbracket w_h \rrbracket + \sum_{F \in \mathcal{F}_h^b} \int_F \gamma v_h w_h, \end{aligned} \quad (4.16)$$

and

$$l_h^R(g; w_h) := \int_{\Omega} f w_h + \int_{\partial\Omega} g w_h.$$

As a result, for the exact solution $u \in V_*$, $a_h^R(u, v_h) = l_h^R(g; v_h)$, for all $v_h \in V_h$. Moreover, we observe that, unlike in the Dirichlet case, the summations in the consistency and symmetry terms are restricted to interfaces. Finally, the case $\gamma \equiv 0$ corresponds to the Neumann problem. For this problem, the data must comply with the compatibility condition $\int_{\Omega} f = -\int_{\partial\Omega} g$, and the solution is

determined up to an additive constant. One possibility is to additionally enforce $\int_{\Omega} u_h = 0$. In practice, the discrete problem can still be formulated using V_h as trial and test space, and the additional constraint can be enforced by postprocessing. Observing that the rank of the problem matrix is one unit less than its size, the discrete solution can be obtained (1) using a direct solver with full pivoting, so that the zero pivot is encountered when processing the last line of the matrix and a solution can be obtained by fixing an arbitrary value for the degree of freedom left or (2) using an iterative solver which only requires matrix-vector products. The most common linear algebra libraries (e.g., the PETSc library [26]) offer specific functionalities to handle this case efficiently.

4.2.3 Basic Energy-Error Estimate

Let u solve the weak problem (4.2) and let u_h solve the discrete problem (4.14). The aim of this section is to estimate the approximation error $(u - u_h)$. The convergence analysis is performed in the spirit of Theorem 1.35. We recall that the space V_* is specified in Assumption 4.4 and that $V_{*h} = V_* + V_h$.

4.2.3.1 Discrete Coercivity

We aim at asserting this property using the following norm: For all $v \in V_{*h}$,

$$\|v\|_{\text{sip}} := \left(\|\nabla_h v\|_{[L^2(\Omega)]^d}^2 + |v|_J^2 \right)^{1/2}, \quad (4.17)$$

with the *jump seminorm*

$$|v|_J := (\eta^{-1} s_h(v, v))^{1/2} = \left(\sum_{F \in \mathcal{F}_h} \frac{1}{h_F} \|\llbracket v \rrbracket\|_{L^2(F)}^2 \right)^{1/2}. \quad (4.18)$$

We observe that $\|\cdot\|_{\text{sip}}$ is indeed a norm on V_{*h} , and even on the broken Sobolev space $H^1(\mathcal{T}_h)$. The only nontrivial property to check is whether, for all $v \in H^1(\mathcal{T}_h)$, $\|v\|_{\text{sip}} = 0$ implies $v = 0$. Clearly, $\|v\|_{\text{sip}} = 0$ implies $\|\nabla_h v\|_{[L^2(\Omega)]^d} = 0$ and $|v|_J = 0$. The first property yields $\nabla_h v = 0$ so that v is piecewise constant. The second property implies that the interface and boundary jumps of v vanish. Hence, $v = 0$.

Our first step toward establishing discrete coercivity for the SIP bilinear form is a bound on the consistency term using the jump seminorm $|\cdot|_J$.

Lemma 4.11 (Bound on consistency term). *For all $(v, w_h) \in V_{*h} \times V_h$,*

$$\left| \sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h v\} \cdot \mathbf{n}_F \llbracket w_h \rrbracket \right| \leq \left(\sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_T} h_F \|\nabla v|_{T \cdot \mathbf{n}_F}\|_{L^2(F)}^2 \right)^{1/2} |w_h|_J. \quad (4.19)$$

Proof. For all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$, and $a_i = (\nabla v)|_{T_i} \cdot \mathbf{n}_F$, $i \in \{1, 2\}$, the Cauchy–Schwarz inequality yields

$$\begin{aligned} \int_F \{\{\nabla_h v\} \cdot \mathbf{n}_F\} [w_h] &= \int_F \frac{1}{2} (a_1 + a_2) [w_h] \\ &\leq \left(\frac{1}{2} h_F (\|a_1\|_{L^2(F)}^2 + \|a_2\|_{L^2(F)}^2) \right)^{1/2} h_F^{-1/2} \|[w_h]\|_{L^2(F)}. \end{aligned}$$

Moreover, for all $F \in \mathcal{F}_h^b$ with $F = \partial T \cap \partial \Omega$,

$$\int_F \{\{\nabla_h v\} \cdot \mathbf{n}_F\} [w_h] \leq h_F^{1/2} \|\nabla v|_T \cdot \mathbf{n}_F\|_{L^2(F)} \times h_F^{-1/2} \|[w_h]\|_{L^2(F)}.$$

Summing over mesh faces, using the Cauchy–Schwarz inequality, and regrouping the face contributions for each mesh element yields the assertion. \square

We can now turn to the discrete coercivity of the SIP bilinear form. We recall that N_∂ , defined by (1.12), denotes the maximum number of mesh faces composing the boundary of a generic mesh element and that this quantity is bounded uniformly in h ; cf. Lemma 1.41.

Lemma 4.12 (Discrete coercivity). *For all $\eta > \underline{\eta} := C_{\text{tr}}^2 N_\partial$ where C_{tr} results from the discrete trace inequality (1.37) and the parameter N_∂ is defined by (1.12), the SIP bilinear form defined by (4.12) is coercive on V_h with respect to the $\|\cdot\|_{\text{sip}}$ -norm, i.e.,*

$$\forall v_h \in V_h, \quad a_h^{\text{sip}}(v_h, v_h) \geq C_\eta \|v_h\|_{\text{sip}}^2,$$

with $C_\eta := (\eta - C_{\text{tr}}^2 N_\partial)(1 + \eta)^{-1}$.

Proof. Let $v_h \in V_h$. Since, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, $h_F \leq h_T$ (cf. Definition 4.5), we obtain using the discrete trace inequality (1.40),

$$\begin{aligned} \sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_T} h_F \|\nabla v_h|_T \cdot \mathbf{n}_F\|_{L^2(F)}^2 &\leq \sum_{T \in \mathcal{T}_h} h_T \|\nabla v_h|_T \cdot \mathbf{n}_F\|_{L^2(\partial T)}^2 \\ &\leq C_{\text{tr}}^2 N_\partial \|\nabla_h v_h\|_{[L^2(\Omega)]^d}^2, \end{aligned}$$

whence we infer from (4.19) that

$$\left| \sum_{F \in \mathcal{F}_h} \int_F \{\{\nabla_h v_h\} \cdot \mathbf{n}_F\} [v_h] \right| \leq C_{\text{tr}} N_\partial^{1/2} \|\nabla_h v_h\|_{[L^2(\Omega)]^d} |v_h|_{\text{J}}.$$

As a result,

$$a_h^{\text{sip}}(v_h, v_h) \geq \|\nabla_h v_h\|_{[L^2(\Omega)]^d}^2 - 2C_{\text{tr}} N_\partial^{1/2} \|\nabla_h v_h\|_{[L^2(\Omega)]^d} |v_h|_{\text{J}} + \eta |v_h|_{\text{J}}^2.$$

We now use the following inequality: Let β be a positive real number, let $\eta > \beta^2$; then, for all $x, y \in \mathbb{R}$,

$$x^2 - 2\beta xy + \eta y^2 \geq \frac{\eta - \beta^2}{1 + \eta} (x^2 + y^2).$$

Applying this inequality with $\beta = C_{\text{tr}} N_{\partial}^{1/2}$, $x = \|\nabla_h v_h\|_{[L^2(\Omega)]^d}$, and $y = |v|_J$ yields the assertion. \square

Remark 4.13 (Modifying the local length scale). Recalling Remark 4.6, other choices for the local length scale h_F can be made when defining the stabilization bilinear form s_h . The jump seminorm $|\cdot|_J$ is still defined by $|v|_J := (\eta^{-1} s_h(v, v))^{-1/2}$, and the proof of Lemma 4.12 can be deployed as above as long as the chosen length scale is a lower bound for the diameter of both neighboring elements; otherwise, an additional factor appears in the definition of the minimal threshold $\underline{\eta}$.

Remark 4.14 (Discrete stability without penalty). In one space dimension, the discrete bilinear form a_h^{cs} enjoys discrete inf-sup stability without adding the stabilization bilinear form s_h for polynomial degrees $k \geq 2$; see Burman, Ern, Mozolevski, and Stamm [67]. Furthermore, in two and three space dimensions and using piecewise affine discrete functions supplemented by suitable element bubble functions, discrete inf-sup stability can be proven for the discrete bilinear form a_h^{cs} again without adding the stabilization bilinear form s_h ; see Burman and Stamm [70]. Finally, it is also possible to devise penalty strategies acting only on the low-degree part of the jumps; see, e.g., Hansbo and Larson [184] and Burman and Stamm [71].

Remark 4.15 (Poincaré inequality using the $\|\cdot\|_{\text{sip}}$ -norm). It can be proven (cf. Corollary 5.4) that there exists σ_2 , independent of h , such that

$$\forall v_h \in V_h, \quad \|v_h\|_{L^2(\Omega)} \leq \sigma_2 \|v_h\|_{\text{sip}}. \quad (4.20)$$

More generally, on the broken Sobolev space $H^1(\mathcal{T}_h)$, it is proven by Brenner [51] (see also Arnold [14]) that, for $d \in \{2, 3\}$, there is σ'_2 , independent of h , such that

$$\forall v \in H^1(\mathcal{T}_h), \quad \|v\|_{L^2(\Omega)} \leq \sigma'_2 \|v\|_{\text{sip}}. \quad (4.21)$$

4.2.3.2 Boundedness

We define on V_{*h} the norm

$$\|v\|_{\text{sip},*} := \left(\|v\|_{\text{sip}}^2 + \sum_{T \in \mathcal{T}_h} h_T \|\nabla v|_{T \cdot \mathbf{n}_T}\|_{L^2(\partial T)}^2 \right)^{1/2}. \quad (4.22)$$

Lemma 4.16 (Boundedness). *There is C_{bnd} , independent of h , such that*

$$\forall (v, w_h) \in V_{*h} \times V_h, \quad a_h^{\text{sip}}(v, w_h) \leq C_{\text{bnd}} \|v\|_{\text{sip},*} \|w_h\|_{\text{sip}}. \quad (4.23)$$

Proof. Let $(v, w_h) \in V_{*h} \times V_h$. We observe that

$$\begin{aligned} a_h^{\text{sip}}(v, w_h) &= \int_{\Omega} \nabla_h v \cdot \nabla_h w_h - \sum_{F \in \mathcal{F}_h} \int_F \{\{\nabla_h v\}\} \cdot \mathbf{n}_F \llbracket w_h \rrbracket - \sum_{F \in \mathcal{F}_h} \int_F \llbracket v \rrbracket \{\{\nabla_h w_h\}\} \cdot \mathbf{n}_F \\ &\quad + \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} \int_F \llbracket v \rrbracket \llbracket w_h \rrbracket := \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3 + \mathfrak{T}_4. \end{aligned} \quad (4.24)$$

Using the Cauchy–Schwarz inequality yields $|\mathfrak{I}_1| \leq \|\nabla_h v\|_{[L^2(\Omega)]^d} \|\nabla_h w_h\|_{[L^2(\Omega)]^d}$ and $|\mathfrak{I}_4| \leq \eta |v|_J |w_h|_J$. Moreover, owing to the bound (4.19) and since $h_F \leq h_T$,

$$|\mathfrak{I}_2| \leq \left(\sum_{T \in \mathcal{T}_h} h_T \|\nabla v|_{T \cdot \mathbf{n}_T}\|_{L^2(\partial T)}^2 \right)^{1/2} |w_h|_J \leq \|v\|_{\text{sip},*} |w_h|_J \leq \|v\|_{\text{sip},*} \|w_h\|_{\text{sip}},$$

by definition of the $\|\cdot\|_{\text{sip},*}$ -norm. Finally, still owing to the bound (4.19) and proceeding as in the proof of Lemma 4.12 leads to

$$|\mathfrak{I}_3| \leq C_{\text{tr}} N_\partial^{1/2} |v|_J \|\nabla_h w_h\|_{[L^2(\Omega)]^d} \leq C_{\text{tr}} N_\partial^{1/2} \|v\|_{\text{sip}} \|w_h\|_{\text{sip}}.$$

Collecting the above bounds yields (4.23) with $C_{\text{bnd}} = 2 + \eta + C_{\text{tr}} N_\partial^{1/2}$. \square

4.2.3.3 $\|\cdot\|_{\text{sip}}$ -Norm Error Estimate and Convergence

A straightforward consequence of the above results together with Theorem 1.35 is the following error estimate.

Theorem 4.17 ($\|\cdot\|_{\text{sip}}$ -norm error estimate). *Let $u \in V_*$ solve (4.2). Let u_h solve (4.14) with a_h^{sip} defined by (4.12) and penalty parameter as in Lemma 4.12. Then, there is C , independent of h , such that*

$$\|u - u_h\|_{\text{sip}} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{\text{sip},*}. \quad (4.25)$$

To infer a convergence result from (4.25), we assume that the exact solution is smooth enough and use Lemmata 1.58 and 1.59. The resulting estimate is optimal both for the broken gradient and the jump seminorm.

Corollary 4.18 (Convergence rate in $\|\cdot\|_{\text{sip}}$ -norm). *Besides the hypotheses of Theorem 4.17, assume $u \in H^{k+1}(\Omega)$. Then, there holds*

$$\|u - u_h\|_{\text{sip}} \leq C_u h^k, \quad (4.26)$$

with $C_u = C \|u\|_{H^{k+1}(\Omega)}$ and C independent of h .

Remark 4.19 (Bound on the jumps). The contribution of the jump seminorm to the error $\|u - u_h\|_{\text{sip}}$ can be controlled by the contribution of the broken gradient under some assumptions. For instance, Lemma 5.30 shows that, on matching simplicial meshes and for a large enough penalty parameter, there holds, up to a positive factor independent of h ,

$$|u - u_h|_J = |u_h|_J \lesssim \|\nabla_h(u - u_h)\|_{[L^2(\Omega)]^d} + \mathcal{R}_{\text{osc},\Omega},$$

where the data oscillation term $\mathcal{R}_{\text{osc},\Omega}$, defined by (5.34), converges to zero at order h^{k+1} if $f \in H^k(\Omega)$ and at order h^{k+2} if $f \in H^{k+1}(\Omega)$. We also refer the reader to Bonito and Nochetto [46] for a similar bound on the jumps on general meshes, and to Ainsworth and Rankin [7, 9] for a sharper condition on the penalty parameter on triangular meshes with hanging nodes.

4.2.3.4 Analysis Using Only the $\|\cdot\|_{\text{sip},*}$ -Norm

The convergence analysis of elliptic problems is often performed using a single norm. Such an approach is possible here by working only with the $\|\cdot\|_{\text{sip}}$ -norm which turns out to be uniformly equivalent to the $\|\cdot\|_{\text{sip},*}$ -norm on V_h .

Lemma 4.20 (Uniform equivalence of $\|\cdot\|_{\text{sip}}$ - and $\|\cdot\|_{\text{sip},*}$ -norms on V_h). *The $\|\cdot\|_{\text{sip}}$ - and $\|\cdot\|_{\text{sip},*}$ -norms are uniformly equivalent on V_h . Specifically,*

$$C_{\text{sip}} \|v_h\|_{\text{sip},*} \leq \|v_h\|_{\text{sip}} \leq \|v_h\|_{\text{sip},*} \quad \forall v_h \in V_h,$$

with C_{sip} independent of h .

Proof. The upper bound is immediate, while the lower bound results from the discrete trace inequality (1.40) and the uniform bound on N_{∂} . \square

A consequence of Lemma 4.20 is discrete coercivity on V_h in the form

$$\forall v_h \in V_h, \quad a_h^{\text{sip}}(v_h, v_h) \geq C'_\eta \|v_h\|_{\text{sip},*}^2,$$

with C'_η independent of h . Moreover, an inspection at the proof of Lemma 4.16 leads to boundedness on $V_{*h} \times V_h$ in the form

$$\forall (v, w_h) \in V_{*h} \times V_h, \quad a_h^{\text{sip}}(v, w_h) \leq C'_{\text{bnd}} \|v\|_{\text{sip},*} \|w_h\|_{\text{sip},*},$$

with C'_{bnd} independent of h . Using the above results leads to the following convergence results in the $\|\cdot\|_{\text{sip},*}$ -norm.

Theorem 4.21 ($\|\cdot\|_{\text{sip},*}$ -norm error estimate). *Under the hypotheses of Theorem 4.17, there is C , independent of h , such that*

$$\|u - u_h\|_{\text{sip},*} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{\text{sip},*}. \quad (4.27)$$

Corollary 4.22 (Convergence rate in $\|\cdot\|_{\text{sip},*}$ -norm). *Besides the hypotheses of Theorem 4.17, assume $u \in H^{k+1}(\Omega)$. Then, there holds*

$$\|u - u_h\|_{\text{sip},*} \leq C_u h^k, \quad (4.28)$$

with $C_u = C \|u\|_{H^{k+1}(\Omega)}$ and C independent of h .

Remark 4.23 (Comparison with $\|\cdot\|_{\text{sip}}$ -norm estimates). The discrete coercivity of a_h^{sip} is naturally expressed using the $\|\cdot\|_{\text{sip}}$ -norm, whereas using the $\|\cdot\|_{\text{sip},*}$ -norm leads to the inclusion in the error estimate of the additional factor C_{sip} related to norm equivalence (cf. Lemma 4.20). Therefore, estimates (4.25) and (4.26) deliver a sharper bound on the broken gradient and the jump seminorm. However, estimates (4.27) and (4.28) convey additional information regarding the convergence of the normal gradient at mesh element boundaries.

4.2.4 L^2 -Norm Error Estimate

Using the broken Poincaré inequality (4.21), the $\|\cdot\|_{\text{sip}}$ -norm estimate (4.26) yields the L^2 -norm estimate

$$\|u - u_h\|_{L^2(\Omega)} \leq \sigma'_2 C_u h^k.$$

This estimate is suboptimal by one power in h . To remedy this drawback and recover optimality, it is possible to resort to a duality argument (the so-called Aubin–Nitsche argument [17]) under the following assumption.

Definition 4.24 (Elliptic regularity). We say that *elliptic regularity* holds true for the model problem (4.2) if there is C_{ell} , only depending on Ω , such that, for all $\psi \in L^2(\Omega)$, the solution to the problem:

$$\text{Find } \zeta \in H_0^1(\Omega) \text{ s.t. } a(\zeta, v) = \int_{\Omega} \psi v \text{ for all } v \in H_0^1(\Omega),$$

is in V_* and satisfies

$$\|\zeta\|_{H^2(\Omega)} \leq C_{\text{ell}} \|\psi\|_{L^2(\Omega)}.$$

Elliptic regularity can be asserted if, for instance, the polygonal domain Ω is convex; see Grisvard [177]. To derive an L^2 -norm error estimate, we extend the SIP bilinear form to $V_{*h} \times V_{*h}$, so that both arguments of a_h^{sip} can belong to V_* .

Theorem 4.25 (L^2 -norm error estimate). *Let $u \in V_*$ solve (4.2). Let u_h solve (4.14) with a_h^{sip} defined by (4.12). Assume elliptic regularity. Then, there is C , independent of h , such that*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch \|u - u_h\|_{\text{sip},*}. \quad (4.29)$$

Proof. We consider the auxiliary problem

$$\text{Find } \zeta \in H_0^1(\Omega) \text{ s.t. } a(\zeta, v) = \int_{\Omega} (u - u_h)v \text{ for all } v \in H_0^1(\Omega),$$

and use elliptic regularity to infer $\|\zeta\|_{H^2(\Omega)} \leq C_{\text{ell}} \|u - u_h\|_{L^2(\Omega)}$. Since $\zeta \in V_*$, $[\![\nabla \zeta]\!] \cdot \mathbf{n}_F = 0$ on all $F \in \mathcal{F}_h^i$ and $[\![\zeta]\!] = 0$ on all $F \in \mathcal{F}_h$. Hence, (4.13) implies

$$a_h^{\text{sip}}(\zeta, u - u_h) = \int_{\Omega} (-\Delta \zeta)(u - u_h).$$

Exploiting the symmetry of a_h^{sip} and since $-\Delta \zeta = u - u_h$, we obtain

$$a_h^{\text{sip}}(u - u_h, \zeta) = \|u - u_h\|_{L^2(\Omega)}^2.$$

Furthermore, since consistency implies Galerkin orthogonality (cf. Remark 1.32) and letting π_h^1 be the L^2 -orthogonal projection onto $\mathbb{P}_d^1(\mathcal{T}_h) \subset V_h$ (since $k \geq 1$), we infer

$$a_h^{\text{sip}}(u - u_h, \pi_h^1 \zeta) = 0.$$

Hence, using the boundedness of a_h^{sip} on $V_{*h} \times V_{*h}$ which results from the fact that $a_h^{\text{sip}}(v, w) \lesssim \|v\|_{\text{sip},*} \|w\|_{\text{sip},*}$ for all $v, w \in V_{*h}$, the approximation properties of π_h^1 in the $\|\cdot\|_{\text{sip},*}$ -norm, and the regularity of ζ , we obtain, up to multiplicative factors independent of h ,

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)}^2 &= a_h^{\text{sip}}(u - u_h, \zeta - \pi_h^1 \zeta) \\ &\lesssim \|u - u_h\|_{\text{sip},*} \|\zeta - \pi_h^1 \zeta\|_{\text{sip},*} \\ &\lesssim \|u - u_h\|_{\text{sip},*} h \|\zeta\|_{H^2(\mathcal{T}_h)} \\ &\lesssim \|u - u_h\|_{\text{sip},*} h \|u - u_h\|_{L^2(\Omega)}. \end{aligned}$$

Simplifying by $\|u - u_h\|_{L^2(\Omega)}$ yields (4.29). \square

A straightforward consequence of (4.28) and (4.29) is the following convergence result for smooth solutions.

Corollary 4.26 (Convergence rate in L^2 -norm). *Besides the hypotheses of Theorem 4.17, assume elliptic regularity and $u \in H^{k+1}(\Omega)$. Then, there holds*

$$\|u - u_h\|_{L^2(\Omega)} \leq C_u h^{k+1}, \quad (4.30)$$

with $C_u = C\|u\|_{H^{k+1}(\Omega)}$ and C independent of h .

Estimate (4.30) is optimal. We emphasize that the symmetry of a_h^{sip} has been used in the proof of Theorem 4.25.

Remark 4.27 (Adjoint-consistency). Following Arnold, Brezzi, Cockburn, and Marini [16], the property $a_h^{\text{sip}}(u - u_h, \zeta) = \int_{\Omega} (-\Delta \zeta)(u - u_h)$, which results from symmetry and consistency, can be termed *adjoint consistency*.

Remark 4.28 (Error estimates in other norms). We refer the reader, e.g., to Chen and Chen [89] and to Guzmán [181] for pointwise error estimates on the discrete solution and its broken gradient using weighted broken Sobolev norms.

4.2.5 Analysis for Low-Regularity Solutions

This section is devoted to the analysis of the SIP method under a regularity assumption on the exact solution that is weaker than Assumption 4.4.

Assumption 4.29 (Regularity of exact solution and space V_*). *We assume that $d \geq 2$ and that there is $p \in (\frac{2d}{d+2}, 2]$ such that, for the exact solution u ,*

$$u \in V_* := V \cap W^{2,p}(\Omega).$$

In the spirit of Sect. 1.3, we set $V_{*h} := V_* + V_h$.

Assumption 4.29 requires $p > 1$ for $d = 2$ and $p > \frac{6}{5}$ for $d = 3$. In particular, we observe that, in two space dimensions, $u \in W^{2,p}(\Omega)$ with $p > 1$ holds true in

polygonal domains; see, e.g., Dauge [119]. Moreover, using Sobolev embeddings (see Evans [153, Sect. 5.6] or Brézis [55, Sect. IX.3]), Assumption 4.29 implies

$$u \in H^{1+\alpha_p}(\Omega), \quad \alpha_p = \frac{d+2}{2} - \frac{d}{p} > 0. \quad (4.31)$$

We still consider the discrete problem (4.14) with the discrete bilinear form a_h^{sip} defined by (4.12). The convergence analysis under the regularity assumption 4.29 has been performed recently by Wihler and Rivière [308] in two space dimensions and, using slightly different techniques, by the authors [132] in the context of heterogeneous diffusion in any space dimension; cf. Sect. 4.5. We follow here the approach of [132], building up on the analysis presented in Sect. 4.2.3 for smooth solutions. In the present context of an exact solution with low-regularity, we assume for simplicity $k = 1$. We also assume $p < 2$ since in the case $p = 2$, Assumption 4.29 amounts to Assumption 4.4.

We already know that discrete coercivity holds true provided the penalty parameter is chosen as in Lemma 4.12. Moreover, owing to Lemma 4.3, the discrete bilinear form a_h^{sip} can be extended to $V_{*h} \times V_h$, and consistency can be asserted as in Lemma 4.8. Thus, it only remains to prove boundedness. To this purpose, we need to redefine the $\|\cdot\|_{\text{sip},*}$ -norm since functions in V_* are such that, for all $T \in \mathcal{T}_h$, $\nabla v|_T \cdot \mathbf{n}_T$ is in $L^p(\partial T)$, but not necessarily in $L^2(\partial T)$. Thus, we now define on V_{*h} the norm

$$\|v\|_{\text{sip},*} := \left(\|v\|_{\text{sip}}^p + \sum_{T \in \mathcal{T}_h} h_T^{1+\gamma_p} \|\nabla v|_T \cdot \mathbf{n}_T\|_{L^p(\partial T)}^p \right)^{1/p}, \quad (4.32)$$

where $\gamma_p := \frac{1}{2}d(p-2)$. We observe that, for $p = 2$, we recover the previous definition (4.22) of the $\|\cdot\|_{\text{sip},*}$ -norm. The value for γ_p is motivated by the following boundedness result.

Lemma 4.30 (Boundedness). *There is C_{bnd} , independent of h , such that*

$$\forall (v, w_h) \in V_{*h} \times V_h, \quad a_h^{\text{sip}}(v, w_h) \leq C_{\text{bnd}} \|v\|_{\text{sip},*} \|w_h\|_{\text{sip}}. \quad (4.33)$$

Proof. Let $(v, w_h) \in V_{*h} \times V_h$. We need to bound the four terms $\mathfrak{T}_1, \dots, \mathfrak{T}_4$ in (4.24). Proceeding as in the proof of Lemma 4.16, we obtain

$$|\mathfrak{T}_1 + \mathfrak{T}_3 + \mathfrak{T}_4| \leq C \|v\|_{\text{sip}} \|w_h\|_{\text{sip}},$$

with C independent of h , so that it only remains to bound the consistency term \mathfrak{T}_2 . To this purpose, we proceed similarly to the proof of (4.19), but use Hölder's inequality instead of the Cauchy-Schwarz inequality. For all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$, and $a_i = (\nabla v)|_{T_i} \cdot \mathbf{n}_F$, $i \in \{1, 2\}$, Hölder's inequality yields

$$\begin{aligned} \int_F \{\nabla_h v\} \cdot \mathbf{n}_F \llbracket w_h \rrbracket &= \int_F \frac{1}{2} (a_1 + a_2) \llbracket w_h \rrbracket \\ &\leq \left(\frac{1}{2} h_F^{1+\gamma_p} (\|a_1\|_{L^p(F)}^p + \|a_2\|_{L^p(F)}^p) \right)^{1/p} h_F^{-\beta_p} \|\llbracket w_h \rrbracket\|_{L^q(F)}, \end{aligned}$$

with $\beta_p = \frac{1+\gamma_p}{p}$ and $q = \frac{p}{p-1}$. Moreover, for all $F \in \mathcal{F}_h^b$ with $F = \partial T \cap \partial\Omega$,

$$\int_F \{\{\nabla_h v\} \cdot \mathbf{n}_F\} [w_h] \leq \left(h_F^{1+\gamma_p} \|\nabla v|_T \cdot \mathbf{n}_T\|_{L^p(F)}^p \right)^{1/p} h_F^{-\beta_p} \| [w_h] \|_{L^q(F)}.$$

Owing to the inverse inequality (1.43) and since $\beta_p - \frac{1}{2} = (d-1)(\frac{1}{q} - \frac{1}{2})$, we infer

$$h_F^{-\beta_p} \| [w_h] \|_{L^q(F)} \leq C_{\text{inv},q,2} h_F^{-1/2} \| [w_h] \|_{L^2(F)},$$

where $C_{\text{inv},q,2}$ is independent of h and can be bounded uniformly in q (cf. Remark 1.51). Combining the above bounds, summing over mesh faces, and using Hölder's inequality yields

$$\begin{aligned} \left| \sum_{F \in \mathcal{F}_h} \int_F \{\{\nabla_h v\} \cdot \mathbf{n}_F\} [w_h] \right| &\leq \left(\sum_{T \in \mathcal{T}_h} h_T^{1+\gamma_p} \|\nabla v|_T \cdot \mathbf{n}_T\|_{L^p(\partial T)}^p \right)^{1/p} \\ &\quad \times C_{\text{inv},q,2} \left(\sum_{F \in \mathcal{F}_h} \left(h_F^{-1/2} \| [w_h] \|_{L^2(F)} \right)^q \right)^{1/q}. \end{aligned}$$

Since $q \geq 2$, we obtain

$$\left(\sum_{F \in \mathcal{F}_h} \left(h_F^{-1/2} \| [w_h] \|_{L^2(F)} \right)^q \right)^{1/q} \leq \left(\sum_{F \in \mathcal{F}_h} \left(h_F^{-1/2} \| [w_h] \|_{L^2(F)} \right)^2 \right)^{1/2} = |w_h|_J.$$

Hence,

$$\left| \sum_{F \in \mathcal{F}_h} \int_F \{\{\nabla_h v\} \cdot \mathbf{n}_F\} [w_h] \right| \leq \left(\sum_{T \in \mathcal{T}_h} h_T^{1+\gamma_p} \|\nabla v|_T \cdot \mathbf{n}_T\|_{L^p(\partial T)}^p \right)^{1/p} C_{\text{inv},q,2} |w_h|_J,$$

whence we infer (4.33). \square

To state a convergence result, we need optimal polynomial approximation for functions in $V_* = W^{2,p}(\Omega)$. For simplicity, we restricted the presentation of Sect. 1.4.4 to the Hilbertian setting. In the present non-Hilbertian setting, we make the following assumption.

Assumption 4.31 (Optimal polynomial approximation in $W^{2,p}(T)$). *The mesh sequence $(\mathcal{T}_h)_{h \in \mathcal{H}}$ is such that, for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, and all $v \in W^{2,p}(T)$, there holds*

$$|v - \pi_h v|_{W^{m,p}(T)} \leq C_{\text{app}} h_T^{2-m} |v|_{W^{2,p}(T)} \quad m \in \{0, 1, 2\}, \quad (4.34a)$$

$$|v - \pi_h v|_{H^m(T)} \leq C_{\text{app}} h_T^{1+\alpha_p-m} |v|_{W^{2,p}(T)} \quad m \in \{0, 1\}, \quad (4.34b)$$

with C_{app} independent of both T and h , while α_p is defined by (4.31).

Assumption 4.31 can be asserted for mesh sequences with star-shaped property or finitely-shaped property; cf. Lemmata 1.61 and 1.62. We can now turn to our main convergence result.

Theorem 4.32 ($\|\cdot\|_{\text{sip}}$ -norm error estimate and convergence rate). *Let $u \in V_*$ solve (4.2). Let u_h solve (4.14) with a_h^{sip} defined by (4.12) and penalty parameter as in Lemma 4.12. Then, there is C , independent of h , such that*

$$\|u - u_h\|_{\text{sip}} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{\text{sip},*}, \quad (4.35)$$

where the $\|\cdot\|_{\text{sip},*}$ -norm is defined by (4.32). Moreover, under Assumption 4.31, there holds

$$\|u - u_h\|_{\text{sip}} \leq C_u h^{\alpha_p}, \quad (4.36)$$

with $C_u = C|u|_{W^{2,p}(\Omega)}$ and C independent of h .

Proof. Estimate (4.35) is a direct consequence of Theorem 1.35 since we established discrete coercivity, consistency, and boundedness. We now take $v_h = \pi_h u$ in (4.35). We first observe that, for all $T \in \mathcal{T}_h$, using (4.34a) together with the continuous trace inequality (1.18) yields

$$\|\nabla(u - \pi_h u)|_{T \cdot \mathbf{n}_T}\|_{L^p(\partial T)} \lesssim h_T^{1-1/p} |u|_{W^{2,p}(T)},$$

where $a \lesssim b$ means the inequality $a \leq Cb$ with generic positive C independent of h and T . Since $\frac{1+\gamma_p}{p} + 1 - \frac{1}{p} = \alpha_p$, we infer

$$\left(\sum_{T \in \mathcal{T}_h} h_T^{1+\gamma_p} \|\nabla(u - \pi_h u)|_{T \cdot \mathbf{n}_T}\|_{L^p(\partial T)}^p \right)^{1/p} \lesssim h^{\alpha_p} |u|_{W^{2,p}(\Omega)}.$$

Moreover, using (4.34b) together with the continuous trace inequality (1.19) yields

$$\|u - \pi_h u\|_{\text{sip}} \lesssim h^{\alpha_p} |u|_{W^{2,p}(\Omega)}.$$

Combining the two above bounds leads to (4.36). \square

The convergence rate in the error estimate (4.36) is optimal both for the broken gradient and the jump seminorm.

4.3 Liftings and Discrete Gradients

Liftings are operators that map scalar-valued functions defined on mesh faces to vector-valued functions defined on mesh elements. In the context of dG methods, liftings act on interface and boundary jumps. They were introduced by Bassi, Rebay, and coworkers [34, 35] in the context of compressible flows and analyzed by Brezzi, Manzini, Marini, Pietra, and Russo [58, 59] in the context of the Poisson problem (see also Perugia and Schötzau [257] for the hp -analysis). Liftings have many useful applications. They can be combined

with the broken gradient to define discrete gradients. Discrete gradients play an important role in the design and analysis of dG methods. Indeed, they can be used to formulate the discrete problem locally on each mesh element using numerical fluxes. Moreover, as detailed in Sect. 5.1, they are instrumental in the derivation of discrete functional analysis results, that, in turn, play a central role in the convergence analysis to minimal regularity solutions (cf. Sect. 5.2). Liftings can also be employed to define the stabilization bilinear form [35], yielding a more convenient lower bound for the penalty parameter η ; cf. Sect. 5.3.2.

4.3.1 Liftings: Definition and Stability

As before, we assume that the mesh \mathcal{T}_h belongs to an admissible mesh sequence. For any mesh face $F \in \mathcal{F}_h$ and for any integer $l \geq 0$, we define the (local) lifting operator

$$\mathbf{r}_F^l : L^2(F) \longrightarrow [\mathbb{P}_d^l(\mathcal{T}_h)]^d$$

as follows: For all $\varphi \in L^2(F)$,

$$\int_{\Omega} \mathbf{r}_F^l(\varphi) \cdot \boldsymbol{\tau}_h = \int_F \llbracket \boldsymbol{\tau}_h \rrbracket \cdot \mathbf{n}_F \varphi \quad \forall \boldsymbol{\tau}_h \in [\mathbb{P}_d^l(\mathcal{T}_h)]^d. \quad (4.37)$$

We observe that the support of $\mathbf{r}_F^l(\varphi)$ consists of the one or two mesh elements of which F is part of the boundary; using the set \mathcal{T}_F defined by (1.13) yields

$$\text{supp}(\mathbf{r}_F^l) = \bigcup_{T \in \mathcal{T}_F} \bar{T}. \quad (4.38)$$

Moreover, whenever the mesh face F is a portion of a hyperplane (this happens, for instance, when working with simplicial meshes or with general meshes consisting of convex elements), $\mathbf{r}_F^l(\varphi)$ is colinear to the normal vector \mathbf{n}_F .

Lemma 4.33 (Bound on local lifting). *Let $F \in \mathcal{F}_h$ and let $l \geq 0$. For all $\varphi \in L^2(F)$, there holds*

$$\|\mathbf{r}_F^l(\varphi)\|_{[L^2(\Omega)]^d} \leq C_{\text{tr}} h_F^{-1/2} \|\varphi\|_{L^2(F)}. \quad (4.39)$$

Proof. Let $\varphi \in L^2(F)$. Equation (4.37), the fact that $h_F \leq h_T$ for all $T \in \mathcal{T}_F$, and the discrete trace inequality (1.37) yield

$$\begin{aligned} \|\mathbf{r}_F^l(\varphi)\|_{[L^2(\Omega)]^d}^2 &= \int_{\Omega} \mathbf{r}_F^l(\varphi) \cdot \mathbf{r}_F^l(\varphi) = \int_F \llbracket \mathbf{r}_F^l(\varphi) \rrbracket \cdot \mathbf{n}_F \varphi \\ &\leq \left(\frac{1}{h_F} \int_F |\varphi|^2 \right)^{1/2} \times \left(h_F \int_F |\llbracket \mathbf{r}_F^l(\varphi) \rrbracket|^2 \right)^{1/2} \\ &\leq h_F^{-1/2} \|\varphi\|_{L^2(F)} \times C_{\text{tr}} \left(\text{card}(\mathcal{T}_F)^{-1} \sum_{T \in \mathcal{T}_F} \int_T |\mathbf{r}_F^l(\varphi)|^2 \right)^{1/2}, \end{aligned}$$

whence (4.39) follows since $\text{card}(\mathcal{T}_F)^{-1} \leq 1$ and since $\sum_{T \in \mathcal{T}_F} \int_T |\mathbf{r}_F^l(\varphi)|^2 = \|\mathbf{r}_F^l(\varphi)\|_{[L^2(\Omega)]^d}^2$ owing to (4.38). \square

For any integer $l \geq 0$ and for any function $v \in H^1(\mathcal{T}_h)$, we define the (global) lifting of its interface and boundary jumps as

$$R_h^l(\llbracket v \rrbracket) := \sum_{F \in \mathcal{F}_h} r_F^l(\llbracket v \rrbracket) \in [\mathbb{P}_d^l(\mathcal{T}_h)]^d, \quad (4.40)$$

being implicitly understood that r_F^l acts on the function $\llbracket v \rrbracket_F$ (which is in $L^2(F)$ since $v \in H^1(\mathcal{T}_h)$).

Lemma 4.34 (Bound on global lifting). *Let $l \geq 0$. For all $v \in H^1(\mathcal{T}_h)$, there holds*

$$\|R_h^l(\llbracket v \rrbracket)\|_{[L^2(\Omega)]^d} \leq N_\partial^{1/2} \left(\sum_{F \in \mathcal{F}_h} \|r_F^l(\llbracket v \rrbracket)\|_{[L^2(\Omega)]^d}^2 \right)^{1/2}, \quad (4.41)$$

so that

$$\|R_h^l(\llbracket v \rrbracket)\|_{[L^2(\Omega)]^d} \leq C_{\text{tr}} N_\partial^{1/2} |v|_J. \quad (4.42)$$

Proof. Let $v \in H^1(\mathcal{T}_h)$. Owing to (4.38), we infer $(R_h^l(\llbracket v \rrbracket))|_T = \sum_{F \in \mathcal{F}_T} (r_F^l(\llbracket v \rrbracket))|_T$, so that using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \|R_h^l(\llbracket v \rrbracket)\|_{[L^2(\Omega)]^d}^2 &= \sum_{T \in \mathcal{T}_h} \int_T \left| \sum_{F \in \mathcal{F}_T} r_F^l(\llbracket v \rrbracket) \right|^2 \\ &\leq \sum_{T \in \mathcal{T}_h} \text{card}(\mathcal{F}_T) \sum_{F \in \mathcal{F}_T} \int_T |r_F^l(\llbracket v_h \rrbracket)|^2 \\ &\leq \max_{T \in \mathcal{T}_h} \text{card}(\mathcal{F}_T) \sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_T} \int_T |r_F^l(\llbracket v_h \rrbracket)|^2 \\ &= \max_{T \in \mathcal{T}_h} \text{card}(\mathcal{F}_T) \sum_{F \in \mathcal{F}_h} \|r_F^l(\llbracket v \rrbracket)\|_{[L^2(\Omega)]^d}^2, \end{aligned}$$

and the bound (4.41) follows using the definition (1.12) of N_∂ . Finally, (4.42) results from (4.41) and the fact that

$$\left(\sum_{F \in \mathcal{F}_h} \|r_F^l(\llbracket v \rrbracket)\|_{[L^2(\Omega)]^d}^2 \right)^{1/2} \leq C_{\text{tr}} |v|_J,$$

owing to Lemma 4.33. \square

To illustrate in the case $l = 0$ (piecewise constant liftings), we obtain, for all $v \in H^1(\mathcal{T}_h)$ and all $T \in \mathcal{T}_h$,

$$R_h^0(\llbracket v \rrbracket)|_T = \sum_{F \in \mathcal{F}_T} \frac{|F|_{d-1}}{|T|_d} (v_F - v_T) \mathbf{n}_{T,F}, \quad (4.43)$$

where $\mathbf{n}_{T,F}$ is the outward normal to T on F , $v_T := v|_T$, and $v_F := \frac{1}{2}(v_T + v|_{T'})$ whenever $F = \partial T \cap \partial T'$, $T \neq T'$, while $v_F := 0$ if $F \in \mathcal{F}_h^b$. The (opposite of the) above expression has been used as a gradient reconstruction in the context of finite volume methods replacing v_F by a consistent trace reconstruction (see Eymard, Gallouët, and Herbin [158]); cf. also formula (5.28).

4.3.2 Discrete Gradients: Definition and Stability

For any integer $l \geq 0$, we define the discrete gradient operator

$$G_h^l : H^1(\mathcal{T}_h) \longrightarrow [L^2(\Omega)]^d,$$

as follows: For all $v \in H^1(\mathcal{T}_h)$,

$$G_h^l(v) := \nabla_h v - R_h^l(\llbracket v \rrbracket). \quad (4.44)$$

Proposition 4.35 (Bound on discrete gradient). *Let $l \geq 0$. For all $v \in H^1(\mathcal{T}_h)$, there holds*

$$\|G_h^l(v)\|_{[L^2(\Omega)]^d} \leq (1 + C_{\text{tr}}^2 N_\partial)^{1/2} \|v\|_{\text{sip}},$$

where the $\|\cdot\|_{\text{sip}}$ -norm is defined by (4.17).

Proof. Let $v \in H^1(\mathcal{T}_h)$. Using the triangle inequality together with (4.42) yields

$$\begin{aligned} \|G_h^l(v)\|_{[L^2(\Omega)]^d} &\leq \|\nabla_h v\|_{[L^2(\Omega)]^d} + \|R_h^l(\llbracket v \rrbracket)\|_{[L^2(\Omega)]^d} \\ &\leq \|\nabla_h v\|_{[L^2(\Omega)]^d} + C_{\text{tr}} N_\partial^{1/2} |v|_J, \end{aligned}$$

whence the assertion. \square

4.3.3 Reformulation of the SIP Bilinear Form

Let $l \in \{k-1, k\}$ and set, as in Sect. 4.2, $V_h = \mathbb{P}_d^k(\mathcal{T}_h)$ with $k \geq 1$ and \mathcal{T}_h belonging to an admissible mesh sequence. Following Brezzi, Manzini, Marini, Pietra, and Russo [58], it is interesting to observe that the bilinear form a_h^{cs} defined by (4.8) can be equivalently written as follows: For all $v_h, w_h \in V_h$,

$$a_h^{\text{cs}}(v_h, w_h) = \int_{\Omega} \nabla_h v_h \cdot \nabla_h w_h - \int_{\Omega} \nabla_h v_h \cdot R_h^l(\llbracket w_h \rrbracket) - \int_{\Omega} \nabla_h w_h \cdot R_h^l(\llbracket v_h \rrbracket). \quad (4.45)$$

This results from definitions (4.37) and (4.40) and the fact that $\nabla_h v_h$ and $\nabla_h w_h$ are in $[\mathbb{P}_d^l(\mathcal{T}_h)]^d$ since $l \geq k-1$, so that, for all $F \in \mathcal{F}_h$,

$$\int_F \{\nabla_h v_h\} \cdot \mathbf{n}_F \llbracket w_h \rrbracket = \int_{\Omega} \nabla_h v_h \cdot \mathbf{r}_F^l(\llbracket w_h \rrbracket).$$

Starting from (4.45) and using the definition (4.44) of the discrete gradient, we infer, for all $v_h, w_h \in V_h$,

$$a_h^{\text{cs}}(v_h, w_h) = \int_{\Omega} G_h^l(v_h) \cdot G_h^l(w_h) - \int_{\Omega} R_h^l(\llbracket v_h \rrbracket) \cdot R_h^l(\llbracket w_h \rrbracket).$$

As a result, recalling that the SIP bilinear form considered in Sect. 4.2 is such that $a_h^{\text{sip}} = a_h^{\text{cs}} + s_h$ with s_h defined by (4.11), we obtain, for all $v_h, w_h \in V_h$,

$$a_h^{\text{sip}}(v_h, w_h) = \int_{\Omega} G_h^l(v_h) \cdot G_h^l(w_h) + \hat{s}_h^{\text{sip}}(v_h, w_h), \quad (4.46)$$

with

$$\hat{s}_h^{\text{sip}}(v_h, w_h) := \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} \int_F \llbracket v_h \rrbracket \llbracket w_h \rrbracket - \int_{\Omega} \mathbf{R}_h^l(\llbracket v_h \rrbracket) \cdot \mathbf{R}_h^l(\llbracket w_h \rrbracket). \quad (4.47)$$

The most natural choice for l appears to be $l = k - 1$ since the broken gradient is in $[\mathbb{P}_d^{k-1}(\mathcal{T}_h)]^d$. The choice $l = k$ can facilitate the implementation of the method in that it allows one to use the same polynomial basis for computing the liftings and assembling the matrix.

The interest in using discrete gradients to formulate dG methods has been recognized recently in various contexts, e.g., by Lew, Neff, Sulsky, and Ortiz [232] and Ten Eyck and Lew [293] for linear and nonlinear elasticity, Buffa and Ortner [61] and Burman and Ern [65] for nonlinear variational problems, and the authors [131] for the Navier–Stokes equations; see also Agélas, Di Pietro, Eymard, and Masson [6]. The expression (4.46) of the SIP bilinear form plays a central role in Sect. 5.2 when analyzing the convergence to minimal regularity solutions. This expression is also useful in Sect. 4.4 in the context of a mixed dG approximation.

It is interesting to notice the following straightforward consequence of the bound (4.42).

Proposition 4.36 (Discrete coercivity). *For all $v_h \in V_h$,*

$$a_h^{\text{sip}}(v_h, v_h) \geq \|G_h^l(v_h)\|_{[L^2(\Omega)]^d}^2 + (\eta - C_{\text{tr}}^2 N_{\partial}) |v_h|_{\mathcal{J}}^2.$$

In view of this result, the expression (4.46) for a_h^{sip} consists of two terms, both yielding a nonnegative contribution whenever $w_h = v_h$ and, as in Lemma 4.12, $\eta > C_{\text{tr}}^2 N_{\partial}$. The first term can be seen as the discrete counterpart of the exact bilinear form a (such that $a(v, w) = \int_{\Omega} \nabla v \cdot \nabla w$) and provides a control on the discrete gradient in $[L^2(\Omega)]^d$. The role of the second term is to strengthen the discrete stability of the method.

Remark 4.37 (Extension to broken Sobolev spaces). We emphasize that the definition (4.46) of a_h^{sip} is equivalent to (4.12) only at the discrete level. Differences occur when extending the definitions (4.12) and (4.46) to larger spaces, e.g., broken Sobolev spaces. As discussed in Sect. 4.2.1, the SIP bilinear form defined by (4.12) cannot be extended to the minimum regularity space $H^1(\Omega)$ because traces of gradients on mesh faces are used. Instead, the bilinear form defined by (4.46) can be extended to the broken Sobolev space $H^1(\mathcal{T}_h)$. We denote this extension by \tilde{a}_h^{sip} . Incidentally, \tilde{a}_h^{sip} is no longer consistent. For convergence analysis to smooth solutions, Strang’s First Lemma (see [285] or, e.g., Braess [49, p. 106]) dedicated to nonconsistent finite element methods can be used, whereby the consistency error is estimated for $u \in H^{k+1}(\Omega)$ as follows: For all $v_h \in V_h$,

$$\tilde{a}_h^{\text{sip}}(u - u_h, v_h) = \sum_{F \in \mathcal{F}_h} \int_F \{\nabla u - \pi_h(\nabla u)\} \cdot \mathbf{n}_F \llbracket v_h \rrbracket \leq C_u h^k |v_h|_{\mathcal{J}},$$

where π_h denotes the L^2 -orthogonal projection onto V_h . As a result, the consistency error tends optimally to zero as the meshsize goes to zero.

4.3.4 Numerical Fluxes

DG methods can be viewed as high-order finite volume methods. The aim of this section is to identify the local conservation properties associated with dG methods. Such properties are important when the diffusive flux is to be used as an advective velocity in a transport problem, as discussed, e.g., by Dawson, Sun, and Wheeler [121] in the context of coupled porous media flow and contaminant transport. While most discretization methods possess local conservation properties, the specificity of dG methods, together with finite volume and mixed finite element methods, is to achieve local conservation at the element level as opposed to vertex-centered or face-centered macro-elements; see, e.g., Eymard, Hilhorst, and Vohralík [161].

Let $T \in \mathcal{T}_h$ and let $\xi \in \mathbb{P}_d^k(T)$. Integration by parts shows that, for the exact solution u ,

$$\int_T f\xi = - \int_T (\Delta u)\xi = \int_T \nabla u \cdot \nabla \xi - \int_{\partial T} (\nabla u \cdot \mathbf{n}_T)\xi.$$

Therefore, defining on each mesh face $F \in \mathcal{F}_h$ the exact flux as

$$\Phi_F(u) := -\nabla u \cdot \mathbf{n}_F, \quad (4.48)$$

and recalling the notation $\epsilon_{T,F} = \mathbf{n}_T \cdot \mathbf{n}_F$ introduced in Sect. 2.2.3, we infer

$$\int_T \nabla u \cdot \nabla \xi + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F \Phi_F(u)\xi = \int_T f\xi.$$

This is a local conservation property satisfied by the exact solution. Our goal is to identify a similar relation satisfied by the discrete solution u_h solving (4.14). Using $v_h = \xi\chi_T$ as test function in (4.14) (where χ_T denotes the characteristic function of T), observing that $\nabla_h(\xi\chi_T) = (\nabla\xi)\chi_T$, and recalling the definition (4.12) of a_h^{sup} , we obtain

$$\begin{aligned} \int_T f\xi &= a_h^{\text{sup}}(u_h, \xi\chi_T) = \int_T \nabla u_h \cdot \nabla \xi - \sum_{F \in \mathcal{F}_T} \int_F \{\nabla_h u_h\} \cdot \mathbf{n}_F [\xi\chi_T] \\ &\quad - \sum_{F \in \mathcal{F}_T} \int_F \{(\nabla\xi)\chi_T\} \cdot \mathbf{n}_F [u_h] + \sum_{F \in \mathcal{F}_T} \frac{\eta}{h_F} \int_F [u_h] [\xi\chi_T]. \end{aligned}$$

Let $l \in \{k-1, k\}$. The first and third terms on the right-hand side sum up to $\int_T G_h^{k-1}(u_h) \cdot \nabla \xi$ since $\nabla \xi \in [\mathbb{P}_d^{k-1}(T)]^d$ and $l \geq k-1$, while in the second and fourth terms, we observe that $[\xi\chi_T] = \epsilon_{T,F}\xi$. As a result, for all $T \in \mathcal{T}_h$ and all $\xi \in \mathbb{P}_d^k(T)$,

$$\int_T G_h^l(u_h) \cdot \nabla \xi + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F \phi_F(u_h)\xi = \int_T f\xi, \quad (4.49)$$

with the numerical flux $\phi_F(u_h)$ defined as

$$\phi_F(u_h) := -\{\nabla_h u_h\} \cdot \mathbf{n}_F + \frac{\eta}{h_F} [u_h]. \quad (4.50)$$

We notice that the two contributions to $\phi_F(u_h)$ in (4.50) respectively stem from the consistency term and the penalty term (cf. Definition 4.7). Equation (4.49) is the local conservation property satisfied by the dG approximation. Interestingly, the expression (4.50) is consistent with (4.48) since, for the exact solution u , $\phi_F(u) = \Phi_F(u)$. We also observe that the local conservation property (4.49) is richer than that encountered in finite volume methods, which can be recovered by just taking $\xi \equiv 1$, i.e.,

$$\sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F \phi_F(u_h) = \int_T f. \quad (4.51)$$

4.4 Mixed dG Methods

In this section, we discuss mixed dG methods, that is, dG approximations to the mixed formulation (4.5) with the homogeneous Dirichlet boundary condition (4.1b). Other boundary conditions can be considered. Such methods produce an approximation u_h for the potential u and an approximation σ_h for the diffusive flux σ .

Definition 4.38 (Discrete potential and discrete diffusive flux). Consistently with Definition 4.1, the scalar-valued function u_h is termed the *discrete potential* and the vector-valued function σ_h the *discrete diffusive flux*.

First, we reformulate the SIP method of Sect. 4.2 as a mixed dG method and show how the discrete diffusive flux can be eliminated locally. Then, we formulate more general mixed dG methods in terms of local problems using numerical fluxes for the discrete potential and the discrete diffusive flux following Bassi, Rebay, and coworkers [34, 35]. This leads, in particular, to the LDG methods introduced by Cockburn and Shu [112]. In these methods, the discrete diffusive flux can also be eliminated locally. Finally, we discuss hybrid mixed dG methods where additional degrees of freedom are introduced at interfaces, thereby allowing one to eliminate locally both the discrete potential and the discrete diffusive flux.

4.4.1 The SIP Method As a Mixed dG Method

One possible weak formulation of the mixed formulation (4.5) with the homogeneous Dirichlet boundary condition (4.1b) consists in finding $(\sigma, u) \in X := [L^2(\Omega)]^d \times H_0^1(\Omega)$ such that

$$\begin{cases} m(\sigma, \tau) + b(\tau, u) = 0 & \forall \tau \in [L^2(\Omega)]^d, \\ -b(\sigma, v) = \int_{\Omega} f v & \forall v \in H_0^1(\Omega), \end{cases} \quad (4.52)$$

where, for all $\sigma, \tau \in [L^2(\Omega)]^d$ and all $v \in H_0^1(\Omega)$, we have introduced the bilinear forms

$$m(\sigma, \tau) := \int_{\Omega} \sigma \cdot \tau, \quad b(\tau, v) := \int_{\Omega} \tau \cdot \nabla v.$$

It is easily seen that $(\sigma, u) \in X$ solves (4.52) if and only if $\sigma = -\nabla u$ and u solves the weak problem (4.2).

At the discrete level, a mixed dG approximation can be designed as follows. We consider a polynomial degree $k \geq 1$ for the approximation of the potential and choose the polynomial degree l for the approximation of the diffusive flux such that $l \in \{k-1, k\}$. The relevant discrete spaces are

$$\Sigma_h := [\mathbb{P}_d^l(\mathcal{T}_h)]^d, \quad U_h := \mathbb{P}_d^k(\mathcal{T}_h), \quad X_h := \Sigma_h \times U_h.$$

The discrete problem consists in finding $(\sigma_h, u_h) \in X_h$ such that

$$\begin{cases} m(\sigma_h, \tau_h) + b_h(\tau_h, u_h) = 0 & \forall \tau_h \in \Sigma_h, \\ -b_h(\sigma_h, v_h) + \hat{s}_h^{\text{sip}}(u_h, v_h) = \int_{\Omega} f v_h & \forall v_h \in U_h, \end{cases} \quad (4.53)$$

with discrete bilinear form

$$b_h(\tau_h, v_h) := \int_{\Omega} \tau_h \cdot G_h^l(v_h),$$

where the discrete gradient operator G_h^l is defined by (4.44) and the stabilization bilinear form \hat{s}_h^{sip} by (4.47).

Proposition 4.39 (Elimination of discrete diffusive flux). *The pair $(\sigma_h, u_h) \in X_h$ solves (4.53) if and only if*

$$\sigma_h = -G_h^l(u_h), \quad (4.54)$$

and $u_h \in U_h$ is such that

$$\int_{\Omega} G_h^l(u_h) \cdot G_h^l(v_h) + \hat{s}_h^{\text{sip}}(u_h, v_h) = \int_{\Omega} f v_h \quad \forall v_h \in U_h. \quad (4.55)$$

Proof. The first equation in (4.53) yields

$$\int_{\Omega} (\sigma_h + G_h^l(u_h)) \cdot \tau_h = 0 \quad \forall \tau_h \in \Sigma_h.$$

Recalling that $G_h^l(u_h) = \nabla_h u_h - \mathbf{R}_h^l(\llbracket u_h \rrbracket)$ and since $l \geq k-1$, we infer that $G_h^l(u_h) \in \Sigma_h$; therefore, (4.54) is satisfied. Substituting this relation into the second equation of (4.53) yields (4.55). The converse is straightforward. \square

Proposition 4.39 shows that the mixed dG method (4.53) is in fact equivalent to a problem in the sole unknown u_h . In particular, the above choice for b_h and \hat{s}_h^{sip} yields the SIP method of Sect. 4.2; cf. (4.46).

Remark 4.40 ($H(\text{div}; \Omega)$ -conformity of discrete diffusive flux). One drawback of mixed dG approximations, and in particular (4.53), is that the discrete diffusive flux $\sigma_h = -G_h^l(u_h)$ is not in $H(\text{div}; \Omega)$ because its normal component is, in general, discontinuous across interfaces. This point is further examined in Sect. 5.5 where we discuss a cost-effective, locally conservative diffusive flux reconstruction obtained by postprocessing the discrete potential.

4.4.2 Numerical Fluxes

In what follows, we focus for simplicity on equal-order approximations for the potential and the diffusive flux, that is, we set $l = k$ so that $\Sigma_h := [\mathbb{P}_d^k(\mathcal{T}_h)]^d$, while, as before, $U_h := \mathbb{P}_d^k(\mathcal{T}_h)$. Similarly to Sect. 4.3.4, we can derive a local formulation by localizing test functions to a single mesh element. Let $T \in \mathcal{T}_h$, let $\zeta \in [\mathbb{P}_d^k(T)]^d$, and let $\xi \in \mathbb{P}_d^k(T)$. Integrating by parts in T , splitting the boundary integral on ∂T as a sum over the mesh faces $F \in \mathcal{F}_T$, and setting $\epsilon_{T,F} = \mathbf{n}_T \cdot \mathbf{n}_F$, we infer for the exact solution that

$$\begin{aligned} \int_T \sigma \cdot \zeta - \int_T u \nabla \cdot \zeta + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F u_F (\zeta \cdot \mathbf{n}_F) &= 0, \\ - \int_T \sigma \cdot \nabla \xi + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F (\sigma_F \cdot \mathbf{n}_F) \xi &= \int_T f \xi, \end{aligned}$$

since $\sigma = -\nabla u$ and $\nabla \cdot \sigma = f$. The traces u_F and $\sigma_F \cdot \mathbf{n}_F$ are single-valued on each interface; cf. Lemma 4.3.

At the discrete level, the general form of the mixed dG approximation is derived by introducing numerical fluxes for the discrete potential and for the discrete diffusive flux. These two numerical fluxes, which are denoted by \hat{u}_F and $\hat{\sigma}_F$ for all $F \in \mathcal{F}_h$, are single-valued on each $F \in \mathcal{F}_h$. The numerical flux \hat{u}_F is scalar-valued and the numerical flux $\hat{\sigma}_F$ is vector-valued. We obtain, for all $T \in \mathcal{T}_h$, all $\zeta \in [\mathbb{P}_d^k(T)]^d$, and all $\xi \in \mathbb{P}_d^k(T)$,

$$\int_T \sigma_h \cdot \zeta - \int_T u_h \nabla \cdot \zeta + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F \hat{u}_F (\zeta \cdot \mathbf{n}_F) = 0, \quad (4.56a)$$

$$- \int_T \sigma_h \cdot \nabla \xi + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F (\hat{\sigma}_F \cdot \mathbf{n}_F) \xi = \int_T f \xi. \quad (4.56b)$$

Lemma 4.41 (Numerical fluxes for SIP). *For the SIP method, the numerical fluxes are given by*

$$\hat{u}_F = \begin{cases} \llbracket u_h \rrbracket & \forall F \in \mathcal{F}_h^i, \\ 0 & \forall F \in \mathcal{F}_h^b, \end{cases} \quad (4.57a)$$

$$\hat{\sigma}_F = -\llbracket \nabla_h u_h \rrbracket + \eta h_F^{-1} \llbracket u_h \rrbracket \mathbf{n}_F \quad \forall F \in \mathcal{F}_h. \quad (4.57b)$$

Proof. The assertion is obtained by testing the first equation in (4.53) with $\tau_h = \zeta \chi_T$, where χ_T denotes the characteristic function of T , and testing the second equation with $v_h = \xi \chi_T$. \square

A first possible variant of the SIP method consists in keeping the definition (4.57a) for the numerical flux \hat{u}_F and defining the numerical flux $\hat{\sigma}_F$ as

$$\hat{\sigma}_F = \llbracket \sigma_h \rrbracket + \eta h_F^{-1} \llbracket u_h \rrbracket \mathbf{n}_F.$$

The resulting dG method belongs to the class of LDG methods. The discrete diffusive flux σ_h can still be eliminated locally (since the numerical flux \hat{u}_F only depends on u_h), and the discrete potential $u_h \in U_h$ is such that

$$a_h^{\text{ldg}}(u_h, v_h) = \int_{\Omega} f v_h \quad \forall v_h \in U_h,$$

with the discrete bilinear form

$$\begin{aligned} a_h^{\text{ldg}}(u_h, v_h) &= \int_{\Omega} \nabla_h u_h \cdot \nabla_h v_h - \sum_{F \in \mathcal{F}_h} \int_F (\{\nabla_h u_h\} \cdot \mathbf{n}_F \llbracket v_h \rrbracket + \{\nabla_h v_h\} \cdot \mathbf{n}_F \llbracket u_h \rrbracket) \\ &\quad + \int_{\Omega} R_h^k(\llbracket u_h \rrbracket) \cdot R_h^k(\llbracket v_h \rrbracket) + \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} \int_F \llbracket u_h \rrbracket \llbracket v_h \rrbracket \\ &= \int_{\Omega} G_h^k(u_h) \cdot G_h^k(v_h) + \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} \int_F \llbracket u_h \rrbracket \llbracket v_h \rrbracket. \end{aligned}$$

A nice feature of the discrete bilinear form a_h^{ldg} is that discrete coercivity holds on U_h with respect to the $\|\cdot\|_{\text{sip}}$ -norm for any $\eta > 0$ (a simple choice is $\eta = 1$). The drawback is that the elementary *stencil* associated with the term $\int_{\Omega} R_h^k(\llbracket u_h \rrbracket) \cdot R_h^k(\llbracket v_h \rrbracket)$ consists of a given mesh element, its neighbors, and the neighbors of its neighbors in the sense of faces; cf. Fig. 4.2. Such a stencil is considerably larger than that associated with the SIP method; cf. Fig. 4.1.

More general forms of the LDG method can be designed with the numerical fluxes

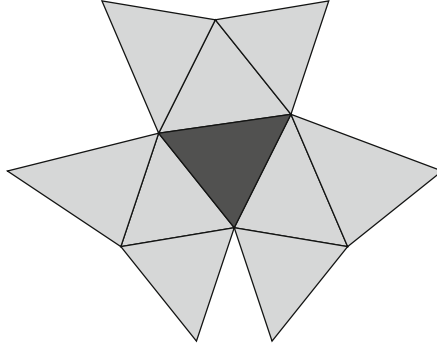


Fig. 4.2: Example of LDG stencil of an element $T \in \mathcal{T}_h$ when \mathcal{T}_h is a matching triangular mesh; the mesh element is highlighted in *dark grey*, and all the nine other elements, highlighted in *light grey*, also belong to the stencil (compare with Fig. 4.1)

$$\begin{aligned}\hat{u}_F &= \begin{cases} \llbracket u_h \rrbracket + \Upsilon \cdot \mathbf{n}_F \llbracket u_h \rrbracket & \forall F \in \mathcal{F}_h^i, \\ 0 & \forall F \in \mathcal{F}_h^b, \end{cases} \\ \hat{\sigma}_F &= \begin{cases} \llbracket \sigma_h \rrbracket - \Upsilon \llbracket \sigma_h \rrbracket \cdot \mathbf{n}_F + \eta h_F^{-1} \llbracket u_h \rrbracket \mathbf{n}_F & \forall F \in \mathcal{F}_h^i, \\ \sigma_h + \eta h_F^{-1} u_h \mathbf{n} & \forall F \in \mathcal{F}_h^b, \end{cases}\end{aligned}$$

where Υ is vector-valued and $\eta > 0$ is scalar-valued (in LDG methods, ηh_F^{-1} is often denoted by C_{11} and Υ by C_{12}). Since the numerical flux \hat{u}_F only depends on u_h , the discrete diffusive flux σ_h can be eliminated locally. The above form of the diffusive fluxes ensures symmetry and discrete stability for the resulting dG method. A simple choice for the penalty parameter is again $\eta = 1$, while the auxiliary vector-parameter Υ can be freely chosen. LDG methods for the Poisson problem have been extensively analyzed by Castillo, Cockburn, Perugia, and Schötzau [80] (see also Castillo, Cockburn, Schötzau, and Schwab [81] for the hp -analysis of LDG methods applied to diffusion-advection problems). In [80], various choices of the penalty parameter C_{11} are discussed; the above choice $C_{11} = \eta h_F^{-1}$ leads to the same energy-norm error estimates as for the SIP method. A particular choice for the vector-parameter Υ leading to superconvergence on Cartesian grids has been studied by Cockburn, Kanschat, Perugia, and Schötzau [100]. Variants of the LDG method aiming at reducing the stencil have been discussed by Sherwin, Kirby, Peiró, Taylor, and Zienkiewicz [277], Peraire and Persson [256], and Castillo [79].

A further variant of the SIP and LDG methods consists in considering the numerical fluxes

$$\begin{aligned}\hat{u}_F &= \begin{cases} \llbracket u_h \rrbracket + \eta_\sigma \llbracket \sigma_h \rrbracket \cdot \mathbf{n}_F & \forall F \in \mathcal{F}_h^i, \\ 0 & \forall F \in \mathcal{F}_h^b, \end{cases} \\ \hat{\sigma}_F &= \llbracket \sigma_h \rrbracket + \eta_u \llbracket u_h \rrbracket \mathbf{n}_F \quad \forall F \in \mathcal{F}_h.\end{aligned}$$

Here, the penalty parameters η_u and η_σ are positive user-dependent real numbers, and a simple choice is to set $\eta_u = \eta_\sigma = 1$. This method is analyzed in Sect. 7.3 in the more general context of Friedrichs' systems. Because the numerical flux \hat{u}_F depends on σ_h , (4.56a) can no longer be used to express locally the discrete diffusive flux σ_h in terms of the discrete potential u_h . This precludes the local elimination of σ_h and, therefore, enhances the computational cost of the approximation method. The approach presents, however, some advantages since it can be used with polynomial degree $k = 0$ and there is no minimal threshold on the penalty parameters (apart from being positive). Moreover, the approximation on the diffusive flux is more accurate yielding convergence rates in the L^2 -norm of order $h^{k+1/2}$ for smooth solutions, as opposed to the convergence rates of order h^k delivered by the SIP method (cf. (4.26)).

Finally, we mention that an even more general presentation can allow for two-valued numerical fluxes at interfaces; see Arnold, Brezzi, Cockburn, and Marini [16] for a unified analysis of dG methods. Two-valued numerical fluxes are

obtained, for instance, when rewriting the nonsymmetric dG methods discussed in Sect. 5.3.1 as mixed dG methods.

4.4.3 Hybrid Mixed dG Methods

The key idea in hybrid mixed dG methods is to introduce additional degrees of freedom at interfaces, thereby allowing one to eliminate locally both the discrete potential and the discrete diffusive flux. Herein, we focus on the HDG methods introduced by Cockburn, Gopalakrishnan, and Lazarov [97]; see also Causin and Sacco [83] for a different approach based on a discontinuous Petrov–Galerkin formulation, Droniou and Eymard [135] for similar ideas in the context of hybrid mixed finite volume schemes, and Ewing, Wang, and Yang for hybrid primal dG methods [154].

In the HDG method, the additional degrees of freedom are used to enforce the continuity of the normal component of the discrete diffusive flux. These additional degrees of freedom act as Lagrange multipliers in the discrete problem and can be interpreted as single-valued traces of the discrete potential on interfaces. We introduce the discrete space

$$\Lambda_h := \bigoplus_{F \in \mathcal{F}_h^i} \mathbb{P}_{d-1}^k(F).$$

A function $\mu_h \in \Lambda_h$ is such that, for all $F \in \mathcal{F}_h^i$, $\mu_h|_F \in \mathbb{P}_{d-1}^k(F)$. The discrete unknowns $(\sigma_h, u_h, \lambda_h) \in \Sigma_h \times U_h \times \Lambda_h$ satisfy the following local problems: For all $T \in \mathcal{T}_h$, all $\zeta \in [\mathbb{P}_d^k(T)]^d$, and all $\xi \in \mathbb{P}_d^k(T)$,

$$\int_T \sigma_h \cdot \zeta - \int_T u_h \nabla \cdot \zeta + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F \hat{u}_F(\zeta \cdot \mathbf{n}_F) = 0, \quad (4.58a)$$

$$- \int_T \sigma_h \cdot \nabla \xi + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F (\hat{\sigma}_{T,F} \cdot \mathbf{n}_F) \xi = \int_T f \xi, \quad (4.58b)$$

while normal diffusive flux continuity is enforced by setting, for all $F \in \mathcal{F}_T \cap \mathcal{F}_h^i$ and all $\mu \in \mathbb{P}_{d-1}^k(F)$,

$$\int_F \llbracket \hat{\sigma}_{T,F} \rrbracket \cdot \mathbf{n}_F \mu = 0. \quad (4.59)$$

Here, the numerical fluxes are such that

$$\hat{u}_F = \begin{cases} \lambda_h & \forall F \in \mathcal{F}_h^i, \\ 0 & \forall F \in \mathcal{F}_h^b, \end{cases} \quad (4.60a)$$

$$\hat{\sigma}_{T,F} = \sigma_h|_T + \tau_T(u_h|_T - \hat{u}_F) \mathbf{n}_T \quad \forall F \in \mathcal{F}_h, \quad (4.60b)$$

with penalty parameter τ_T defined elementwise. We observe that (4.59) indeed enforces $\llbracket \hat{\sigma}_{T,F} \rrbracket \cdot \mathbf{n}_F = 0$ for all $F \in \mathcal{F}_h^i$ since $\llbracket \hat{\sigma}_{T,F} \rrbracket \cdot \mathbf{n}_F \in \mathbb{P}_{d-1}^k(F)$. As a result, the quantity $(\hat{\sigma}_{T,F} \cdot \mathbf{n}_F)$ in (4.58b) is actually single-valued.

Lemma 4.42 (HDG as mixed dG method). *Let $(\sigma_h, u_h, \lambda_h) \in \Sigma_h \times U_h \times \Lambda_h$ solve (4.58) and (4.59). Then, the pair $(\sigma_h, u_h) \in \Sigma_h \times U_h$ solves the local problems of the mixed dG formulation (4.56) with numerical fluxes such that, for all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$,*

$$\hat{u}_F = \{\!\!\{u_h\}\!\!\} + C_{12} \llbracket u_h \rrbracket \mathbf{n}_F + C_{22} \llbracket \sigma_h \rrbracket \cdot \mathbf{n}_F, \quad (4.61a)$$

$$\hat{\sigma}_F = \{\!\!\{\sigma_h\}\!\!\} + C_{11} \llbracket u_h \rrbracket \mathbf{n}_F - C_{12} \llbracket \sigma_h \rrbracket \cdot \mathbf{n}_F, \quad (4.61b)$$

with the parameters

$$C_{11} = \frac{\tau_1 \tau_2}{\tau_1 + \tau_2}, \quad C_{12} = \frac{\tau_1 - \tau_2}{2(\tau_1 + \tau_2)} \mathbf{n}_F, \quad C_{22} = \frac{1}{\tau_1 + \tau_2},$$

where $\tau_i := \tau_{T_i}$, $i \in \{1, 2\}$. Moreover, for all $F \in \mathcal{F}_h^b$ with $F = \partial T \cap \partial \Omega$, $\hat{u}_F = 0$ and $\hat{\sigma}_F = \sigma_h + \tau T u_h$.

Proof. Since $\llbracket \hat{\sigma}_{T,F} \rrbracket \cdot \mathbf{n}_F = 0$, we infer from (4.60b) that

$$\llbracket \sigma_h \rrbracket \cdot \mathbf{n}_F + 2 \{\!\!\{\tau u_h\}\!\!\} - 2 \{\!\!\{\tau\}\!\!\} \hat{u}_F = 0.$$

Observing that $\{\!\!\{\tau u_h\}\!\!\} = \{\!\!\{\tau\}\!\!\} \{\!\!\{u_h\}\!\!\} + \frac{1}{4} \llbracket \tau \rrbracket \llbracket u_h \rrbracket$, we obtain

$$\hat{u}_F = \{\!\!\{u_h\}\!\!\} + \frac{1}{4} \frac{\llbracket \tau \rrbracket}{\{\!\!\{\tau\}\!\!\}} \llbracket u_h \rrbracket + \frac{1}{2 \{\!\!\{\tau\}\!\!\}} \llbracket \sigma_h \rrbracket \cdot \mathbf{n}_F,$$

which yields (4.61a). Moreover, since the normal component of $\hat{\sigma}_{T,F}$ is single-valued, we infer

$$\hat{\sigma}_{T,F} \cdot \mathbf{n}_F = \{\!\!\{\sigma_h\}\!\!\} \cdot \mathbf{n}_F + \frac{1}{2} \llbracket \tau u_h \rrbracket - \frac{1}{2} \llbracket \tau \rrbracket \hat{u}_F.$$

Observing that $\llbracket \tau u_h \rrbracket = \llbracket \tau \rrbracket \{\!\!\{u_h\}\!\!\} + \{\!\!\{\tau\}\!\!\} \llbracket u_h \rrbracket$, we obtain

$$\hat{\sigma}_{T,F} \cdot \mathbf{n}_F = \{\!\!\{\sigma_h\}\!\!\} \cdot \mathbf{n}_F + \frac{1}{2} \llbracket \tau \rrbracket (\{\!\!\{u_h\}\!\!\} - \hat{u}_F) + \frac{1}{2} \{\!\!\{\tau\}\!\!\} \llbracket u_h \rrbracket.$$

Using (4.61a) to evaluate \hat{u}_F in this expression and rearranging terms leads to

$$\hat{\sigma}_{T,F} \cdot \mathbf{n}_F = \{\!\!\{\sigma_h\}\!\!\} \cdot \mathbf{n}_F + \frac{\tau_1 \tau_2}{\tau_1 + \tau_2} \llbracket u_h \rrbracket - \frac{\tau_1 - \tau_2}{2(\tau_1 + \tau_2)} \llbracket \sigma_h \rrbracket \cdot \mathbf{n}_F.$$

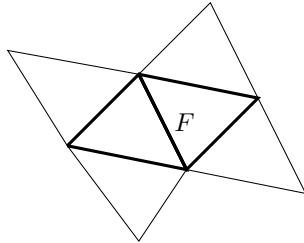


Fig. 4.3: Stencil $\mathcal{S}(F)$ for HDG methods

An inspection of (4.61b) shows that $\hat{\sigma}_{T,F \cdot \mathbf{n}_F} = \hat{\sigma}_{F \cdot \mathbf{n}_F}$, and this concludes the proof. \square

We observe that the numerical flux \hat{u}_F in (4.61a) depends on σ_h since $C_{22} \neq 0$. As a result, the discrete diffusive flux cannot be eliminated locally to derive a discrete problem for the sole discrete potential. Instead, a computationally efficient implementation of HDG methods consists in using (4.58) to eliminate locally both the discrete potential and the discrete diffusive flux by static condensation (similarly to mixed finite element methods; see, e.g., Arnold and Brezzi [15]), so as to obtain, using (4.59), a discrete problem where the sole unknown is $\lambda_h \in \Lambda_h$. For a given interface $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$, the stencil associated with this interface is (cf. Fig. 4.3)

$$\mathcal{S}(F) = \{F' \in \mathcal{F}_h^i \mid F' \in \mathcal{F}_{T_1} \cup \mathcal{F}_{T_2}\}.$$

For matching simplicial meshes, the set $\mathcal{S}(F)$ contains five interfaces for $d = 2$ and seven interfaces for $d = 3$.

HDG methods for elliptic problems have been analyzed by Cockburn, Dong, and Guzmán [95] and Cockburn, Guzmán, and Wang [98] where error estimates in various norms are derived for various choices of the penalty parameter τ . In particular, L^2 -norm error estimates of order h^{k+1} can be derived both for the potential and the diffusive flux for smooth solutions and polynomial order $k \geq 0$. Moreover, for $k \geq 1$, a postprocessed potential converging at order h^{k+2} can be derived, similarly to classical mixed finite element methods. The extension of HDG methods to diffusion-advection methods is investigated by Cockburn, Dong, Guzmán, Restelli, and Sacco [96] and Nguyen, Peraire, and Cockburn [245].

4.5 Heterogeneous Diffusion

In this section, we consider a model problem with heterogeneous diffusion. To approximate this problem using dG methods, we revisit the design and analysis of the SIP method considered in Sect. 4.2 for the Poisson problem. Following Dryja [136], we use diffusion-dependent weights to formulate the consistency and symmetry terms in the discrete bilinear form and we penalize interface and boundary jumps using a diffusion-dependent parameter scaling as the harmonic mean of the diffusion coefficient. Such a penalty strategy is particularly important in heterogeneous diffusion-advection-reaction equations (cf. Sect. 4.6) where the diffusion coefficient takes locally small values leading to so-called advection-dominated regimes. In this context, the exact solution exhibits sharp inner layers which, in practice, are not resolved by the underlying meshes, so that excessive penalty at such layers triggers spurious oscillations. Using the harmonic mean of the diffusion coefficient to penalize jumps turns out to tune automatically the amount of penalty and thereby avoid such oscillations. Incidentally, we also observe that, in finite volume and mixed finite element schemes, the harmonic mean of the diffusion coefficient is often considered at interfaces.

4.5.1 The Continuous Setting

Let $\kappa \in L^\infty(\Omega)$ be the diffusion coefficient and assume that κ is uniformly bounded from below in Ω by a positive real number. The anisotropic case, where κ is actually $\mathbb{R}^{d,d}$ -valued, is examined in Sect. 4.5.6. We are interested in the problem

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

with source term $f \in L^2(\Omega)$. The weak form of this problem is

$$\text{Find } u \in V \text{ s.t. } a(u, v) = \int_{\Omega} f v \text{ for all } v \in V, \quad (4.62)$$

with energy space $V = H_0^1(\Omega)$ and bilinear form

$$a(u, v) := \int_{\Omega} \kappa \nabla u \cdot \nabla v.$$

Owing to the above assumptions on the diffusion coefficient κ , the Lax–Milgram Lemma implies that (4.62) is well-posed. The case where κ is constant in Ω yields, up to rescaling, the Poisson problem; the latter can thus be viewed as a prototype for homogeneous diffusion problems.

Adopting the terminology used for the Poisson problem (cf. Definition 4.1), the \mathbb{R}^d -valued function

$$\sigma := -\kappa \nabla u$$

is termed the *diffusive flux*. By construction, $\sigma \in H(\text{div}; \Omega)$.

In practice, the diffusion coefficient has more regularity than just belonging to $L^\infty(\Omega)$. Henceforth, we make the following assumption.

Assumption 4.43 (Partition of Ω). *There is a partition $P_\Omega := \{\Omega_i\}_{1 \leq i \leq N_\Omega}$ of Ω such that*

- (i) *Each Ω_i , $1 \leq i \leq N_\Omega$, is a polyhedron;*
- (ii) *The restriction of κ to each Ω_i , $1 \leq i \leq N_\Omega$, is constant.*

Remark 4.44 (Motivation for assumption 4.43). In groundwater flow applications, the partition P_Ω results for instance from the partitioning of the porous medium into various geological layers.

From a physical viewpoint, the normal component of the diffusive flux σ is continuous across any interface $\partial\Omega_i \cap \partial\Omega_j$ with positive $(d-1)$ -dimensional Hausdorff measure. Assuming $\kappa|_{\Omega_i} \neq \kappa|_{\Omega_j}$, this implies that the normal component of ∇u cannot be continuous across this interface. This fact modifies the regularity that can be expected for the exact solution in heterogeneous diffusion problems with respect to the Poisson problem.

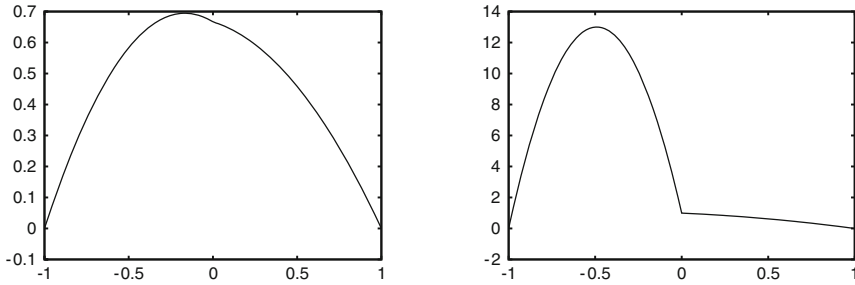


Fig. 4.4: Exact solution with diffusion heterogeneity parameter $\alpha = 0.5$ (left) and $\alpha = 0.01$ (right); the two panels use different vertical scales

4.5.1.1 One-Dimensional Example

Let $\Omega = (-1, 1)$ be partitioned into two subdomains $\Omega_1 = (-1, 0)$ and $\Omega_2 = (0, 1)$ such that $\kappa|_{\Omega_1} = \alpha$ and $\kappa|_{\Omega_2} = 1$ with positive parameter α . The exact solution of (4.62) with $f \equiv 1$ is

$$u(x) = \begin{cases} a_1(1+x)^2 + b_1(1+x) & \text{if } x \in \Omega_1, \\ a_2(x-1)^2 + b_2(x-1) & \text{if } x \in \Omega_2, \end{cases}$$

where $a_1 = -\frac{1}{2\alpha}$, $a_2 = -\frac{1}{2}$, $b_1 = \frac{1+3\alpha}{2\alpha(1+\alpha)}$, and $b_2 = -\frac{\alpha+3}{2(1+\alpha)}$. Figure 4.4 presents the exact solutions obtained with $\alpha = 0.5$ (mild diffusion heterogeneity) and $\alpha = 0.01$ (strong diffusion heterogeneity). As expected, the exact solution is only continuous at $x = 0$, but not differentiable, and the jump in the derivative of the exact solution is more pronounced in the case of strong diffusion heterogeneity. Interestingly, the maximum value attained by the exact solution in Ω is substantially affected by the diffusion heterogeneity.

4.5.1.2 Two-Dimensional Example

In dimension $d \geq 2$, discontinuities in the diffusion coefficient can cause severe singularities in the exact solution. Exact solutions of two-dimensional heterogeneous diffusion problems with zero right-hand side are explicitly constructed by Kellogg [210]. A typical situation is the case where $\Omega = (-1, 1)^2$ is divided into four quadrants, and the diffusion coefficient takes the value κ_1 in the first and third quadrants and the value κ_2 in the second and fourth quadrants. Then, it is possible to construct an exact solution with zero source term and suitable nonhomogeneous Dirichlet boundary conditions such that, in polar coordinates, $u(r, \theta) = r^\gamma v(\theta)$ with a smooth function v . The exponent $\gamma > 0$ can be made as small as desired by taking large values of the ratio κ_1/κ_2 . Figure 4.5 illustrates the exact solution for $\kappa_1/\kappa_2 = 5$ (left) and $\kappa_1/\kappa_2 = 100$ (right). In dimension 2, regularity results take the form $u \in H^{1+\epsilon}(\Omega)$ with $\epsilon > 0$ but arbitrary small. In dimension 3, regularity results have been obtained by Nicaise and Sändig [247] in some particular situations.

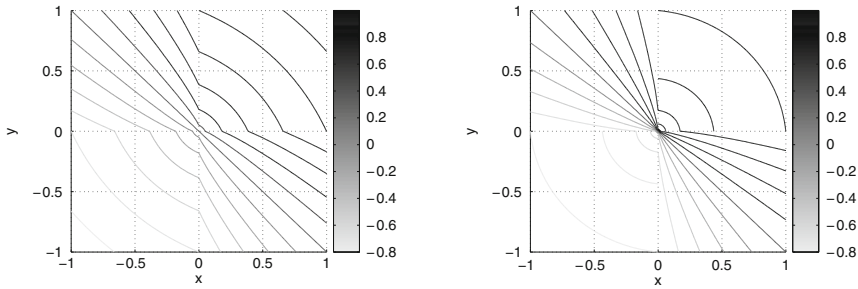


Fig. 4.5: Exact solution for heterogeneous diffusion problem; $\Omega = (-1, 1)^2$ is divided into four quadrants, and the diffusion coefficient takes the value κ_1 in the first and third quadrants and the value κ_2 in the second and fourth quadrants; *left*: $\kappa_1/\kappa_2 = 5$; *right*: $\kappa_1/\kappa_2 = 100$ (courtesy M. Vohralík)

4.5.2 Discretization

We aim at approximating the exact solution u of (4.62) by a dG method using the discrete space

$$V_h := \mathbb{P}_d^k(\mathcal{T}_h),$$

where $\mathbb{P}_d^k(\mathcal{T}_h)$ is defined by (1.15) with polynomial degree $k \geq 1$ and \mathcal{T}_h belonging to an admissible mesh sequence. We consider the discrete problem:

$$\text{Find } u_h \in V_h \text{ s.t. } a_h^{\text{swip}}(u_h, v_h) = \int_{\Omega} f v_h \text{ for all } v_h \in V_h, \quad (4.63)$$

with the discrete bilinear form a_h^{swip} yet to be designed.

4.5.2.1 Mesh Compatibility

An important assumption on the mesh sequence $\mathcal{T}_{\mathcal{H}} := (\mathcal{T}_h)_{h \in \mathcal{H}}$ is its compatibility with the partition P_{Ω} .

Assumption 4.45 (Mesh compatibility). *We suppose that the admissible mesh sequence $\mathcal{T}_{\mathcal{H}}$ is such that, for each $h \in \mathcal{H}$, each $T \in \mathcal{T}_h$ is a subset of only one set Ω_i of the partition P_{Ω} . In this situation, the meshes are said to be compatible with the partition P_{Ω} .*

An example of compatible mesh is presented in Fig. 4.6. The motivation for the above assumption is to prevent jumps of the diffusion coefficient κ to occur inside mesh elements. Indeed, owing to Assumption 4.43, the diffusion coefficient is piecewise constant on each mesh \mathcal{T}_h . This fact is often used in what follows. The present setting can be enlarged, at the price of additional technicalities, by assuming that the diffusion coefficient is piecewise smooth (e.g., piecewise Lipschitz continuous). However, it is not reasonable to envisage a high-order dG method to approximate an heterogeneous diffusion problem if the mesh is not

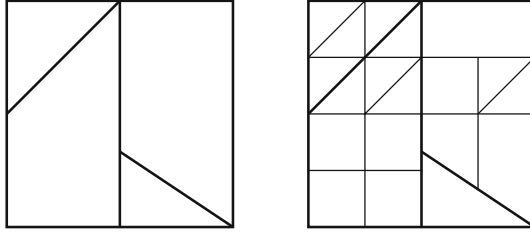


Fig. 4.6: Partition P_Ω (left) and compatible mesh (right)

compatible with the singularities of the diffusion coefficient. Indeed, the exact solution is not expected to be sufficiently smooth across these singularities to exploit the local degrees of freedom in the polynomial space.

4.5.2.2 Weighted Averages

While we keep Definitions 1.17 and 4.2 for interface and boundary jumps respectively, it is convenient to introduce weighted averages.

Definition 4.46 (Weighted averages). To any interface $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$, we assign two nonnegative real numbers $\omega_{T_1,F}$ and $\omega_{T_2,F}$ such that

$$\omega_{T_1,F} + \omega_{T_2,F} = 1.$$

Then, for any scalar-valued function v defined on Ω that is smooth enough to admit a possibly two-valued trace on all $F \in \mathcal{F}_h^i$, we define its *weighted average* on F such that, for a.e. $x \in F$,

$$\{v\}_{\omega,F}(x) := \omega_{T_1,F} v|_{T_1}(x) + \omega_{T_2,F} v|_{T_2}(x).$$

On boundary faces $F \in \mathcal{F}_h^b$ with $F = \partial T \cap \partial \Omega$, we set $\{v\}_{\omega,F}(x) := v|_T(x)$. When v is vector-valued, the weighted average operator acts componentwise on the function v . Whenever no confusion can arise, the subscript F and the variable x are omitted and we simply write $\{v\}_\omega$.

Clearly, the usual (arithmetic) average of Definition 1.17 at interfaces corresponds to the particular choice $\omega_{T_1,F} = \omega_{T_2,F} = 1/2$. Henceforth, we consider a specific diffusion-dependent choice for the weights, namely, for all $F \in \mathcal{F}_h^i$, $F = \partial T_1 \cap \partial T_2$,

$$\omega_{T_1,F} := \frac{\kappa_2}{\kappa_1 + \kappa_2}, \quad \omega_{T_2,F} := \frac{\kappa_1}{\kappa_1 + \kappa_2},$$

where $\kappa_i = \kappa|_{T_i}$, $i \in \{1, 2\}$. In particular, the case of homogeneous diffusion yields the usual (arithmetic) averages.

4.5.2.3 The SWIP Bilinear Form

In the context of heterogeneous diffusion problems, we modify the SIP bilinear form defined by (4.12) as follows: For all $(v_h, y_h) \in V_h \times V_h$,

$$\begin{aligned} a_h^{\text{swip}}(v_h, y_h) := & \int_{\Omega} \kappa \nabla_h v_h \cdot \nabla_h y_h + \sum_{F \in \mathcal{F}_h} \eta \frac{\gamma_{\kappa, F}}{h_F} \int_F \llbracket v_h \rrbracket \llbracket y_h \rrbracket \\ & - \sum_{F \in \mathcal{F}_h} \int_F (\llbracket \kappa \nabla_h v_h \rrbracket_{\omega} \cdot \mathbf{n}_F \llbracket y_h \rrbracket + \llbracket v_h \rrbracket \llbracket \kappa \nabla_h y_h \rrbracket_{\omega} \cdot \mathbf{n}_F). \end{aligned} \quad (4.64)$$

The quantity $\eta > 0$ denotes a user-dependent penalty parameter which is independent of the diffusion coefficient, while the diffusion-dependent penalty parameter $\gamma_{\kappa, F}$ is such that for all $F \in \mathcal{F}_h^i$, $F = \partial T_1 \cap \partial T_2$,

$$\gamma_{\kappa, F} := \frac{2\kappa_1 \kappa_2}{\kappa_1 + \kappa_2},$$

where, as above, $\kappa_i = \kappa|_{T_i}$, $i \in \{1, 2\}$, while, for all $F \in \mathcal{F}_h^b$, $F = \partial T \cap \partial \Omega$,

$$\gamma_{\kappa, F} := \kappa|_T.$$

We notice that the above choice for the penalty parameter $\gamma_{\kappa, F}$ on interfaces corresponds to the *harmonic mean* of the values of the diffusion coefficient on either side of the interface. Furthermore, we observe that, for all $F \in \mathcal{F}_h^i$,

$$\gamma_{\kappa, F} \leq 2 \min(\kappa_1, \kappa_2). \quad (4.65)$$

This property is used in the convergence analysis of Sect. 4.6.3 in the context of diffusion-advection-reaction problems; cf., in particular, Remark 4.65.

The bilinear form a_h^{swip} defined by (4.64) is termed the Symmetric Weighted Interior Penalty (SWIP) bilinear form. It has been introduced by Dryja [136] for heterogeneous diffusion problems and analyzed (in the more general context of diffusion-advection-reaction problems) by Di Pietro, Ern, and Guermond [133] and Ern, Stephansen, and Zunino [150]. The two differences with respect to the more usual SIP bilinear form are the use of (diffusion-dependent) weighted averages to formulate the consistency and symmetry terms and the presence of the diffusion-dependent penalty parameter. Whenever κ is constant in Ω , the usual (arithmetic) averages are recovered in the consistency and symmetry terms. The possibility of using non-arithmetic averages in dG methods has been pointed out and used in various contexts, e.g., by Stenberg [282], by Heinrich and co-workers [188–190], and by Hansbo and Hansbo [182] in the context of unfitted finite element methods based on Nitsche’s method. The idea of connecting the actual value of the weights to the diffusion coefficient was also considered by Burman and Zunino [73] in the context of mortaring techniques for a singularly perturbed diffusion-advection equation.

Lemma 4.47 (Reformulation of SWIP bilinear form). *There holds, for all $(v_h, y_h) \in V_h \times V_h$,*

$$\begin{aligned} a_h^{\text{swip}}(v_h, y_h) = & - \sum_{T \in \mathcal{T}_h} \int_T \nabla \cdot (\kappa \nabla v_h) y_h + \sum_{F \in \mathcal{F}_h} \eta \frac{\gamma_{\kappa, F}}{h_F} \int_F \llbracket v_h \rrbracket \llbracket y_h \rrbracket \\ & + \sum_{F \in \mathcal{F}_h^i} \int_F \llbracket \kappa \nabla_h v_h \rrbracket \cdot \mathbf{n}_F \llbracket y_h \rrbracket_{\overline{\omega}} - \sum_{F \in \mathcal{F}_h} \int_F \llbracket v_h \rrbracket \llbracket \kappa \nabla_h y_h \rrbracket_{\omega} \cdot \mathbf{n}_F, \end{aligned} \quad (4.66)$$

where $\llbracket y_h \rrbracket_{\overline{\omega}}$ is the skew-weighted average value of y_h defined as

$$\llbracket y_h \rrbracket_{\overline{\omega}} := \omega_{T_2, F} y_h|_{T_1} + \omega_{T_1, F} y_h|_{T_2}.$$

Proof. Integrating by parts the first term in (4.64) yields

$$\int_{\Omega} \kappa \nabla_h v_h \cdot \nabla_h y_h = - \sum_{T \in \mathcal{T}_h} \int_T \nabla \cdot (\kappa \nabla v_h) y_h + \sum_{T \in \mathcal{T}_h} \int_{\partial T} \kappa (\nabla v_h \cdot \mathbf{n}_T) y_h. \quad (4.67)$$

Rearranging the second term on the right-hand side as a sum over mesh faces leads to

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} \kappa (\nabla v_h \cdot \mathbf{n}_T) y_h = \sum_{F \in \mathcal{F}_h^i} \int_F \llbracket (\kappa \nabla_h v_h) y_h \rrbracket \cdot \mathbf{n}_F + \sum_{F \in \mathcal{F}_h^b} \int_F \kappa (\nabla v_h \cdot \mathbf{n}) y_h.$$

We now observe that, for all $F \in \mathcal{F}_h^i$,

$$\llbracket (\kappa \nabla_h v_h) y_h \rrbracket = \llbracket \kappa \nabla_h v_h \rrbracket_{\omega} \llbracket y_h \rrbracket + \llbracket \kappa \nabla_h v_h \rrbracket \llbracket y_h \rrbracket_{\overline{\omega}}.$$

To prove this identity, we set $a_i = (\kappa \nabla_h v_h)|_{T_i}$, $b_i = y_h|_{T_i}$, $\omega_i = \omega_{T_i, F}$, $i \in \{1, 2\}$, so that

$$\begin{aligned} \llbracket (\kappa \nabla_h v_h) y_h \rrbracket &= a_1 b_1 - a_2 b_2 \\ &= (\omega_1 a_1 + \omega_2 a_2)(b_1 - b_2) + (a_1 - a_2)(\omega_2 b_1 + \omega_1 b_2) \\ &= \llbracket \kappa \nabla_h v_h \rrbracket_{\omega} \llbracket y_h \rrbracket + \llbracket \kappa \nabla_h v_h \rrbracket \llbracket y_h \rrbracket_{\overline{\omega}}, \end{aligned}$$

since $\omega_1 + \omega_2 = 1$. As a result and accounting for boundary faces,

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} \kappa (\nabla v_h \cdot \mathbf{n}_T) y_h = \sum_{F \in \mathcal{F}_h} \int_F \llbracket \kappa \nabla_h v_h \rrbracket_{\omega} \cdot \mathbf{n}_F \llbracket y_h \rrbracket + \sum_{F \in \mathcal{F}_h^i} \int_F \llbracket \kappa \nabla_h v_h \rrbracket \cdot \mathbf{n}_F \llbracket y_h \rrbracket_{\overline{\omega}}.$$

Combining this expression with (4.64) and (4.67) yields the assertion. \square

4.5.3 Error Estimates for Smooth Solutions

In this section, we present the convergence analysis for the discrete problem (4.63) in the case where the exact solution is smooth enough to match the following assumption.

Assumption 4.48 (Regularity of exact solution and space V_*). *We assume that the exact solution u is such that*

$$u \in V_* := V \cap H^2(P_\Omega).$$

In the spirit of Sect. 1.3, we set $V_{*h} := V_* + V_h$.

Assumption 4.48 implies that, for all $T \in \mathcal{T}_h$, letting $\sigma_T := -(\kappa \nabla u)|_T$ and $\sigma_{\partial T} = \sigma_T \cdot \mathbf{n}_T$ on ∂T , the trace $\sigma_{\partial T}|_F$ is in $L^2(F)$ for all $F \in \mathcal{F}_T$. Using Lemma 1.23 for the jumps of the potential and proceeding as in the proof of Lemma 1.24 for the jumps of the diffusive flux, we infer that the exact solution satisfies

$$[[u]] = 0 \quad \forall F \in \mathcal{F}_h, \quad (4.68a)$$

$$[[\kappa \nabla u]] \cdot \mathbf{n}_F = 0 \quad \forall F \in \mathcal{F}_h^i. \quad (4.68b)$$

The convergence analysis is performed in the spirit of Theorem 1.35 by establishing discrete coercivity, consistency, and boundedness for a_h^{swip} . The discrete bilinear form a_h^{swip} is extended to $V_{*h} \times V_h$.

Lemma 4.49 (Consistency). *Assume $u \in V_*$. Then, for all $v_h \in V_h$,*

$$a_h^{\text{swip}}(u, v_h) = \int_\Omega f v_h.$$

Proof. The result is a direct consequence of (4.66) and (4.68). \square

To formulate discrete stability in the context of heterogeneous diffusion, we modify the $\|\cdot\|_{\text{swip}}$ -norm considered for the Poisson problem (cf. (4.17)) as follows: For all $v \in V_{*h}$,

$$\|v\|_{\text{swip}} := \left(\|\kappa^{1/2} \nabla_h v\|_{[L^2(\Omega)]^d}^2 + |v|_{J, \kappa}^2 \right)^{1/2}, \quad (4.69)$$

with the diffusion-dependent jump seminorm

$$|v|_{J, \kappa} = \left(\sum_{F \in \mathcal{F}_h} \frac{\gamma_{\kappa, F}}{h_F} \|[[v]]\|_{L^2(F)}^2 \right)^{1/2}. \quad (4.70)$$

Before addressing the discrete coercivity of the SWIP bilinear form, we derive a bound on the consistency term.

Lemma 4.50 (Bound on consistency term). *For all $(v, y_h) \in V_{*h} \times V_h$,*

$$\left| \sum_{F \in \mathcal{F}_h} \int_F \{ \kappa \nabla_h v \} \omega \cdot \mathbf{n}_F [[y_h]] \right| \leq \left(\sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_T} h_F \|\kappa^{1/2} \nabla v|_T \cdot \mathbf{n}_F\|_{L^2(F)}^2 \right)^{1/2} |y_h|_{J, \kappa}. \quad (4.71)$$

Proof. For all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$, $\omega_i = \omega_{T_i, F}$, $\kappa_i = \kappa|_{T_i}$, and $a_i = \kappa_i^{1/2}(\nabla v)|_{T_i \cdot \mathbf{n}_F}$, $i \in \{1, 2\}$, the Cauchy–Schwarz inequality yields

$$\begin{aligned} \int_F \{\{\kappa \nabla_h v\}\}_{\omega \cdot \mathbf{n}_F} \llbracket y_h \rrbracket &= \int_F (\omega_1 \kappa_1^{1/2} a_1 + \omega_2 \kappa_2^{1/2} a_2) \llbracket y_h \rrbracket \\ &\leq \left(\frac{1}{2} h_F (\|a_1\|_{L^2(F)}^2 + \|a_2\|_{L^2(F)}^2) \right)^{1/2} \\ &\quad \times \left(2(\omega_1^2 \kappa_1 + \omega_2^2 \kappa_2) \frac{1}{h_F} \|\llbracket y_h \rrbracket\|_{L^2(F)}^2 \right)^{1/2}, \end{aligned}$$

and since $2(\omega_1^2 \kappa_1 + \omega_2^2 \kappa_2) = \gamma_{\kappa, F}$, we infer

$$\begin{aligned} \int_F \{\{\kappa \nabla_h v\}\}_{\omega \cdot \mathbf{n}_F} \llbracket y_h \rrbracket &\leq \left(\frac{1}{2} h_F (\|a_1\|_{L^2(F)}^2 + \|a_2\|_{L^2(F)}^2) \right)^{1/2} \\ &\quad \times \left(\frac{\gamma_{\kappa, F}}{h_F} \right)^{1/2} \|\llbracket y_h \rrbracket\|_{L^2(F)}. \end{aligned}$$

Moreover, for all $F \in \mathcal{F}_h^b$ with $F = \partial T \cap \partial \Omega$,

$$\int_F \{\{\kappa \nabla_h v\}\}_{\omega \cdot \mathbf{n}_F} \llbracket y_h \rrbracket \leq h_F^{1/2} \|(\kappa^{1/2} \nabla v)|_{T \cdot \mathbf{n}_F}\|_{L^2(F)} \times \left(\frac{\gamma_{\kappa, F}}{h_F} \right)^{1/2} \|\llbracket y_h \rrbracket\|_{L^2(F)}.$$

Summing over mesh faces, using the Cauchy–Schwarz inequality, and regrouping the face contributions for each mesh element yields the assertion. \square

We now establish the discrete coercivity of the SWIP bilinear form under the usual assumption that the penalty parameter η is large enough. An important point is that the minimal threshold on the penalty parameter is independent of the diffusion coefficient (it is actually the same as for the Poisson problem).

Lemma 4.51 (Discrete coercivity). *For all $\eta > \underline{\eta}$ with $\underline{\eta}$ defined in Lemma 4.12, the SWIP bilinear form defined by (4.64) is coercive on V_h with respect to the $\|\cdot\|_{\text{swip}}$ -norm, i.e.,*

$$\forall v_h \in V_h, \quad a_h^{\text{swip}}(v_h, v_h) \geq C_\eta \|v_h\|_{\text{swip}}^2,$$

with C_η defined in Lemma 4.12.

Proof. Let $v_h \in V_h$. Owing to the discrete trace inequality (1.40), the fact that $h_F \leq h_T$ for all $T \in \mathcal{T}_h$ and for all $F \in \mathcal{F}_T$, and since κ is piecewise constant on \mathcal{T}_h , we infer from the bound (4.71) that

$$\left| \sum_{F \in \mathcal{F}_h} \int_F \{\{\kappa \nabla_h v_h\}\}_{\omega \cdot \mathbf{n}_F} \llbracket v_h \rrbracket \right| \leq C_{\text{tr}} N_\partial^{1/2} \|\kappa^{1/2} \nabla_h v_h\|_{[L^2(\Omega)]^d} |v_h|_{J, \kappa}.$$

We conclude as in the proof of Lemma 4.12. \square

A straightforward consequence of the Lax–Milgram Lemma is that the discrete problem (4.63) is well-posed.

Our last step in the convergence analysis is to prove the boundedness of the SWIP bilinear form. To formulate this result, we define on V_{*h} the norm

$$\|v\|_{\text{swip},*} := \left(\|v\|_{\text{swip}}^2 + \sum_{T \in \mathcal{T}_h} h_T \|\kappa^{1/2} \nabla v|_T \cdot \mathbf{n}_T\|_{L^2(\partial T)}^2 \right)^{1/2}.$$

Lemma 4.52 (Boundedness). *There is C_{bnd} , independent of h and κ , such that*

$$\forall (v, y_h) \in V_{*h} \times V_h, \quad a_h^{\text{swip}}(v, y_h) \leq C_{\text{bnd}} \|v\|_{\text{swip},*} \|y_h\|_{\text{swip}}.$$

Proof. Let $(v, y_h) \in V_{*h} \times V_h$ and observe that

$$\begin{aligned} a_h^{\text{swip}}(v, y_h) &:= \int_{\Omega} \kappa \nabla_h v \cdot \nabla_h y_h + \sum_{F \in \mathcal{F}_h} \eta \frac{\gamma_{\kappa, F}}{h_F} \int_F \llbracket v \rrbracket \llbracket y_h \rrbracket \\ &\quad - \sum_{F \in \mathcal{F}_h} \int_F \{ \kappa \nabla_h v \}_\omega \cdot \mathbf{n}_F \llbracket y_h \rrbracket - \sum_{F \in \mathcal{F}_h} \int_F \llbracket v \rrbracket \{ \kappa \nabla_h y_h \}_\omega \cdot \mathbf{n}_F \\ &= \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3 + \mathfrak{T}_4. \end{aligned} \tag{4.72}$$

Using the Cauchy–Schwarz inequality yields

$$|\mathfrak{T}_1 + \mathfrak{T}_2| \leq (1 + \eta) \|v\|_{\text{swip}} \|y_h\|_{\text{swip}}.$$

Moreover, owing to the bound (4.71),

$$|\mathfrak{T}_3| \leq \|v\|_{\text{swip},*} \|y_h\|_{J, \kappa} \leq \|v\|_{\text{swip},*} \|y_h\|_{\text{swip}}$$

by definition of the $\|\cdot\|_{\text{swip},*}$ -norm. Finally, still owing to the bound (4.71) and proceeding as in the proof of Lemma 4.51 leads to

$$|\mathfrak{T}_4| \leq C_{\text{tr}} N_{\partial}^{1/2} |v|_{J, \kappa} \|\kappa^{1/2} \nabla_h y_h\|_{[L^2(\Omega)]^d} \leq C_{\text{tr}} N_{\partial}^{1/2} \|v\|_{\text{swip}} \|y_h\|_{\text{swip}}.$$

Collecting the above bounds yields the assertion with $C_{\text{bnd}} = 2 + \eta + C_{\text{tr}} N_{\partial}^{1/2}$. \square

A straightforward consequence of Theorem 1.35, together with Lemmata 1.58 and 1.59, is the following convergence result.

Theorem 4.53 ($\|\cdot\|_{\text{swip}}$ -norm error estimate and convergence rate). *Let $u \in V_*$ solve (4.62). Let u_h solve (4.63) with a_h^{swip} defined by (4.64) and penalty parameter as in Lemma 4.12. Then, there is C , independent of h and κ , such that*

$$\|u - u_h\|_{\text{swip}} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{\text{swip},*}.$$

Moreover, if $u \in H^{k+1}(P_{\Omega})$,

$$\|u - u_h\|_{\text{swip}} \leq C_u \|\kappa\|_{L^{\infty}(\Omega)}^{1/2} h^k,$$

with $C_u = C \|u\|_{H^{k+1}(P_{\Omega})}$ and C independent of h and κ .

Since the quantity C in the error estimates is independent of the diffusion coefficient κ , the approximation method is robust with respect to diffusion heterogeneities (observing that the $\|\cdot\|_{\text{swip}}$ -norm depends on κ). The convergence rate in the $\|\cdot\|_{\text{swip}}$ -norm is optimal, both for the broken gradient and the jump seminorm.

4.5.4 Error Estimates for Low-Regularity Solutions

In this section, following [132], we present the convergence analysis for the discrete problem (4.63) for an exact solution with low-regularity.

Assumption 4.54 (Regularity of exact solution and space V_*). *We assume that $d \geq 2$ and that there is $p \in (\frac{2d}{d+2}, 2]$ such that, for the exact solution u ,*

$$u \in V_* := V \cap W^{2,p}(P_\Omega).$$

In the spirit of Sect. 1.3, we set $V_{*h} := V_* + V_h$.

Assumption 4.48 implies that, for all $T \in \mathcal{T}_h$, letting $\sigma_T := -(\kappa \nabla u)|_T$ and $\sigma_{\partial T} = \sigma_T \cdot \mathbf{n}_T$ on ∂T , the trace $\sigma_{\partial T}|_F$ is in $L^p(F)$ for all $F \in \mathcal{F}_T$. We adapt the analysis of Sect. 4.2.5 for the Poisson problem to the present setting with heterogeneous diffusion.

We already know that discrete coercivity holds true provided the penalty parameter is chosen as in Lemma 4.12. Moreover, since the jump conditions (4.68) still hold true, consistency can be asserted. Thus, it only remains to prove boundedness, which we do by redefining on V_{*h} the $\|\cdot\|_{\text{swip},*}$ -norm as

$$\|v\|_{\text{swip},*} := \left(\|v\|_{\text{swip}}^p + \sum_{T \in \mathcal{T}_h} h_T^{1+\gamma_p} \|\kappa^{1/2} \nabla v|_T \cdot \mathbf{n}_T\|_{L^p(\partial T)}^p \right)^{1/p}, \quad (4.73)$$

where $\gamma_p := \frac{1}{2}d(p-2)$. We observe that, for $p = 2$, we recover the previous definition of the $\|\cdot\|_{\text{swip},*}$ -norm. The value for γ_p is motivated by the following boundedness result.

Lemma 4.55 (Boundedness). *There is C_{bnd} , independent of h and κ , such that*

$$\forall (v, w_h) \in V_{*h} \times V_h, \quad a_h^{\text{swip}}(v, w_h) \leq C_{\text{bnd}} \|v\|_{\text{swip},*} \|w_h\|_{\text{swip}}.$$

Proof. Let $(v, w_h) \in V_{*h} \times V_h$. We need to bound the four terms $\mathfrak{T}_1, \dots, \mathfrak{T}_4$ in (4.72). Proceeding as in the proof of Lemma 4.52, we obtain

$$|\mathfrak{T}_1 + \mathfrak{T}_3 + \mathfrak{T}_4| \leq C \|v\|_{\text{swip}} \|w_h\|_{\text{swip}},$$

with C independent of h and κ , so that it only remains to bound the consistency term \mathfrak{T}_2 . For all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$, and $a_i = (\kappa^{1/2} \nabla v)|_{T_i} \cdot \mathbf{n}_F$, $i \in \{1, 2\}$,

Hölder's inequality yields

$$\begin{aligned} \int_F \{\kappa \nabla_h v\}_{\omega \cdot \mathbf{n}_F} [w_h] &= \int_F (\omega_1 \kappa_1^{1/2} a_1 + \omega_2 \kappa_2^{1/2} a_2) [w_h] \\ &\leq \left(\frac{1}{2} h_F^{1+\gamma_p} (\|a_1\|_{L^p(F)}^p + \|a_2\|_{L^p(F)}^p) \right)^{1/p} \\ &\quad \times 2^{1/p} \left((\omega_1^q \kappa_1^{q/2} + \omega_2^q \kappa_2^{q/2}) h_F^{-q\beta_p} \|[w_h]\|_{L^q(F)}^q \right)^{1/q}, \end{aligned}$$

with $\beta_p = \frac{1+\gamma_p}{p}$ and $q = \frac{p}{p-1}$. Since $q \geq 2$, we obtain

$$(\omega_1^q \kappa_1^{q/2} + \omega_2^q \kappa_2^{q/2}) = \frac{(\kappa_1 \kappa_2)^{q/2}}{(\kappa_1 + \kappa_2)^q} (\kappa_1^{q/2} + \kappa_2^{q/2}) \leq \frac{(\kappa_1 \kappa_2)^{q/2}}{(\kappa_1 + \kappa_2)^q} (\kappa_1 + \kappa_2)^{q/2} = 2^{-q/2} \gamma_{\kappa, F}^{q/2}.$$

Hence, since $2^{1/p-1/2} \leq 2$,

$$\begin{aligned} \int_F \{\kappa \nabla_h v\}_{\omega \cdot \mathbf{n}_F} [w_h] &\leq \left(\frac{1}{2} h_F^{1+\gamma_p} (\|a_1\|_{L^p(F)}^p + \|a_2\|_{L^p(F)}^p) \right)^{1/p} \\ &\quad \times 2 \gamma_{\kappa, F}^{1/2} h_F^{-\beta_p} \|[w_h]\|_{L^q(F)}. \end{aligned}$$

Moreover, for all $F \in \mathcal{F}_h^b$ with $F = \partial T \cap \partial \Omega$,

$$\int_F \{\kappa \nabla_h v\}_{\omega \cdot \mathbf{n}_F} [w_h] \leq \left(h_F^{1+\gamma_p} \|\kappa^{1/2} \nabla v|_T \cdot \mathbf{n}_F\|_{L^p(F)}^p \right)^{1/p} \gamma_{\kappa, F}^{1/2} h_F^{-\beta_p} \|[w_h]\|_{L^q(F)}.$$

We can now conclude as in the proof of Lemma 4.30. \square

A straightforward consequence of Theorem 1.35 is the following convergence result. The achieved convergence rates are optimal, both for the broken gradient and the jump seminorm.

Theorem 4.56 ($\|\cdot\|_{\text{swip}}$ -norm error estimate and convergence rate). *Let $u \in V_*$ solve (4.62). Let u_h solve (4.63) with a_h^{swip} defined by (4.64) and penalty parameter as in Lemma 4.12. Then, there is C , independent of h and κ , such that*

$$\|u - u_h\|_{\text{swip}} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{\text{swip},*},$$

where the $\|\cdot\|_{\text{swip},*}$ -norm is defined by (4.73). Moreover, under Assumption 4.31, there holds

$$\|u - u_h\|_{\text{swip}} \leq C_u h^{\alpha_p},$$

with $C_u = C|u|_{W^{2,p}(P_\Omega)}$, C independent of h and κ , and $\alpha_p = \frac{d+2}{2} - \frac{d}{p} > 0$.

4.5.5 Numerical Fluxes

As for the Poisson problem in Sect. 4.3.4, it is possible to derive a local formulation of the discrete problem (4.63) by localizing test functions to mesh elements. To this purpose, we first modify the definition of the lifting operators and discrete gradients (cf. Sects. 4.3.1 and 4.3.2) to account for diffusion heterogeneities. For any face $F \in \mathcal{F}_h$ and for any integer $l \geq 0$, we define the (local) lifting operator $\mathbf{r}_{F,\kappa}^l : L^2(F) \rightarrow [\mathbb{P}_d^l(\mathcal{T}_h)]^d$ as follows: For all $\varphi \in L^2(F)$,

$$\int_{\Omega} \kappa \mathbf{r}_{F,\kappa}^l(\varphi) \cdot \boldsymbol{\tau}_h = \int_F \{\!\!\{ \kappa \boldsymbol{\tau}_h \}\!\!\}_\omega \cdot \mathbf{n}_F \varphi \quad \forall \boldsymbol{\tau}_h \in [\mathbb{P}_d^l(\mathcal{T}_h)]^d. \quad (4.74)$$

Clearly, if κ does not jump across F (so that κ is constant in the support of $\mathbf{r}_{F,\kappa}^l(\varphi)$), definitions (4.37) and (4.74) produce the same result, but this is no longer the case in the presence of diffusion heterogeneities. Then, for any function $v \in H^1(\mathcal{T}_h)$, we define the (global) lifting of its interface and boundary jumps as

$$\mathbf{R}_{h,\kappa}^l(\llbracket v \rrbracket) := \sum_{F \in \mathcal{F}_h} \mathbf{r}_{F,\kappa}^l(\llbracket v \rrbracket) \in [\mathbb{P}_d^l(\mathcal{T}_h)]^d, \quad (4.75)$$

being implicitly understood that $\mathbf{r}_{F,\kappa}^l$ acts on the function $\llbracket v \rrbracket_F$ (which is in $L^2(F)$ since $v \in H^1(\mathcal{T}_h)$). If κ is constant in Ω , definitions (4.40) and (4.75) produce the same result. Finally, the definition (4.44) of the discrete gradient is extended to the heterogeneous diffusion case by setting, for all $v \in H^1(\mathcal{T}_h)$,

$$G_{h,\kappa}^l(v) := \nabla_h v - \mathbf{R}_{h,\kappa}^l(\llbracket v \rrbracket) \in [L^2(\Omega)]^d.$$

Let $T \in \mathcal{T}_h$ and let $\xi \in \mathbb{P}_d^k(T)$. Then, using $v_h = \xi \chi_T$ as test function in (4.63) where χ_T is the characteristic function of T , proceeding as in Sect. 4.3.4, and using the above definitions, we infer

$$\int_T \kappa G_{h,\kappa}^l(u_h) \cdot \nabla \xi + \sum_{F \in \mathcal{F}_T} \epsilon_{T,F} \int_F \phi_F(u_h) \xi = \int_T f \xi,$$

with $l \in \{k-1, k\}$, $\epsilon_{T,F} = \mathbf{n}_T \cdot \mathbf{n}_F$, and the numerical flux $\phi_F(u_h)$ defined as

$$\phi_F(u_h) := -\{\!\!\{ \kappa \nabla_h u_h \}\!\!\}_\omega \cdot \mathbf{n}_F + \eta \frac{\gamma_{\kappa,F}}{h_F} \llbracket u_h \rrbracket.$$

Remark 4.57 (Harmonic means). For all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$ and $\kappa_i = \kappa|_{T_i}$, $i \in \{1, 2\}$, we observe that

$$\begin{aligned} -\{\!\!\{ \kappa \nabla_h u_h \}\!\!\}_\omega \cdot \mathbf{n}_F &= -\frac{\kappa_2}{\kappa_1 + \kappa_2} \kappa_1 (\nabla u_h)|_{T_1} \cdot \mathbf{n}_F - \frac{\kappa_1}{\kappa_1 + \kappa_2} \kappa_2 (\nabla u_h)|_{T_2} \cdot \mathbf{n}_F \\ &= -\frac{2\kappa_1 \kappa_2}{\kappa_1 + \kappa_2} \{\!\!\{ \nabla_h u_h \}\!\!\} \cdot \mathbf{n}_F. \end{aligned}$$

Thus, recalling that the jump seminorm of u_h tends to zero as $h \rightarrow 0$, the leading-order term in the numerical flux $\phi_F(u_h)$ uses the harmonic mean of the diffusion

coefficient. A motivation for using harmonic means can be given in the context of heat transfer where κ represents the thermal conductivity, u the temperature, and $-\kappa\nabla u$ the heat flux. Consider an interface between a poorly conductive medium (where κ is relatively small) and a highly conductive medium (where κ is much larger), so that, at this interface, the harmonic mean of κ is close to the value in the poorly conductive medium. Then, the heat transfer through this interface is essentially governed by the poorly conductive medium.

4.5.6 Anisotropy

The above developments can be extended to the anisotropic case, that is, when for a.e. $x \in \Omega$, $\kappa(x)$ is a symmetric tensor in $\mathbb{R}^{d,d}$. Assuming that the lowest eigenvalue of κ is uniformly bounded from below in Ω by a positive real number, the model problem (4.62) is well-posed.

The SWIP bilinear form defined by (4.64) can be used to approximate heterogeneous anisotropic diffusion problems. Specifically, the weights $\{\omega_{T_1,F}, \omega_{T_2,F}\}$ and the penalty parameter $\gamma_{\kappa,F}$ are evaluated on any interface $F \in \mathcal{F}_h^i$ by using the normal component of the diffusion tensor on both sides of that interface, that is, for all $F \in \mathcal{F}_h^i$, $F = \partial T_1 \cap \partial T_2$, we now let $\kappa_i := \mathbf{n}_F^t(\kappa|_{T_i})\mathbf{n}_F$, $i \in \{1, 2\}$, and we set as before

$$\omega_{T_1,F} := \frac{\kappa_2}{\kappa_1 + \kappa_2}, \quad \omega_{T_2,F} := \frac{\kappa_1}{\kappa_1 + \kappa_2}, \quad \gamma_{\kappa,F} := \frac{2\kappa_1\kappa_2}{\kappa_1 + \kappa_2}.$$

Moreover, for all $F \in \mathcal{F}_h^b$, $F = \partial T \cap \partial\Omega$, we set $\gamma_{\kappa,F} := \mathbf{n}^t(\kappa|_T)\mathbf{n}$. With these modifications, the convergence analysis proceeds as in the isotropic case. Since κ takes symmetric positive definite values, it is in particular possible to define $\kappa^{1/2}$ as the symmetric positive definite matrix such that $\kappa^{1/2}\kappa^{1/2} = \kappa$. We refer the reader to [133, 150] for a detailed presentation of the convergence analysis.

4.6 Diffusion-Advection-Reaction

In this section, we consider a model diffusion-advection-reaction problem. The design and analysis of the dG approximation combine the ideas of Sect. 4.5 to handle the diffusion part and those of Sect. 2.3 to handle the advection-reaction part. One issue of particular interest is the robustness of the approximation in the singularly perturbed regime where advection-reaction effects dominate over diffusion effects. In particular, we address at the end of this section the situation where the diffusion coefficient can actually vanish locally, so that a first-order PDE in some part of the domain is coupled to an elliptic PDE in the remaining part.

4.6.1 The Continuous Setting

Let $\kappa \in L^\infty(\Omega)$ and assume that κ is uniformly bounded from below in Ω by a positive real number; the singular limit where κ can actually vanish locally in

some parts of Ω is addressed in Sect. 4.6.4. Moreover, we keep Assumption 4.43 so as to localize possible jumps in the diffusion coefficient. Let $\beta \in [\text{Lip}(\Omega)]^d$ be the advective velocity and let $\tilde{\mu} \in L^\infty(\Omega)$ be the reaction coefficient. We are interested in the problem:

$$\begin{aligned} \nabla \cdot (-\kappa \nabla u + \beta u) + \tilde{\mu} u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

with source term $f \in L^2(\Omega)$. The weak form of this problem reads

$$\text{Find } u \in V \text{ s.t. } a(u, v) = \int_{\Omega} f v \text{ for all } v \in V, \quad (4.76)$$

with energy space $V = H_0^1(\Omega)$ and bilinear form

$$a(u, v) := \int_{\Omega} \kappa \nabla u \cdot \nabla v - \int_{\Omega} u \beta \cdot \nabla v + \int_{\Omega} \tilde{\mu} u v.$$

We observe that the advective term is written in conservative form. The \mathbb{R}^d -valued function

$$\Phi(u) = -\kappa \nabla u + \beta u$$

is termed the diffusive-advective flux. By construction, $\Phi(u)$ is in $H(\text{div}; \Omega)$. The diffusion-advection-reaction can be rewritten as

$$-\nabla \cdot \Phi(u) + \tilde{\mu} u = f,$$

and the bilinear form a as

$$a(u, v) = \int_{\Omega} -\Phi(u) \cdot \nabla v + \int_{\Omega} \tilde{\mu} u v. \quad (4.77)$$

Since $u \in H^1(\Omega)$ and β is smooth, it is equivalent to consider the advective term in its non-conservative form, i.e.,

$$-\nabla \cdot (\kappa \nabla u) + \beta \cdot \nabla u + \mu u = f,$$

with $\mu := \tilde{\mu} + \nabla \cdot \beta$. However, if κ vanishes locally, the exact solution can feature discontinuities, and the two forms are no longer equivalent. The conservative form is more natural from a physical viewpoint since it expresses a basic conservation principle. Indeed, integrating the diffusion-advection-reaction equation over a control volume $V \subset \Omega$, we obtain formally

$$\int_{\partial V} \Phi(u) \cdot \mathbf{n}_V + \int_V \tilde{\mu} u = \int_V f,$$

where \mathbf{n}_V denotes the outward normal to ∂V . This equation expresses the fact that the variation of u in the control volume V due to the diffusive and advective exchanges through ∂V plus the quantity of u generated/depleted by reaction over V is equal to the integral of the source term f over V .

As in Sect. 2.1, we assume that there is a real number $\mu_0 > 0$ such that

$$\Lambda := \tilde{\mu} + \frac{1}{2} \nabla \cdot \beta = \mu - \frac{1}{2} \nabla \cdot \beta \geq \mu_0 \text{ a.e. in } \Omega.$$

Hence, using integration by parts, the bilinear form a is coercive on V ,

$$\forall v \in V, \quad a(v, v) = \|\kappa^{1/2} \nabla v\|_{[L^2(\Omega)]^d}^2 + \|\Lambda^{1/2} v\|_{L^2(\Omega)}^2.$$

Owing to the Lax–Milgram Lemma, (4.76) is therefore well-posed.

4.6.2 Discretization

We aim at approximating the exact solution u of (4.76) by a dG method using the discrete space

$$V_h := \mathbb{P}_d^k(\mathcal{T}_h),$$

where $\mathbb{P}_d^k(\mathcal{T}_h)$ is defined by (1.15) with polynomial degree $k \geq 1$ and \mathcal{T}_h belonging to an admissible mesh sequence. We keep Assumption 4.45 on the compatibility of the meshes with the partition P_Ω associated with the diffusion coefficient κ . Moreover, concerning the regularity of the exact solution, we assume that (cf. Assumption 4.48)

$$u \in V_* := V \cap H^2(P_\Omega),$$

and we set, as before, $V_{*h} = V_* + V_h$. It is also possible to analyze the dG approximation in the case of low-regularity exact solutions matching only Assumption 4.54.

The dG method considered herein combines the SWIP bilinear form of Sect. 4.5 to handle the diffusion term and the upwind dG method of Sect. 2.3 to handle the advection-reaction terms. Thus, we let, for all $(v, w_h) \in V_{*h} \times V_h$,

$$a_h^{\text{dar}}(v, w_h) = a_h^{\text{swip}}(v, w_h) + a_h^{\text{upw}}(v, w_h), \quad (4.78)$$

where (cf. (4.64))

$$\begin{aligned} a_h^{\text{swip}}(v, w_h) &:= \int_{\Omega} \kappa \nabla_h v \cdot \nabla_h w_h + \sum_{F \in \mathcal{F}_h} \eta \frac{\gamma_{\kappa, F}}{h_F} \int_F \llbracket v \rrbracket \llbracket w_h \rrbracket \\ &\quad - \sum_{F \in \mathcal{F}_h} \int_F (\llbracket \kappa \nabla_h v \rrbracket_{\omega \cdot \mathbf{n}_F} \llbracket w_h \rrbracket + \llbracket v \rrbracket \llbracket \kappa \nabla_h w_h \rrbracket_{\omega \cdot \mathbf{n}_F}), \end{aligned}$$

and (cf. (2.34))

$$\begin{aligned} a_h^{\text{upw}}(v, w_h) &= \int_{\Omega} [\tilde{\mu} v w_h + \nabla_h \cdot (\beta v) w_h] + \int_{\partial \Omega} (\beta \cdot \mathbf{n})^{\ominus} v w_h \\ &\quad - \sum_{F \in \mathcal{F}_h^i} \int_F (\beta \cdot \mathbf{n}_F) \llbracket v \rrbracket \llbracket w_h \rrbracket + \sum_{F \in \mathcal{F}_h^i} \int_F \gamma_{\beta, F} \llbracket v \rrbracket \llbracket w_h \rrbracket, \end{aligned}$$

or equivalently, after integrating by parts the advective derivative (cf. (2.35)),

$$\begin{aligned} a_h^{\text{upw}}(v, w_h) &= \int_{\Omega} [\tilde{\mu} v w_h - v(\beta \cdot \nabla_h w_h)] + \int_{\partial\Omega} (\beta \cdot \mathbf{n})^{\oplus} v w_h \\ &\quad + \sum_{F \in \mathcal{F}_h^i} \int_F (\beta \cdot \mathbf{n}_F) \llbracket v \rrbracket \llbracket w_h \rrbracket + \sum_{F \in \mathcal{F}_h^i} \int_F \gamma_{\beta, F} \llbracket v \rrbracket \llbracket w_h \rrbracket. \end{aligned}$$

In what follows, we set

$$\gamma_{\beta, F} := \frac{1}{2} |\beta \cdot \mathbf{n}_F|.$$

It is also possible to multiply $\gamma_{\beta, F}$ by a positive user-dependent parameter as in Sect. 2.3 (cf., e.g., (2.33)), but the present choice is needed for consistency reasons in Sect. 4.6.4 in the singular limit of locally vanishing diffusion; cf. Remark 4.69. We also observe that the penalty terms can be grouped to obtain

$$\sum_{F \in \mathcal{F}_h^i} \left(\eta \frac{\gamma_{\kappa, F}}{h_F} + \frac{1}{2} |\beta \cdot \mathbf{n}_F| \right) \int_F \llbracket v \rrbracket \llbracket w_h \rrbracket.$$

In the diffusion-dominated regime where $h_F |\beta \cdot \mathbf{n}_F| \lesssim \gamma_{\kappa, F}$, the amount of penalty introduced by the SWIP bilinear form suffices for discrete stability, and it is possible to drop upwinding for the advective terms (that is, to approximate the advective term using centered fluxes). The ratio $h_F |\beta \cdot \mathbf{n}_F| / \gamma_{\kappa, F}$ is termed a local Péclet number. In practice, local Péclet numbers are often large, generally because the diffusion coefficient is (locally) small, so that upwinding is necessary. In this situation, the exact solution features inner and outflow layers where it varies quite sharply, and practical meshes may not be fine enough to resolve these layers; we refer the reader, e.g., to Roos, Stynes, and Tobiska [274] for a general overview on singularly perturbed diffusion-advection-reaction problems and stabilized finite element approximations.

4.6.3 Error Estimates

To approximate the model problem (4.76), we consider the discrete problem:

$$\text{Find } u_h \in V_h \text{ s.t. } a_h^{\text{dar}}(u_h, v_h) = \int_{\Omega} f v_h \text{ for all } v_h \in V_h, \quad (4.79)$$

where a_h^{dar} is the discrete bilinear form defined by (4.78). The convergence analysis is performed by establishing discrete stability, consistency, and boundedness for a_h^{dar} . We begin with consistency.

Lemma 4.58 (Consistency). *Assume $u \in V_*$. Then, for all $w_h \in V_h$,*

$$a_h^{\text{dar}}(u, w_h) = \int_{\Omega} f w_h.$$

Proof. The proof of Lemma 4.49 yields, for all $w_h \in V_h$,

$$a_h^{\text{swip}}(u, w_h) = \sum_{T \in \mathcal{T}_h} \int_T \nabla \cdot (-\kappa \nabla u) w_h.$$

Moreover, adapting the proof of Lemma 2.27, we infer

$$a_h^{\text{upw}}(u, w_h) = \sum_{T \in \mathcal{T}_h} \int_T \nabla \cdot (\beta u) w_h + \int_{\Omega} \tilde{\mu} u w_h.$$

Summing up and observing that $\nabla \cdot (-\kappa \nabla u + \beta u) + \tilde{\mu} u = f$ in all $T \in \mathcal{T}_h$ yields the assertion. \square

4.6.3.1 Analysis Based on Discrete Coercivity

The convergence analysis is performed in the spirit of Theorem 2.31 by combining consistency (cf. Lemma 4.58) with discrete coercivity and boundedness on orthogonal subscales for the discrete bilinear form a_h^{dar} . We recall that in the context of the advection-reaction equation, we introduced in Sect. 2.1 the reference time τ_c and the reference velocity β_c such that

$$\tau_c := \{\max(\|\mu\|_{L^\infty(\Omega)}, L_\beta)\}^{-1}, \quad \beta_c := \|\beta\|_{[L^\infty(\Omega)]^d},$$

where L_β is the Lipschitz module of β (cf. (2.5)). We define on V_{*h} the norm

$$\|v\|_{\text{dab}} := \left(\|v\|_{\text{swip}}^2 + |v|_\beta^2 + \tau_c^{-1} \|v\|_{L^2(\Omega)}^2 \right)^{1/2}, \quad (4.80)$$

where the $\|\cdot\|_{\text{swip}}$ -norm is defined by (4.69) and (4.70) while the $|\cdot|_\beta$ -seminorm is defined as

$$|v|_\beta := \left(\int_{\partial\Omega} \frac{1}{2} |\beta \cdot \mathbf{n}| v^2 + \sum_{F \in \mathcal{F}_h^i} \int_F \frac{1}{2} |\beta \cdot \mathbf{n}_F| \|v\|^2 \right)^{1/2}.$$

The two rightmost terms in (4.80) form the stability norm (cf. (2.37)) considered in Sect. 2.3.2 for the advection-reaction equation.

Lemma 4.59 (Discrete coercivity). *For all $\eta > \underline{\eta}$ with $\underline{\eta}$ defined in Lemma 4.12, the discrete bilinear form a_h^{dar} defined by (4.78) is coercive on V_h , i.e.,*

$$\forall v_h \in V_h, \quad a_h^{\text{dar}}(v_h, v_h) \geq \min(1, \tau_c \mu_0, C_\eta) \|v_h\|_{\text{dab}}^2,$$

with C_η defined in Lemma 4.12.

Proof. Let $v_h \in V_h$. Lemma 4.51 yields

$$a_h^{\text{swip}}(v_h, v_h) \geq C_\eta \|v_h\|_{\text{swip}}^2.$$

Moreover, owing to Lemma 2.27,

$$a_h^{\text{upw}}(v_h, v_h) \geq \min(1, \tau_c \mu_0) \left(|v_h|_\beta^2 + \tau_c^{-1} \|v_h\|_{L^2(\Omega)}^2 \right).$$

Combining these lower bounds yields the assertion. \square

A straightforward consequence of the Lax–Milgram Lemma is that the discrete problem (4.79) is well-posed.

The last ingredient is boundedness on orthogonal subscales for the discrete bilinear form a_h^{dar} . To this purpose, we define on V_{*h} the norm

$$\|v\|_{\text{dab},*} := \left(\|v\|_{\text{dab}}^2 + \sum_{T \in \mathcal{T}_h} \beta_c \|v\|_{L^2(\partial T)}^2 + \sum_{T \in \mathcal{T}_h} h_T \|\kappa^{1/2} \nabla v \cdot \mathbf{n}_T\|_{L^2(\partial T)}^2 \right)^{1/2}.$$

Lemma 4.60 (Boundedness on orthogonal subscales). *There is C_{bnd} , independent of h and the data κ , β , and $\tilde{\mu}$, such that*

$$\forall (v, w_h) \in V_* \times V_h, \quad a_h^{\text{dar}}(v - \pi_h v, w_h) \leq C_{\text{bnd}} \|v - \pi_h v\|_{\text{dab},*} \|w_h\|_{\text{dab}},$$

where π_h denotes the L^2 -orthogonal projection onto V_h .

Proof. Combine Lemma 4.52 with Lemma 2.30. (The fact that the first argument in a_h^{dar} is L^2 -orthogonal to V_h is only needed to apply Lemma 2.30.) \square

Proceeding as in the proof of Theorem 2.31 leads to the following error estimate.

Theorem 4.61 (Error estimate). *Let $u \in V_*$ solve (4.76). Let u_h solve (4.79) with a_h^{dar} defined by (4.78) and penalty parameter as in Lemma 4.12. Then, there is C , independent of h and the data κ , β , and $\tilde{\mu}$, such that*

$$\|u - u_h\|_{\text{dab}} \leq C \max(1, \tau_c^{-1} \mu_0^{-1}, C_\eta^{-1}) \|u - \pi_h u\|_{\text{dab},*}. \quad (4.81)$$

A convergence rate can be inferred from (4.81) using Lemmata 1.58 and 1.59 if the exact solution is smooth enough. Namely, if $u \in H^{k+1}(\Omega)$, (4.81) yields

$$\|u - u_h\|_{\text{dab}} \leq C'_u \max(1, \tau_c^{-1} \mu_0^{-1}, C_\eta^{-1}) (\bar{\kappa}^{1/2} + \beta_c^{1/2} h^{1/2} + \tau_c^{-1/2} h) h^k, \quad (4.82)$$

with $\bar{\kappa} := \|\kappa\|_{L^\infty(\Omega)}$, $C'_u = C' \|u\|_{H^{k+1}(\Omega)}$, and C' independent of h and the data κ , β , and $\tilde{\mu}$. The estimate can be simplified by dropping the last term under the reasonable assumption that $h \leq \beta_c \tau_c$; cf. (2.41). Moreover, observing that $h\beta_c/\bar{\kappa}$ represents a Péclet number and recalling the definition (4.80) of the $\|\cdot\|_{\text{dab}}$ -norm, we conclude that in the advection-dominated regime, the convergence rate of $\|u - u_h\|_\beta + \tau_c^{-1/2} \|u - u_h\|_{L^2(\Omega)}$ is of order $h^{k+1/2}$ (as for the pure advection-reaction problem; cf. Sect. 2.3.2), while in the diffusion-dominated regime, the convergence rate of $\|u - u_h\|_{\text{swip}}$ is of order h^k (as for the purely diffusive problem; cf. Sect. 4.5.3).

4.6.3.2 Analysis Based on Discrete Inf-Sup Condition

As shown in [133, 150], the above convergence analysis can be improved by including a bound on the advective derivative of the error. To this purpose, we need to tighten the discrete stability norm. Indeed, using the $\|\cdot\|_{\text{swip}}$ -norm contribution to the $\|\cdot\|_{\text{dab}}$ -norm to bound the advective derivative leads to an error bound

that scales unfavorably with the Péclet number. Instead, we define on V_{*h} the norm

$$\|v\|_{\text{da}\sharp} := \left(\|v\|_{\text{dab}}^2 + \sum_{T \in \mathcal{T}_h} \beta_c^{-1} h_T \|\beta \cdot \nabla v\|_{L^2(T)}^2 \right)^{1/2}.$$

As in Sect. 2.3.3, asserting discrete stability in the $\|\cdot\|_{\text{da}\sharp}$ -norm requires proving a discrete inf-sup condition.

Lemma 4.62 (Discrete inf-sup stability). *There is C_{sta} , independent of h and the data κ , β , and $\tilde{\mu}$, such that*

$$\forall v_h \in V_h, \quad C_{\text{sta}} \min(1, \tau_c \mu_0, C_\eta) \|v_h\|_{\text{da}\sharp} \leq \sup_{w_h \in V_h \setminus \{0\}} \frac{a_h^{\text{dar}}(v_h, w_h)}{\|w_h\|_{\text{da}\sharp}}.$$

Proof. The proof is similar to that of Lemma 2.35. Let $v_h \in V_h$ and set $\mathbb{S} = \sup_{w_h \in V_h \setminus \{0\}} \frac{a_h^{\text{dar}}(v_h, w_h)}{\|w_h\|_{\text{da}\sharp}}$. Lemma 4.59 implies that

$$\min(1, \tau_c \mu_0, C_\eta) \|v_h\|_{\text{dab}}^2 \leq a_h^{\text{dar}}(v_h, v_h) \leq \mathbb{S} \|v_h\|_{\text{da}\sharp}.$$

To bound the contribution of the advective derivative in the expression for $\|v_h\|_{\text{da}\sharp}$, we consider the function $w_h \in V_h$ such that, for all $T \in \mathcal{T}_h$, $w_h|_T = \beta_c^{-1} h_T \langle \beta \rangle_T \cdot \nabla v_h$ where $\langle \beta \rangle_T$ denotes the mean value of β over T . To alleviate the notation, we abbreviate as $a \lesssim b$ the inequality $a \leq Cb$ with positive C independent of h and the data κ , β , and $\tilde{\mu}$.

(i) Let us bound $\|w_h\|_{\text{da}\sharp}$ by $\|v_h\|_{\text{da}\sharp}$. As in the proof of Lemma 2.35, we obtain

$$|w_h|_\beta^2 + \tau_c^{-1} \|w_h\|_{L^2(\Omega)}^2 + \sum_{T \in \mathcal{T}_h} \beta_c^{-1} h_T \|\beta \cdot \nabla w_h\|_{L^2(T)}^2 \lesssim \|v_h\|_{\text{da}\sharp}^2.$$

Moreover, owing to the inverse inequality (1.36) and the fact that $\kappa|_T$ and $\langle \beta \rangle_T$ are constant in any mesh element $T \in \mathcal{T}_h$,

$$\begin{aligned} \|\kappa^{1/2} \nabla_h w_h\|_{[L^2(\Omega)]^d}^2 &= \sum_{T \in \mathcal{T}_h} \kappa|_T \beta_c^{-2} h_T^2 \|\nabla(\langle \beta \rangle_T \cdot \nabla v_h)\|_{L^2(T)}^2 \\ &\lesssim \sum_{T \in \mathcal{T}_h} \kappa|_T \|\nabla v_h\|_{[L^2(T)]^d}^2 = \|\kappa^{1/2} \nabla_h v_h\|_{[L^2(\Omega)]^d}^2. \end{aligned}$$

In addition, for all $F \in \mathcal{F}_h^i$ with $F = \partial T_1 \cap \partial T_2$,

$$\begin{aligned} \frac{\gamma_{\kappa, F}}{h_F} \|\llbracket w_h \rrbracket\|_{L^2(F)}^2 &\leq 2 \frac{\gamma_{\kappa, F}}{h_F} \beta_c^{-2} \sum_{i \in \{1, 2\}} h_{T_i}^2 \|\langle \beta \rangle_{T_i} \cdot (\nabla v_h)|_{T_i}\|_{L^2(F)}^2 \\ &\lesssim \sum_{i \in \{1, 2\}} \kappa|_{T_i} \|\nabla v_h\|_{[L^2(T_i)]^d}^2, \end{aligned}$$

where we have used the discrete trace inequality (1.37), the mesh regularity, and the bound (4.65) on $\gamma_{\kappa, F}$. Hence,

$$|w_h|_{J, \kappa} \lesssim \|\kappa^{1/2} \nabla_h v_h\|_{[L^2(\Omega)]^d},$$

and collecting the above bounds yields $\|w_h\|_{\text{da}\sharp} \lesssim \|v_h\|_{\text{da}\sharp}$.

(ii) Proceeding as in step (ii) of the proof of Lemma 2.35, we observe that

$$\begin{aligned} \sum_{T \in \mathcal{T}_h} \beta_c^{-1} h_T \|\beta \cdot \nabla v_h\|_{L^2(T)}^2 &= a_h^{\text{dar}}(v_h, w_h) - a_h^{\text{swip}}(v_h, w_h) - \int_{\Omega} \mu v_h w_h \\ &\quad + \sum_{T \in \mathcal{T}_h} \beta_c^{-1} h_T \int_T (\beta \cdot \nabla v_h) (\beta - \langle \beta \rangle_T) \cdot \nabla v_h \\ &\quad - \int_{\partial\Omega} (\beta \cdot \mathbf{n})^{\ominus} v_h w_h + \sum_{F \in \mathcal{F}_h^i} \int_F (\beta \cdot \mathbf{n}_F) \llbracket v_h \rrbracket \llbracket w_h \rrbracket \\ &\quad - \sum_{F \in \mathcal{F}_h^i} \int_F \frac{1}{2} |\beta \cdot \mathbf{n}_F| \llbracket v_h \rrbracket \llbracket w_h \rrbracket = \mathfrak{T}_1 + \dots + \mathfrak{T}_7. \end{aligned}$$

Clearly, $|\mathfrak{T}_1| \leq \mathbb{S} \|w_h\|_{\text{da}\sharp} \lesssim \mathbb{S} \|v_h\|_{\text{da}\sharp}$ and

$$|\mathfrak{T}_2| = |a_h^{\text{swip}}(v_h, w_h)| \lesssim \|v_h\|_{\text{dab}} \|w_h\|_{\text{dab}} \lesssim \|v_h\|_{\text{dab}} \|v_h\|_{\text{da}\sharp}.$$

Finally, the terms $\mathfrak{T}_3, \dots, \mathfrak{T}_7$ are those already bounded in the proof of Lemma 2.35. As a result,

$$\sum_{T \in \mathcal{T}_h} \beta_c^{-1} h_T \|\beta \cdot \nabla v_h\|_{L^2(T)}^2 \lesssim \mathbb{S} \|v_h\|_{\text{da}\sharp} + \|v_h\|_{\text{dab}} \|v_h\|_{\text{da}\sharp} + \|v_h\|_{\text{dab}}^2.$$

We conclude as in the proof of Lemma 2.35. \square

To formulate a boundedness result, we define on V_{*h} the norm

$$\begin{aligned} \|v\|_{\text{da}\sharp,*} &:= \left(\|v\|_{\text{da}\sharp}^2 + \sum_{T \in \mathcal{T}_h} \beta_c \left(h_T^{-1} \|v\|_{L^2(T)}^2 + \|v\|_{L^2(\partial T)}^2 \right) \right. \\ &\quad \left. + \sum_{T \in \mathcal{T}_h} h_T \|\kappa^{1/2} \nabla v \cdot \mathbf{n}_T\|_{L^2(\partial T)}^2 \right)^{1/2}. \end{aligned}$$

Lemma 4.63 (Boundedness). *There is C_{bnd} , independent of h and the data κ , β , and $\tilde{\mu}$, such that*

$$\forall (v, w_h) \in V_{*h} \times V_h, \quad a_h^{\text{dar}}(v, w_h) \leq C_{\text{bnd}} \|v\|_{\text{da}\sharp,*} \|w_h\|_{\text{da}\sharp}.$$

Proof. Combine Lemma 4.52 with Lemma 2.36. \square

A straightforward consequence of Theorem 1.35 is the following error estimate.

Theorem 4.64 (Error estimate). *Under the hypotheses of Theorem 4.61, there is C , independent of h and the data κ , β , and $\tilde{\mu}$, such that*

$$\|u - u_h\|_{\text{da}\sharp} \leq C \max(1, \tau_c^{-1} \mu_0^{-1}, C_{\eta}^{-1}) \inf_{v_h \in V_h} \|u - v_h\|_{\text{da}\sharp,*}. \quad (4.83)$$

Finally, a convergence rate can be inferred from (4.83) using Lemmata 1.58 and 1.59 if $u \in H^{k+1}(\Omega)$ since (4.83) yields an error estimate with the same upper bound as in (4.82), namely

$$\|u - u_h\|_{\text{da}\sharp} \leq C'_u \max(1, \tau_c^{-1} \mu_0^{-1}, C_\eta^{-1}) (\bar{\kappa}^{-1/2} + \beta_c^{1/2} h^{1/2} + \tau_c^{-1/2} h) h^k.$$

Thus, in the advection-dominated regime, the convergence rate of $|u - u_h|_\beta + \tau_c^{-1/2} \|u - u_h\|_{L^2(\Omega)} + (\sum_{T \in \mathcal{T}_h} \beta_c^{-1} h_T \|\beta \cdot \nabla v\|_{L^2(T)}^2)^{1/2}$ is of order $h^{k+1/2}$ (as for the pure advection-reaction problem; cf. Sect. 2.3.3), while in the diffusion-dominated regime, the convergence rate of $\|u - u_h\|_{\text{swip}}$ is of order h^k (as for the purely diffusive problem; cf. Sect. 4.5.3).

Remark 4.65 (Harmonic means in the penalty term). The bound (4.65) plays an important role in the proof of Lemma 4.62 since it allows one to bound the jump seminorm $|w_h|_{J, \kappa}$. We observe that this bound results from the fact that the harmonic mean of the diffusion coefficient is used to penalize jumps across interfaces in the SWIP bilinear form.

Remark 4.66 (Numerical fluxes). A local formulation using numerical fluxes can be derived for the discrete problem (4.79) by combining the results of Sect. 4.5.5 for the diffusion terms and those of Sect. 2.3.4 for the advection-reaction terms. Specifically, letting $T \in \mathcal{T}_h$ and $\xi \in \mathbb{P}_d^k(T)$, we infer (compare with (4.77))

$$\int_T (\kappa G_{h, \kappa}^l(u_h) - u_h \beta) \cdot \nabla \xi + \int_T \tilde{\mu} u_h \xi + \sum_{F \in \mathcal{F}_T} \epsilon_{T, F} \int_F \phi_F(u_h) \xi = \int_T f \xi,$$

with $l \in \{k-1, k\}$, $\epsilon_{T, F} = n_T \cdot n_F$, and the numerical flux $\phi_F(u_h)$ defined as

$$\phi_F(u_h) := \begin{cases} (-\{\kappa \nabla_h u_h\}_\omega + \beta \{\{u_h\}\}) \cdot n_F + (\eta \frac{\gamma_{\kappa, F}}{h_F} + \frac{1}{2} |\beta \cdot n_F|) \|u_h\| & \text{if } F \in \mathcal{F}_h^i, \\ -\kappa \nabla_h u_h \cdot n + (\beta \cdot n)^\oplus u_h + \eta \frac{\gamma_{\kappa, F}}{h_F} u_h & \text{if } F \in \mathcal{F}_h^b. \end{cases}$$

Remark 4.67 (Anisotropic diffusion). In the case of anisotropic diffusion, the SWIP bilinear form is modified as discussed in Sect. 4.5.6. The convergence analysis based on discrete coercivity can be extended to this case. However, it is not clear how to extend the proof of Lemma 4.62 since the bound on $\|\kappa^{1/2} \nabla_h w_h\|_{[L^2(\Omega)]^d}$ uses the assumption that κ is scalar-valued; see [150] for further discussion.

4.6.4 Locally Vanishing Diffusion

In this section, we are interested in the case where κ only takes nonnegative values in the domain Ω , a typical example being that κ vanishes in some parts of Ω . In the anisotropic case, a more complex situation is that where κ only takes symmetric semidefinite values, for instance because different eigenvalues of κ vanish in different parts of Ω .

4.6.4.1 The Continuous Setting

As before, we keep Assumption 4.43 so as to localize the jumps of κ , and we consider the resulting partition P_Ω . We say that I is a *partition interface* if:

- (a) I has positive $(d - 1)$ -dimensional Hausdorff measure.
- (b) I is part of a hyperplane, say H_I .
- (c) There are distinct Ω_i and Ω_j belonging to P_Ω such that $I = H_I \cap \partial\Omega_i \cap \partial\Omega_j$.

Partition interfaces are collected into the set I_Ω and points in Ω belonging to partition interfaces are collected into the set \mathcal{I}_Ω . Of particular interest are those partition interfaces for which the normal component of the diffusion tensor becomes singular on one of its sides, say Ω_j . Specifically, we set

$$I_{0,\Omega} := \{I \in I_\Omega \mid \mathbf{n}_I^t(\kappa|_{\Omega_i})\mathbf{n}_I > \mathbf{n}_I^t(\kappa|_{\Omega_j})\mathbf{n}_I = 0\},$$

where \mathbf{n}_I denotes a unit normal vector to I , and without loss of generality we assume that \mathbf{n}_I points from Ω_i toward Ω_j . On a (partition) interface $I \in I_{0,\Omega}$, we loosely say that the subdomain Ω_i is the *diffusive side* and the subdomain Ω_j the *nondiffusive side*. Points in Ω belonging to (partition) interfaces in $I_{0,\Omega}$ are collected into the set $\mathcal{I}_{0,\Omega}$. It is important to identify those points in $\mathcal{I}_{0,\Omega}$ where the advective field flows from the diffusive side to the nondiffusive side and to distinguish them from the remaining points, namely

$$\begin{aligned} \mathcal{I}_{0,\Omega}^+ &:= \{x \in \mathcal{I}_{0,\Omega} \mid (\beta \cdot \mathbf{n}_I)(x) > 0\}, \\ \mathcal{I}_{0,\Omega}^- &:= \{x \in \mathcal{I}_{0,\Omega} \mid (\beta \cdot \mathbf{n}_I)(x) < 0\}, \end{aligned}$$

and we assume that $(\beta \cdot \mathbf{n}_I)(x) \neq 0$ for a.e. $x \in \mathcal{I}_{0,\Omega}$. Following Di Pietro, Ern, and Guermond [133], we consider the following diffusion-advection-reaction problem with locally vanishing diffusion

$$\nabla \cdot (-\kappa \nabla u + \beta u) + \tilde{\mu}u = f \quad \text{in } \Omega \setminus \mathcal{I}_{0,\Omega}, \quad (4.84a)$$

$$u = 0 \quad \text{on } \partial\Omega_{\kappa,\beta}, \quad (4.84b)$$

where

$$\partial\Omega_{\kappa,\beta} := \{x \in \partial\Omega \mid \mathbf{n}^t \kappa \mathbf{n} > 0 \text{ or } \beta \cdot \mathbf{n} < 0\},$$

and supplemented with the following conditions on $\mathcal{I}_{0,\Omega}$:

$$[-\kappa \nabla u + \beta u] \cdot \mathbf{n}_I = 0 \quad \text{on } \mathcal{I}_{0,\Omega}, \quad (4.85a)$$

$$[u] = 0 \quad \text{on } \mathcal{I}_{0,\Omega}^+. \quad (4.85b)$$

We observe that (4.84b) enforces a homogeneous Dirichlet condition if $\mathbf{n}^t \kappa \mathbf{n} > 0$ (as for pure diffusion problems) or if $\beta \cdot \mathbf{n} < 0$ (as for advection-reaction problems). Moreover, (4.85a) enforces the continuity of the normal component of the diffusive-advective flux on the whole partition interface $\mathcal{I}_{0,\Omega}$, whereas (4.85b) enforces the continuity of the exact solution only on $\mathcal{I}_{0,\Omega}^+$, that is, where the

advection field flows from the diffusive side toward the nondiffusive side, while the exact solution can jump across $\mathcal{I}_{0,\Omega}^-$. We also notice that combining (4.85a) and (4.85b) yields $[\![\kappa \nabla u]\!] \cdot \mathbf{n}_I = 0$ across $\mathcal{I}_{0,\Omega}^+$ (recall that β is smooth so that its normal component is continuous across partition interfaces). Since $\kappa|_{\Omega_j} \cdot \mathbf{n}_I = 0$ on the nondiffusive side, this yields the homogeneous Neumann condition $(\kappa \nabla u)|_{\Omega_i} \cdot \mathbf{n}_I = 0$ on the diffusive side.

The mathematical analysis of the model problem (4.84) with conditions (4.85) can be found in [133]. We only give here a brief motivation for condition (4.85b). In one space dimension, these conditions were derived by Gastaldi and Quarteroni [165], where it is proven that the solution u_ϵ of the following regularized problem with suitable boundary conditions:

$$(-\kappa u'_\epsilon + \beta u_\epsilon)' + \tilde{\mu} u_\epsilon - \epsilon u''_\epsilon = f, \quad (4.86)$$

converges in $L^2(\Omega)$, as $\epsilon \rightarrow 0$, to the so-called viscosity solution of (4.84a) which satisfies conditions (4.85). As an example, let $\Omega = (0, 1)$ be partitioned into $\Omega_1 = (0, \frac{1}{3})$, $\Omega_2 = (\frac{1}{3}, \frac{2}{3})$, and $\Omega_3 = (\frac{2}{3}, 1)$ and set $f = 0$, $\mu = 0$, $\beta = 1$, $\kappa|_{\Omega_1 \cup \Omega_3} = 1$, and $\kappa|_{\Omega_2} = 0$. Then, $\mathcal{I}_{0,\Omega} = \{\frac{1}{3}, \frac{2}{3}\}$ with $\mathcal{I}_{0,\Omega}^+ = \{\frac{1}{3}\}$ and $\mathcal{I}_{0,\Omega}^- = \{\frac{2}{3}\}$. The viscosity solution to (4.86) with the boundary conditions $u(0) = 1$ and $u(1) = 0$ is (cf. Fig. 4.7)

$$u|_{\Omega_1} = u|_{\Omega_2} = 1, \quad u|_{\Omega_3} = 1 - e^{(x-1)}.$$

This solution satisfies (4.85).

4.6.4.2 Discretization

We set $V_h = \mathbb{P}_d^k(\mathcal{T}_h)$ with $k \geq 1$ and \mathcal{T}_h belonging to an admissible mesh sequence satisfying Assumption 4.45. In addition, we assume that each (mesh) interface

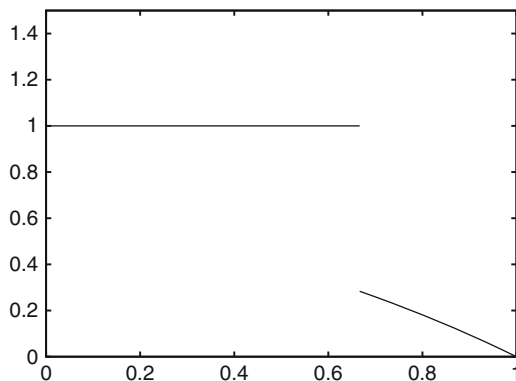


Fig. 4.7: Exact solution with vanishing diffusion

$F \in \mathcal{F}_h^i$ such that $F \cap \mathcal{I}_{0,\Omega}$ has positive $(d-1)$ -dimensional Hausdorff measure is either a subset of $\mathcal{I}_{0,\Omega}^-$ or of $\mathcal{I}_{0,\Omega}^+$. We define \mathcal{F}_h^{i*} as the set of (mesh) interfaces such that $F \cap \mathcal{I}_{0,\Omega}$ is a subset of $\mathcal{I}_{0,\Omega}^-$. Without loss of generality, we assume that the normal \mathbf{n}_F to each $F \in \mathcal{F}_h^{i*}$ points from the diffusive side, say Ω_1 , to the nondiffusive side, say Ω_2 . As a result, the weights at F are such that $\omega_{T_1,F} = 0$ and $\omega_{T_2,F} = 1$. We also assume that each boundary face $F \in \mathcal{F}_h^b$ is either a subset of $\partial\Omega_{\kappa,\beta}$ or of $\partial\Omega \setminus \partial\Omega_{\kappa,\beta}$, and we define \mathcal{F}_h^{b*} as the set of boundary faces that are a subset of $\partial\Omega \setminus \partial\Omega_{\kappa,\beta}$. With these definitions, we obtain

$$\llbracket u \rrbracket = 0 \quad \forall F \in \mathcal{F}_h \setminus (\mathcal{F}_h^{i*} \cup \mathcal{F}_h^{b*}). \quad (4.87)$$

The key property is that the discrete bilinear form a_h^{dar} defined by (4.78) remains consistent even in the singular limit of vanishing diffusion.

Lemma 4.68 (Consistency). *For all $v_h \in V_h$,*

$$a_h^{\text{dar}}(u, v_h) = \int_{\Omega} f v_h.$$

Proof. Let $v_h \in V_h$. Consider first the contribution of the SWIP bilinear form. Owing to (4.66) and (4.87),

$$\begin{aligned} a_h^{\text{swip}}(u, v_h) &= - \sum_{T \in \mathcal{T}_h} \int_T \nabla \cdot (\kappa \nabla u) v_h + \sum_{F \in \mathcal{F}_h} \eta \frac{\gamma_{\kappa,F}}{h_F} \int_F \llbracket u \rrbracket \llbracket v_h \rrbracket \\ &\quad + \sum_{F \in \mathcal{F}_h^i} \int_F \llbracket \kappa \nabla_h u \rrbracket \cdot \mathbf{n}_F \{v_h\}_{\bar{\omega}} - \sum_{F \in \mathcal{F}_h} \int_F \llbracket u \rrbracket \{ \kappa \nabla_h v_h \}_{\omega} \cdot \mathbf{n}_F \\ &= - \sum_{T \in \mathcal{T}_h} \int_T \nabla \cdot (\kappa \nabla u) v_h + \sum_{F \in \mathcal{F}_h^{i*} \cup \mathcal{F}_h^{b*}} \eta \frac{\gamma_{\kappa,F}}{h_F} \int_F \llbracket u \rrbracket \llbracket v_h \rrbracket \\ &\quad + \sum_{F \in \mathcal{F}_h^{i*}} \int_F \llbracket \kappa \nabla_h u \rrbracket \cdot \mathbf{n}_F \{v_h\}_{\bar{\omega}} - \sum_{F \in \mathcal{F}_h^{i*} \cup \mathcal{F}_h^{b*}} \int_F \llbracket u \rrbracket \{ \kappa \nabla_h v_h \}_{\omega} \cdot \mathbf{n}_F, \end{aligned}$$

where we have used the fact that $\llbracket \kappa \nabla u \rrbracket \cdot \mathbf{n}_F = 0$ on all $F \in \mathcal{F}_h^i \setminus \mathcal{F}_h^{i*}$ owing to (4.85). Moreover, for all $F \in \mathcal{F}_h^{i*}$, $\gamma_{\kappa,F} = 0$ and $\{ \kappa \nabla_h v_h \}_{\omega} \cdot \mathbf{n}_F = 0$ owing to the definition of the penalty parameter and the weights. Similarly, owing to the boundary condition (4.84b), $\mathbf{n}^t \kappa \mathbf{n} = 0$ for all $F \in \mathcal{F}_h^{b*}$. As a result,

$$a_h^{\text{swip}}(u, v_h) = - \sum_{T \in \mathcal{T}_h} \int_T \nabla \cdot (\kappa \nabla u) v_h + \sum_{F \in \mathcal{F}_h^{i*}} \int_F \llbracket \kappa \nabla_h u \rrbracket \cdot \mathbf{n}_F v_h|_{\Omega_1},$$

where we have used the fact that $\{v_h\}_{\bar{\omega}} = v_h|_{\Omega_1}$ since $\omega_{T_1,F} = 0$ and $\omega_{T_2,F} = 1$.

Consider now the contribution of the upwind bilinear form, namely

$$\begin{aligned} a_h^{\text{upw}}(u, v_h) &= \sum_{T \in \mathcal{T}_h} \int_T [\tilde{\mu} u v_h + \nabla \cdot (\beta u) v_h] + \int_{\partial \Omega} (\beta \cdot \mathbf{n})^\ominus u v_h \\ &\quad - \sum_{F \in \mathcal{F}_h^i} \int_F (\beta \cdot \mathbf{n}_F) \llbracket u \rrbracket \{v_h\} + \sum_{F \in \mathcal{F}_h^i} \int_F \frac{1}{2} |\beta \cdot \mathbf{n}_F| \llbracket u \rrbracket \{v_h\} \\ &= \sum_{T \in \mathcal{T}_h} \int_T [\tilde{\mu} u v_h + \nabla \cdot (\beta u) v_h] - \sum_{F \in \mathcal{F}_h^{i*}} \int_F (\beta \cdot \mathbf{n}_F) \llbracket u \rrbracket v_h|_{\Omega_1}, \end{aligned}$$

where we have used (4.87), $(\beta \cdot \mathbf{n})^\ominus = 0$ on all $F \in \mathcal{F}_h^{b*}$, and that the upwind side is the nondiffusive side on all $F \in \mathcal{F}_h^{i*}$ so that

$$\begin{aligned} -(\beta \cdot \mathbf{n}_F) \llbracket u \rrbracket \{v_h\} + \frac{1}{2} |\beta \cdot \mathbf{n}_F| \llbracket u \rrbracket \{v_h\} &= -(\beta \cdot \mathbf{n}_F) \llbracket u \rrbracket \left(\{v_h\} + \frac{1}{2} \llbracket v_h \rrbracket \right) \\ &= -(\beta \cdot \mathbf{n}_F) \llbracket u \rrbracket v_h|_{\Omega_1}. \end{aligned}$$

Summing up yields

$$\begin{aligned} a_h^{\text{dar}}(u, v_h) &= \sum_{T \in \mathcal{T}_h} \int_T (\nabla \cdot (-\kappa \nabla u + \beta u) + \tilde{\mu} u) v_h + \sum_{F \in \mathcal{F}_h^{i*}} \int_F \llbracket \kappa \nabla u - \beta u \rrbracket \cdot \mathbf{n}_F v_h|_{\Omega_1} \\ &= \sum_{T \in \mathcal{T}_h} \int_T (\nabla \cdot (-\kappa \nabla u + \beta u) + \tilde{\mu} u) v_h, \end{aligned}$$

owing to (4.85a). The assertion follows. \square

Remark 4.69 (Amount of upwinding). The choice $\gamma_{\beta, F} = \frac{1}{2} |\beta \cdot \mathbf{n}_F|$, corresponding to the usual amount of upwinding, is instrumental in the above proof so as to combine the two terms multiplying $v_h|_{\Omega_1}$ and recover the jump of the total diffusive-advective flux.

The rest of the convergence analysis proceeds as in Sect. 4.6.3 yielding the error estimates (4.81) and (4.83). We observe that the approximate solution exhibits, like the exact solution, a finite jump across $\mathcal{I}_{0, \Omega}^-$. The approximation error on this jump is controlled via the $|\cdot|_\beta$ -seminorm present in the error estimates.

To illustrate with a two-dimensional example, we consider $\Omega = (0, 1)^2$ partitioned into the two subdomains depicted in the left panel of Fig. 4.8. The subdomain Ω_1 is a trapezoidal inclusion. The diffusion is anisotropic and such that

$$\kappa|_{\Omega_1} = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}, \quad \kappa|_{\Omega_2} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The advection field is horizontal and uniform with $\beta = (-5, 0)$, and the reaction coefficient is uniform with $\tilde{\mu} = 1$. The partition interface $\mathcal{I}_{0, \Omega}$ consists of the two vertical sides of Ω_1 , with $\mathcal{I}_{0, \Omega}^+$ equal to the left side and $\mathcal{I}_{0, \Omega}^-$ to the right side. The approximate solution obtained with the above dG method and polynomial degree $k = 1$ is shown in the right panel of Fig. 4.8, showing that the expected behavior of the exact solution is captured accurately. In particular, the jump across $\mathcal{I}_{0, \Omega}^-$ is clearly visible.

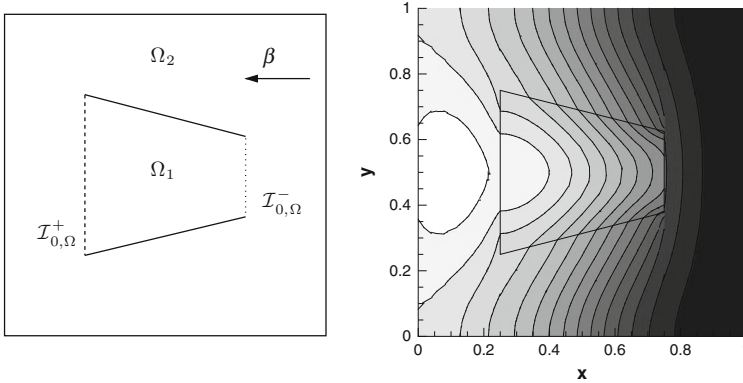


Fig. 4.8: Two-dimensional example of heterogeneous and anisotropic diffusion-advection-reaction problem: problem setting (*left*) and approximate dG solution (*right*)

4.7 An Unsteady Example: The Heat Equation

To illustrate the approximation of unsteady scalar PDEs with diffusion, we consider the heat equation, which we approximate in space using the SIP scheme of Sect. 4.2 and in time using (implicit) A-stable finite difference schemes, e.g., the backward Euler and BDF2 schemes. Implicit time-marching is usually preferred for parabolic problems because explicit schemes lead to the stringent parabolic CFL condition $\delta t \leq Ch^2$ where δt is the time step and h the meshsize.

4.7.1 The Continuous Setting

For given finite time $t_F > 0$, source term f , and initial datum u_0 , we consider the unsteady version of the Poisson problem (4.1), namely,

$$\partial_t u - \Delta u = f \quad \text{in } \Omega \times (0, t_F), \quad (4.88a)$$

$$u = 0 \quad \text{on } \partial\Omega \times (0, t_F), \quad (4.88b)$$

$$u(\cdot, t = 0) = u_0 \quad \text{in } \Omega. \quad (4.88c)$$

Problem (4.88) is termed the *heat equation*.

We recall (cf. Sect. 3.1.1) that, for a function ψ defined on the space-time cylinder $\Omega \times (0, t_F)$, we consider ψ as a function of the time variable with values in a Hilbert space, say V , spanned by functions of the space variable, in such a way that

$$\psi : (0, t_F) \ni t \mapsto \psi(t) \equiv \psi(\cdot, t) \in V.$$

We also recall that, for an integer $l \geq 0$, $C^l(V)$ denotes the space of V -valued functions that are l times continuously differentiable in the interval $[0, t_F]$.

We take the source term f in $C^0(L^2(\Omega))$. Moreover, we are interested in smooth solutions such that

$$u \in C^0(H_0^1(\Omega)) \cap C^1(L^2(\Omega)).$$

This implies, in particular, that the initial datum u_0 is in the energy space $H_0^1(\Omega)$. In addition, since $u \in C^0(H_0^1(\Omega)) \cap C^1(L^2(\Omega))$, $d_t u \in L^2(\Omega)$ for all $t \in (0, t_F)$, so that we consider the following weak formulation of (4.88): For all $t \in [0, t_F]$,

$$(d_t u, v)_{L^2(\Omega)} + a(u, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega), \quad (4.89)$$

where, as in the steady case, $a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v$.

We now establish the basic stability result for (4.89).

Lemma 4.70 (Stability). *Let $u \in C^0(H_0^1(\Omega)) \cap C^1(L^2(\Omega))$ solve (4.89). Then, for all $t \in [0, t_F]$,*

$$\|u(t)\|_{L^2(\Omega)}^2 + \int_0^t \|\nabla u(s)\|_{[L^2(\Omega)]^d}^2 \, ds \leq \|u_0\|_{L^2(\Omega)}^2 + C_{\Omega}^2 \int_0^t \|f(s)\|_{L^2(\Omega)}^2 \, ds,$$

where C_{Ω} results from the Poincaré inequality (4.4).

Proof. For a fixed $t \in (0, t_F)$, selecting $u(t)$ as a test function in (4.89) and using the Cauchy–Schwarz inequality followed by the Poincaré inequality (4.4), we infer

$$\begin{aligned} \frac{1}{2} d_t \|u(t)\|_{L^2(\Omega)}^2 + \|\nabla u(t)\|_{[L^2(\Omega)]^d}^2 &= (f(t), u(t))_{L^2(\Omega)} \\ &\leq C_{\Omega} \|f(t)\|_{L^2(\Omega)} \|\nabla u(t)\|_{[L^2(\Omega)]^d} \\ &\leq \frac{C_{\Omega}^2}{2} \|f(t)\|_{L^2(\Omega)}^2 + \frac{1}{2} \|\nabla u(t)\|_{[L^2(\Omega)]^d}^2. \end{aligned}$$

Rearranging terms and integrating in time yields the assertion. \square

4.7.2 Discretization

As in Chap. 3, we focus on the method of lines in which the time evolution problem (4.89) is first semidiscretized in space yielding a system of coupled ODEs which is then discretized in time. Specifically, we consider space semidiscretization by the SIP dG method of Sect. 4.2 together with a *backward* (also called *implicit*) *Euler scheme* for time discretization. The BDF2 scheme to discretize in time is addressed in Sect. 4.7.4.

4.7.2.1 Space Semidiscretization

Let $V_h = \mathbb{P}_d^k(\mathcal{T}_h)$ with polynomial degree $k \geq 1$ and \mathcal{T}_h belonging to an admissible mesh sequence. The spaces V_* and V_{*h} are defined in Assumption 4.4.

The discrete problem is formulated as follows: For all $t \in [0, t_F]$,

$$(d_t u_h, v_h)_{L^2(\Omega)} + a_h^{\text{sip}}(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h,$$

with bilinear form a_h^{sip} defined by (4.10). We introduce the discrete differential operator $A_h^{\text{sip}} : V_{*h} \rightarrow V_h$ such that, for all $(v, w_h) \in V_{*h} \times V_h$,

$$(A_h^{\text{sip}} v, w_h)_{L^2(\Omega)} = a_h^{\text{sip}}(v, w_h).$$

The discrete operator A_h^{sip} can be used to formulate the space semidiscrete problem in the following form: For all $t \in [0, t_F]$,

$$d_t u_h(t) + A_h^{\text{sip}} u_h(t) = f_h(t), \quad (4.90)$$

with initial condition $u_h(0) = \pi_h u_0$ and source term

$$f_h(t) = \pi_h f(t) \quad \forall t \in [0, t_F],$$

where π_h denotes, as usual, the L^2 -orthogonal projection onto V_h . Choosing a basis for V_h , the space semidiscrete evolution problem (4.90) can be transformed into a system of coupled ODEs for the time-dependent components of $u_h(t)$ on the selected basis. Written in component form, (4.90) leads to the appearance of the mass matrix in front of the time derivative. In the context of dG methods, the mass matrix is block-diagonal; cf. Sect. A.1.2.

We now state the important properties of the discrete operator A_h^{sip} . These properties result from the consistency, discrete coercivity, and boundedness of the SIP bilinear form.

Lemma 4.71 (Properties of A_h^{sip}). *The discrete operator A_h^{sip} satisfies the following properties:*

- (i) Consistency: For the exact solution u , assuming $u \in C^0(V_*)$,

$$\pi_h d_t u(t) + A_h^{\text{sip}} u(t) = f_h(t) \quad \forall t \in [0, t_F].$$

- (ii) Discrete coercivity: For all $v_h \in V_h$,

$$(A_h^{\text{sip}} v_h, v_h)_{L^2(\Omega)} \geq C_{\text{sta}} \|v_h\|_{\text{sip}}^2.$$

- (iii) Boundedness: For all $(v, w_h) \in V_{*h} \times V_h$,

$$(A_h^{\text{sip}} v, w_h)_{L^2(\Omega)} \leq C_{\text{bnd}} \|v\|_{\text{sip},*} \|w_h\|_{\text{sip}}.$$

Here, C_{sta} and C_{bnd} are independent of h and δt , the $\|\cdot\|_{\text{sip}}$ -norm is defined by (4.17), and the $\|\cdot\|_{\text{sip},*}$ -norm by (4.22).

4.7.2.2 Time Discretization

To discretize in time the space semidiscrete problem (4.90), we introduce a partition of $(0, t_F)$ into N intervals of length δt (the time step) so that $N\delta t = t_F$; more generally, a variable time step can be considered. For $n \in \{0, \dots, N\}$, a superscript n indicates the value of a function at the discrete time $t^n := n\delta t$.

For a function $v \in C^1(V)$ with some function space V of the space variable, we introduce the backward Euler operator $\delta_t^{(1)}$ such that, for all $n \in \{1, \dots, N\}$,

$$\delta_t^{(1)} v^n := \frac{v^n - v^{n-1}}{\delta t} \in V, \quad (4.91)$$

thereby providing a first-order finite difference approximation of the time derivative $d_t v^n$. The discrete solution is obtained from the backward Euler scheme,

$$\delta_t^{(1)} u_h^{n+1} + A_h^{\text{sip}} u_h^{n+1} = f_h^{n+1}, \quad (4.92)$$

with the initial condition $u_h^0 = \pi_h u_0$ and the source term $f_h^{n+1} = \pi_h f^{n+1}$ for all $n \in \{0, \dots, N-1\}$. Problem (4.92) can be equivalently rewritten as

$$u_h^{n+1} + \delta t A_h^{\text{sip}} u_h^{n+1} = u_h^n + \delta t f_h^{n+1},$$

thus highlighting the fact that u_h^{n+1} is obtained from u_h^n by solving a linear problem. In what follows, we abbreviate as $a \lesssim b$ the inequality $a \leq Cb$ with positive C independent of h , δt , and f .

4.7.3 Error Estimates

The analysis follows a path similar to that deployed in Sect. 3.1 for the unsteady advection-reaction equation. We first derive the error equation, then establish an energy estimate and finally infer the convergence result. The analysis is much simpler than in Sect. 3.1 since we use an implicit scheme to march in time. Indeed, contrary to explicit schemes, implicit schemes are dissipative.

Letting

$$\xi_h^n := u_h^n - \pi_h u^n, \quad \xi_\pi^n := u^n - \pi_h u^n, \quad (4.93)$$

the approximation error at the discrete time t^n is decomposed as

$$u^n - u_h^n = \xi_\pi^n - \xi_h^n.$$

Lemma 4.72 (Error equation). *Assume $u \in C^0(V_*) \cap C^2(L^2(\Omega))$. Then,*

$$\delta_t^{(1)} \xi_h^{n+1} = -A_h^{\text{sip}} \xi_h^{n+1} + \alpha_h^{n+1}, \quad (4.94)$$

with $\alpha_h^{n+1} := A_h^{\text{sip}} \xi_\pi^{n+1} - \pi_h \theta^{n+1}$ and $\theta^{n+1} := \delta t^{-1} \int_{t^n}^{t^{n+1}} (t^n - t) d_t^2 u(t) dt$.

Proof. A Taylor expansion in time yields

$$u^n = u^{n+1} - \delta t d_t u^{n+1} - \int_{t^n}^{t^{n+1}} (t^n - t) d_t^2 u(t) dt, \quad (4.95)$$

i.e.,

$$\delta_t^{(1)} u^{n+1} = d_t u^{n+1} + \theta^{n+1}.$$

Projecting this equation onto V_h and replacing $\pi_h d_t u^{n+1}$ by $f_h^{n+1} - A_h^{\text{sip}} u^{n+1}$ owing to consistency, we obtain

$$\delta_t^{(1)} \pi_h u^{n+1} + A_h^{\text{sip}} u^{n+1} = f_h^{n+1} + \pi_h \theta^{n+1}. \quad (4.96)$$

Subtracting (4.96) from (4.92) yields the assertion. \square

We now derive an energy estimate for the discrete scheme.

Lemma 4.73 (Energy estimate). *Assume $u \in C^0(V_*) \cap C^2(L^2(\Omega))$. Then, for all $n \in \{0, \dots, N-1\}$, there holds*

$$\begin{aligned} \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 - \|\xi_h^n\|_{L^2(\Omega)}^2 + \|\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \|\xi_h^{n+1}\|_{\text{sip}}^2 \\ \lesssim \delta t (\|\xi_\pi^{n+1}\|_{\text{sip},*}^2 + C_u^2 \delta t^2), \end{aligned} \quad (4.97)$$

with $C_u := \|d_t^2 u\|_{C^0(L^2(\Omega))}$.

Proof. Testing (4.94) with $\delta t \xi_h^{n+1}$, we obtain

$$\begin{aligned} \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 + \delta t (A_h^{\text{sip}} \xi_h^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} \\ = (\xi_h^n, \xi_h^{n+1})_{L^2(\Omega)} + \delta t (A_h^{\text{sip}} \xi_\pi^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} - \delta t (\theta^{n+1}, \xi_h^{n+1})_{L^2(\Omega)}, \end{aligned}$$

since $(\pi_h \theta^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} = (\theta^{n+1}, \xi_h^{n+1})_{L^2(\Omega)}$. Using the algebraic relation $ab = \frac{1}{2}a^2 + \frac{1}{2}b^2 - \frac{1}{2}(a-b)^2$ for the first term on the right-hand side, the above energy equality becomes

$$\begin{aligned} \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 + \|\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2 + 2\delta t (A_h^{\text{sip}} \xi_h^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} \\ = \|\xi_h^n\|_{L^2(\Omega)}^2 + 2\delta t (A_h^{\text{sip}} \xi_\pi^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} - 2\delta t (\theta^{n+1}, \xi_h^{n+1})_{L^2(\Omega)}. \end{aligned}$$

Using discrete coercivity and boundedness of A_h^{sip} together with the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 + \|\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2 + 2\delta t C_{\text{sta}} \|\xi_h^{n+1}\|_{\text{sip}}^2 \\ \leq \|\xi_h^n\|_{L^2(\Omega)}^2 + 2C_{\text{bnd}} \delta t \|\xi_\pi^{n+1}\|_{\text{sip},*} \|\xi_h^{n+1}\|_{\text{sip}} + 2\delta t \|\theta^{n+1}\|_{L^2(\Omega)} \|\xi_h^{n+1}\|_{L^2(\Omega)}. \end{aligned}$$

Owing to the discrete Poincaré inequality (4.20),

$$\begin{aligned} \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 + \|\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2 + 2\delta t C_{\text{sta}} \|\xi_h^{n+1}\|_{\text{sip}}^2 \\ \leq \|\xi_h^n\|_{L^2(\Omega)}^2 + 2C_{\text{bnd}} \delta t \|\xi_\pi^{n+1}\|_{\text{sip},*} \|\xi_h^{n+1}\|_{\text{sip}} + 2\sigma_2 \delta t \|\theta^{n+1}\|_{L^2(\Omega)} \|\xi_h^{n+1}\|_{\text{sip}}. \end{aligned}$$

Using Young’s inequality for the two rightmost terms yields

$$\begin{aligned} \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 + \|\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \|\xi_h^{n+1}\|_{\text{sip}}^2 \\ \leq \|\xi_h^n\|_{L^2(\Omega)}^2 + C \delta t (\|\xi_\pi^{n+1}\|_{\text{sip},*}^2 + \|\theta^{n+1}\|_{L^2(\Omega)}^2). \end{aligned}$$

Finally, proceeding as in the proof of Lemma 3.20, we infer

$$\|\theta^{n+1}\|_{L^2(\Omega)} \lesssim C_u \delta t,$$

whence the assertion. \square

Remark 4.74 (Dissipation in backward Euler scheme). The dissipative nature of the backward Euler scheme is reflected by the presence of the time increment $\|\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2$ on the left-hand side of the energy estimate (4.97). We observe that, up to a factor δt , the increment $\xi_h^{n+1} - \xi_h^n$ can be interpreted as a first-order finite difference approximation of the time derivative of the error component in V_h .

Finally, we arrive at our main convergence result.

Theorem 4.75 (Convergence). *Let $u \in C^0(V_*) \cap C^2(L^2(\Omega))$ solve (4.89) and let $(u_h^n)_{1 \leq n \leq N}$ solve (4.92) with $u_h^0 = \pi_h u_0$. Assume $u \in C^0(H^{k+1}(\Omega))$. Then, there holds*

$$\|u^N - u_h^N\|_{L^2(\Omega)} + \left(C_{\text{sta}} \sum_{n=1}^N \delta t \|u^n - u_h^n\|_{\text{sip}}^2 \right)^{1/2} \lesssim \chi_1 \delta t + \chi_2 h^k, \quad (4.98)$$

with $\chi_1 = t_F^{1/2} \|d_t^2 u\|_{C^0(L^2(\Omega))}$ and $\chi_2 = \|u\|_{C^0(H^{k+1}(\Omega))}$.

Proof. Summing (4.97) for $n \in \{0, \dots, N-1\}$, dropping the nonnegative contribution $\|\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2$, and observing that $\xi_h^0 = 0$, we obtain

$$\|\xi_h^N\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \sum_{n=1}^N \|\xi_h^n\|_{\text{sip}}^2 \lesssim \delta t \sum_{n=1}^N \|\xi_\pi^n\|_{\text{sip},*}^2 + t_F C_u^2 \delta t^2.$$

Recalling the results of Sect. 4.2.3, we infer that, for all $n \in \{1, \dots, N\}$, $\|\xi_\pi^n\|_{\text{sip},*} \lesssim h^k \|u^n\|_{H^{k+1}(\Omega)}$. Hence,

$$\|\xi_h^N\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \sum_{n=1}^N \|\xi_h^n\|_{\text{sip}}^2 \lesssim (\chi_1 \delta t + \chi_2 h^k)^2.$$

Using the triangle inequality $\|u^N - u_h^N\|_{L^2(\Omega)} \leq \|\xi_\pi^N\|_{L^2(\Omega)} + \|\xi_h^N\|_{L^2(\Omega)}$ together with

$$\sum_{n=1}^N \delta t \|u^n - u_h^n\|_{\text{sip}}^2 \leq 2 \sum_{n=1}^N \delta t (\|\xi_h^n\|_{\text{sip}}^2 + \|\xi_\pi^n\|_{\text{sip}}^2),$$

and observing that

$$\|\xi_\pi^N\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \sum_{n=1}^N \|\xi_\pi^n\|_{\text{sip}}^2 \lesssim (\chi_2 h^k)^2,$$

yields the assertion. \square

4.7.4 BDF2 Time Discretization

To improve the convergence rate in time, we can consider higher-order backward approximations of the time derivative $d_t u$. In this section, we briefly examine time discretization using the *second-order backward difference formula (BDF2)*. We show, in particular, that, also in this case, stability is related to the dissipative nature of the scheme. We proceed as in Sect. 4.7.3, whereby we derive the error equation, establish an energy estimate, and finally infer the convergence result.

For a function $v \in C^1(V)$, we introduce the BDF2 operator $\delta_t^{(2)}$ such that, for all $n \in \{2, \dots, N\}$,

$$\delta_t^{(2)} v^n := \frac{3v^n - 4v^{n-1} + v^{n-2}}{2\delta t} \in V,$$

thereby providing a second-order finite difference approximation of the time derivative $d_t v^n$. Then, the discrete solution is obtained from

$$\delta_t^{(1)} u_h^1 + A_h^{\text{sip}} \frac{u_h^0 + u_h^1}{2} = \frac{f_h^0 + f_h^1}{2}, \quad (4.99a)$$

$$\delta_t^{(2)} u_h^{n+1} + A_h^{\text{sip}} u_h^{n+1} = f_h^{n+1} \quad \text{for } n \in \{1, \dots, N-1\}, \quad (4.99b)$$

with $u_h^0 = \pi_h u_0$. The operator $\delta_t^{(2)}$ cannot be used for the first time step $n = 1$, since only the initial value is available. In (4.99a), the value u_h^1 is computed from u_h^0 using the Crank–Nicolson scheme which is also second-order accurate in time.

We first derive the error equation, recalling that the components ξ_h^n and ξ_π^n of the approximation errors are defined by (4.93).

Lemma 4.76 (Error equation). *Assume $u \in C^0(V_*) \cap C^3(L^2(\Omega))$. Then,*

$$\delta_t^{(1)} \xi_h^1 + A_h^{\text{sip}} \frac{\xi_h^0 + \xi_h^1}{2} = \alpha_h^1, \quad (4.100)$$

where $\alpha_h^1 := A_h^{\text{sip}} \frac{\xi_\pi^0 + \xi_\pi^1}{2} - \pi_h \theta^1$, $\theta^1 := -\frac{1}{2} \delta t^{-1} \int_0^{\delta t} t(\delta t - t) d_t^3 u(t) dt$, and, for all $n \in \{1, \dots, N-1\}$,

$$\delta_t^{(2)} \xi_h^{n+1} + A_h^{\text{sip}} \xi_h^{n+1} = \alpha_h^{n+1}, \quad (4.101)$$

where $\alpha_h^{n+1} := A_h^{\text{sip}} \xi_\pi^{n+1} - \pi_h \theta^{n+1}$ and

$$\theta^{n+1} := \delta t^{-1} \int_{t^n}^{t^{n+1}} (t^n - t)^2 d_t^3 u(t) dt - \frac{1}{4} \delta t^{-1} \int_{t^{n-1}}^{t^{n+1}} (t^{n-1} - t)^2 d_t^3 u(t) dt.$$

Proof. We observe that

$$\begin{aligned} \delta_t^{(1)} u^1 &= d_t u^1 - \delta t^{-1} \int_0^{\delta t} t d_t^2 u(t) dt, \\ \delta_t^{(1)} u^1 &= d_t u^0 + \delta t^{-1} \int_0^{\delta t} (\delta t - t) d_t^2 u(t) dt, \end{aligned}$$

so that, integrating by parts in time,

$$\delta_t^{(1)} u^1 = d_t \frac{u^0 + u^1}{2} + \frac{1}{2} \delta t^{-1} \int_0^{\delta t} (\delta t - 2t) d_t^2 u(t) dt = d_t \frac{u^0 + u^1}{2} + \theta^1.$$

Proceeding as in the proof of Lemma 4.72 yields (4.100). Furthermore, for all $n \in \{2, \dots, N\}$, a direct calculation shows that

$$\delta_t^{(2)} u^{n+1} = d_t u^{n+1} + \theta^{n+1},$$

and proceeding again as in the proof of Lemma 4.72 yields (4.101). \square

We can now derive the energy estimate.

Lemma 4.77 (Energy estimate). *Assume $u \in C^0(V_*) \cap C^3(L^2(\Omega))$. Then, there holds*

$$\|\xi_h^1\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \|\xi_h^1\|_{\text{sip}}^2 \lesssim \delta t (\|\xi_\pi^1\|_{\text{sip},*}^2 + \|\xi_\pi^0\|_{\text{sip},*}^2 + C_u^2 \delta t^4), \quad (4.102)$$

and, for all $n \in \{1, \dots, N-1\}$,

$$\begin{aligned} & \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 - \|\xi_h^n\|_{L^2(\Omega)}^2 + \|2\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2 - \|2\xi_h^n - \xi_h^{n-1}\|_{L^2(\Omega)}^2 \\ & + \|\delta_{tt}\xi_h^{n+1}\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \|\xi_h^{n+1}\|_{\text{sip}}^2 \lesssim \delta t (\|\xi_\pi^{n+1}\|_{\text{sip},*}^2 + C_u^2 \delta t^4), \end{aligned} \quad (4.103)$$

with $C_u := \|d_t^3 u\|_{C^0(L^2(\Omega))}$ and $\delta_{tt}\xi_h^{n+1} := \xi_h^{n+1} - 2\xi_h^n + \xi_h^{n-1}$.

Proof. Testing (4.100) with $\delta t \xi_h^1$, observing that $\xi_h^0 = 0$, rearranging terms, and using $\|\theta^1\|_{L^2(\Omega)} \lesssim C_u \delta t^2$ yields (4.102). Furthermore, testing (4.101) with $4\delta t \xi_h^{n+1}$, we infer

$$4\delta t (\delta_t^{(2)} \xi_h^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} + 4\delta t (A_h \xi_h^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} = 4\delta t (\alpha_h^{n+1}, \xi_h^{n+1})_{L^2(\Omega)}.$$

We observe that

$$\begin{aligned} 4\delta t (\delta_t^{(2)} \xi_h^{n+1}, \xi_h^{n+1})_{L^2(\Omega)} &= \|\xi_h^{n+1}\|_{L^2(\Omega)}^2 - \|\xi_h^n\|_{L^2(\Omega)}^2 \\ &+ \|2\xi_h^{n+1} - \xi_h^n\|_{L^2(\Omega)}^2 - \|2\xi_h^n - \xi_h^{n-1}\|_{L^2(\Omega)}^2 \\ &+ \|\delta_{tt}\xi_h^{n+1}\|_{L^2(\Omega)}^2. \end{aligned}$$

Finally, we bound $(\alpha_h^{n+1}, \xi_h^{n+1})_{L^2(\Omega)}$ by proceeding as in the proof of Lemma 4.73 using $\|\theta^{n+1}\|_{L^2(\Omega)} \lesssim C_u \delta t^2$ for all $n \in \{1, \dots, N-1\}$. \square

Remark 4.78 (Dissipation in BDF2 scheme). The dissipative nature of the BDF2 scheme is reflected by the term $\|\delta_{tt}\xi_h^{n+1}\|_{L^2(\Omega)}^2$ on the left-hand side of (4.103). We observe that, up to a factor δt^2 , $\delta_{tt}\xi_h^{n+1}$ can be interpreted as a second-order finite difference approximation of the second-order time derivative of the error component in V_h .

Finally, we arrive at our main convergence result. The proof is skipped since it is similar to that of Theorem 4.75.

Theorem 4.79 (Convergence). *Let $u \in C^0(V_*) \cap C^3(L^2(\Omega))$ solve (4.89) and let $(u_h^n)_{1 \leq n \leq N}$ solve (4.99) with $u_h^0 = \pi_h u_0$. Assume $u \in C^0(H^{k+1}(\Omega))$. Then, there holds*

$$\|u^N - u_h^N\|_{L^2(\Omega)} + \left(C_{\text{sta}} \sum_{n=1}^N \delta t \|u^n - u_h^n\|_{\text{sip}}^2 \right)^{1/2} \lesssim \chi_1 \delta t^2 + \chi_2 h^k,$$

with $\chi_1 = t_F^{1/2} \|d_t^3 u\|_{C^0(L^2(\Omega))}$ and $\chi_2 = \|u\|_{C^0(H^{k+1}(\Omega))}$.

4.7.5 Improved $C^0(L^2(\Omega))$ -Error Estimate

The error estimate (4.98) is suboptimal for the term $\|u^N - u_h^N\|_{L^2(\Omega)}$ by one power in h . Following the ideas of Wheeler [305], a sharper result can be obtained by replacing the L^2 -orthogonal projector in (4.93) by the so-called *elliptic projector* $\pi_{\text{ell}} \in \mathcal{L}(V_*, V_h)$ such that, for all $w \in V_*$, $\pi_{\text{ell}} w \in V_h$ solves

$$a_h^{\text{sip}}(\pi_{\text{ell}} w, v_h) = a_h^{\text{sip}}(w, v_h) \quad \forall v_h \in V_h,$$

or, equivalently,

$$A_h^{\text{sip}} \pi_{\text{ell}} w = A_h^{\text{sip}} w.$$

If elliptic regularity holds (cf. Definition 4.24), there is C , independent of h , such that, for all $w \in V_* \cap H^{k+1}(\Omega)$,

$$\|w - \pi_{\text{ell}} w\|_{L^2(\Omega)} \leq C h^{k+1} \|w\|_{H^{k+1}(\Omega)}. \quad (4.104)$$

Redefining the components of the approximation error as

$$\zeta_\pi^n := u^n - \pi_{\text{ell}} u^n, \quad \zeta_h^n := \pi_{\text{ell}} u^n - u_h^n,$$

the approximation error at the discrete time t^n is now decomposed as

$$u^n - u_h^n = \zeta_\pi^n + \zeta_h^n.$$

We consider again the backward Euler operator $\delta_t^{(1)}$ defined by (4.91).

Lemma 4.80 (Error equation). *Assume $u \in C^0(V_*) \cap C^2(L^2(\Omega))$. Then,*

$$\delta_t^{(1)} \zeta_h^{n+1} + A_h^{\text{sip}} \zeta_h^{n+1} = \alpha_h^{n+1}, \quad (4.105)$$

with $\alpha_h^{n+1} := \pi_h \delta_t^{(1)} \zeta_\pi^{n+1} - \pi_h \theta^{n+1}$ and θ^{n+1} defined in Lemma 4.72.

Proof. Recalling that $\delta_t^{(1)} u^{n+1} = d_t u^{n+1} + \theta^{n+1}$ and observing that $\delta_t^{(1)} u^{n+1} = \delta_t^{(1)} \pi_{\text{ell}} u^{n+1} + \delta_t^{(1)} \zeta_\pi^{n+1}$, we obtain

$$\delta_t^{(1)} \pi_{\text{ell}} u^{n+1} = d_t u^{n+1} + \theta^{n+1} - \delta_t^{(1)} \zeta_\pi^{n+1}.$$

Projecting this equation onto V_h , replacing $\pi_h d_t u^{n+1}$ by $f_h^{n+1} - A_h^{\text{sip}} u^{n+1}$, and observing that $A_h^{\text{sip}} \pi_{\text{ell}} u^{n+1} = A_h^{\text{sip}} u^{n+1}$, we infer

$$\delta_t^{(1)} \pi_{\text{ell}} u^{n+1} = f_h^{n+1} - A_h^{\text{sip}} \pi_{\text{ell}} u^{n+1} + \pi_h \theta^{n+1} - \pi_h \delta_t^{(1)} \zeta_\pi^{n+1}.$$

Subtracting this relation from (4.92) yields the assertion. \square

The difference between (4.105) and (4.94) lies in the residual α_h^{n+1} . When using the elliptic projector, the term $A_h^{\text{sip}} \zeta_\pi^{n+1}$ in α_h^{n+1} is replaced $\pi_h \delta_t^{(1)} \zeta_\pi^{n+1}$. This is a key point, since the latter scales in space as h^{k+1} (cf. the proof of Theorem 4.82 below), while the former only scales as h^k .

The next step is to derive an energy estimate. The proof is skipped since it is similar to that of Lemma 4.73.

Lemma 4.81 (Energy estimate). *For all $n \in \{0, \dots, N-1\}$, there holds*

$$\begin{aligned} \|\zeta_h^{n+1}\|_{L^2(\Omega)}^2 - \|\zeta_h^n\|_{L^2(\Omega)}^2 + \|\zeta_h^{n+1} - \zeta_h^n\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \|\zeta_h^{n+1}\|_{\text{sip}}^2 \\ \lesssim \delta t (\|\delta_t^{(1)} \zeta_\pi^{n+1}\|_{L^2(\Omega)}^2 + C_u^2 \delta t^2), \end{aligned} \quad (4.106)$$

with $C_u := \|d_t^2 u\|_{C^0(L^2(\Omega))}$.

Finally, we arrive at our improved convergence result.

Theorem 4.82 (Convergence). *Let $u \in C^0(V_*) \cap C^2(L^2(\Omega))$ solve (4.89) and additionally assume that $u \in C^1(H^{k+1}(\Omega))$. Then,*

$$\|u^N - u_h^N\|_{L^2(\Omega)} \lesssim \chi_1 \delta t + \chi_2 h^{k+1},$$

with $\chi_1 = t_F^{1/2} \|d_t^2 u\|_{C^0(L^2(\Omega))}$ and $\chi_2 = t_F^{1/2} \|u\|_{C^1(H^{k+1}(\Omega))}$.

Proof. We first observe that

$$\delta_t^{(1)} \zeta_\pi^{n+1} = \pi_{\text{ell}} \delta_t^{(1)} u^{n+1} - \delta_t^{(1)} u^{n+1},$$

so that, owing to (4.104),

$$\|\delta_t^{(1)} \zeta_\pi^{n+1}\|_{L^2(\Omega)} \lesssim h^{k+1} \|\delta_t^{(1)} u^{n+1}\|_{H^{k+1}(\Omega)}.$$

Moreover,

$$\begin{aligned} \|\delta_t^{(1)} u^{n+1}\|_{H^{k+1}(\Omega)}^2 &= \sum_{|\alpha| \leq k+1} \int_\Omega \frac{1}{\delta t^2} \left| \int_{t^n}^{t^{n+1}} \partial^\alpha d_t u(s) \, ds \right|^2 \\ &\leq \sum_{|\alpha| \leq k+1} \int_\Omega \frac{1}{\delta t} \int_{t^n}^{t^{n+1}} |\partial^\alpha d_t u(s)|^2 \, ds \\ &= \frac{1}{\delta t} \int_{t^n}^{t^{n+1}} \|d_t u(s)\|_{H^{k+1}(\Omega)}^2 \, ds \\ &\leq \|d_t u\|_{C^0(H^{k+1}(\Omega))}^2 \leq \|u\|_{C^1(H^{k+1}(\Omega))}^2. \end{aligned}$$

Then, summing (4.106) for $n \in \{0, \dots, N-1\}$, dropping the nonnegative contribution $\|\zeta_h^{n+1} - \zeta_h^n\|_{L^2(\Omega)}^2$, and observing that $\zeta_h^0 = 0$ we obtain

$$\|\zeta_h^N\|_{L^2(\Omega)}^2 + \delta t C_{\text{sta}} \sum_{n=1}^N \|\zeta_h^n\|_{\text{sip}}^2 \lesssim (\chi_1 \delta t + \chi_2 h^{k+1})^2. \quad (4.107)$$

Owing to the triangle inequality, $\|u^N - u_h^N\|_{L^2(\Omega)} \leq \|\zeta_\pi^N\|_{L^2(\Omega)} + \|\zeta_h^N\|_{L^2(\Omega)}$ and we conclude using (4.104). \square

Remark 4.83 (Superconvergence of $\delta t \sum_{n=1}^N \|\zeta_h^n\|_{\text{sip}}^2$). The error decomposition based on ζ_h^n and ζ_π^n can also be used to derive an energy-error estimate of order h^k in space. The bound (4.107) shows that $\left(\sum_{n=1}^N \delta t \|\zeta_h^n\|_{\text{sip}}^2\right)^{1/2}$ scales as h^{k+1} , and, therefore, superconverges. The leading order term in the energy-error estimate is the projection term $\left(\sum_{n=1}^N \delta t \|\zeta_\pi^n\|_{\text{sip}}\right)^{1/2}$ which scales as h^k .

Mathematical Aspects of Discontinuous Galerkin
Methods

Di Pietro, D.A.; Ern, A.

2012, XVII, 384 p. 34 illus., Softcover

ISBN: 978-3-642-22979-4