Calcolo (2018) 55:40
https://doi.org/10.1007/s10092-018-0282-3

A third Strang lemma and an Aubin–Nitsche trick for schemes in fully discrete formulation

Daniele A. Di Pietro[1] · **Jérôme Droniou**[2]

Received: 16 August 2018 / Accepted: 27 August 2018

Abstract
In this work, we present an abstract error analysis framework for the approximation of linear partial differential equation problems in weak formulation. We consider approximation methods in fully discrete formulation, where the discrete and continuous spaces are possibly not embedded in a common space. A proper notion of consistency is designed, and, under a classical inf–sup condition, it is shown to bound the approximation error. This error estimate result is in the spirit of Strang's first and second lemmas, but applicable in situations not covered by these lemmas (because of a fully discrete approximation space). An improved estimate is also established in a weaker norm, using the Aubin–Nitsche trick. We then apply these abstract estimates to an anisotropic heterogeneous diffusion model and two classical families of schemes for this model: virtual element and finite volume methods. For each of these methods, we show that the abstract results yield new error estimates with a precise and mild dependency on the local anisotropy ratio. A key intermediate step to derive such estimates for virtual element methods is proving optimal approximation properties of the oblique elliptic projector in weighted Sobolev seminorms. This is a result whose interest goes beyond the specific model and methods considered here. We also obtain, to our knowledge, the first clear notion of consistency for finite volume methods, which leads to a generic error estimate involving the fluxes and valid for a wide range of finite volume schemes. An important application is the first error estimate for multi-point flux approximation L and G methods.

Keywords Strang lemma · Consistency · Error estimate · Aubin–Nitsche trick · Virtual element methods · Finite volume methods · Oblique elliptic projector

✉ Jérôme Droniou
jerome.droniou@monash.edu

Daniele A. Di Pietro
daniele.di-pietro@umontpellier.fr

[1] Institut Montpelliérain Alexander Grothendieck, Univ. Montpellier, CNRS, Montpellier, France

[2] School of Mathematical Sciences, Monash University, Melbourne, Australia

Published online: 03 September 2018

Mathematics Subject Classification 65N08 · 65N12 · 65N15 · 65N30

1 Introduction

The second Strang lemma [43,44] is probably the most famous error estimate result for Finite Element Methods, and is used as a starting point for the analysis of non-conforming methods in many reference textbooks; see, e.g., [16,35]. In recent years, it has been generalised to novel technologies including, e.g., Discontinuous Galerkin (DG) [22, Section 1.3] and Virtual Element methods (VEM) [13, Theorem 2]. In a nutshell, given Hilbert spaces V and V_h, a bilinear form $a(\cdot,\cdot)$ (resp. $a_h(\cdot,\cdot)$) and a linear form $\ell(\cdot)$ (resp. $\ell_h(\cdot)$) defined on V (resp. V_h), and considering the continuous and discrete problems

$$\text{Find } u \in V \text{ such that } \quad a(u,v) = \ell(v) \quad \forall v \in V$$

and

$$\text{Find } u_h \in V_h \text{ such that } \quad a_h(u_h,v_h) = \ell_h(v_h) \quad \forall v_h \in V_h, \tag{1}$$

the second Strang lemma provides, under boundedness and inf–sup conditions on a_h, a bound on a proper norm of $u - u_h$ in terms of quantities measuring the approximation properties of V by V_h and the consistency of the discrete problem.

This result has two major constraints:

(i) V and V_h must be subspaces of a common space of functions, to ensure that the sum $V + V_h$ is well defined,
(ii) $a_h(\cdot,\cdot)$ must be extended to $V + V_h$, such that its restriction to V is consistent with $a(\cdot,\cdot)$ in an appropriate way.

The first constraint is not an issue for Finite Element, DG methods or VEM, whose natural unknowns are functions, but it is not satisfied by a number of other methods such as, e.g., Hybrid High-Order [23], Mimetic Finite Differences [9], cell- and face-centred Finite Volume methods (such as Hybrid Mimetic Mixed methods [27,30,37]), even though, in some of these methods, some components of vectors in V_h represent functions on the mesh cells, other components can represent unknowns/functions on the mesh faces, at the mesh vertices, etc.

Even for methods that clearly satisfy (i), the second constraint can raise some challenges. For example, in DG methods, the extension of a_h can often be made only in $V_* + V_h$, where V_* is a strict subspace of V; see, e.g., [22, Section 1.3.3]. Possible ways of circumventing the difficulties linked to the insufficient regularity of the exact solution have been proposed, e.g., in [39] (trimmed error estimates) and, more recently, in [34] (mollified error estimates). Other difficulties may be inherent to the approach used to construct the discretisation. In VEM, the discrete bilinear form $a_h(\cdot,\cdot)$ often contains contributions defined in an algebraic way that make its extension to V not obvious; the Strang-like estimate of [13] circumvents this question of extension of $a_h(\cdot,\cdot)$, at the expense of additional terms, by extending instead the

continuous form $a(\cdot,\cdot)$ to $V + V_h$. Another example can be found in the family of cell-centered Finite Volume methods [3,26,33,38]: even if the unknowns can be considered, in these methods, as piecewise constant functions on the mesh, and their formulation can be written as (1), the resulting bilinear form $a_h(\cdot,\cdot)$ is written in a fully discrete form that makes its extension to a space of functions, and the subsequent analysis, more involved.

In this work, we propose a new error analysis framework, for problems written in weak (Petrov–Galerkin) form, that is free from the two constraints mentioned above. The main idea is to estimate a *discrete* approximation error $I_h u - u_h$, where $I_h : V \to V_h$ is a well chosen interpolant of functions onto the discrete space. With this definition of the approximation error, the exact solution u *need not* be plugged into the discrete bilinear form to write the error equation. Instead, its role is played by $I_h u$. The discrete approximation error can then be estimated solely in terms of the (discrete) norm in V_h under a *stability* assumption on a_h (an inf–sup condition) and in terms of only one *consistency* measure involving $I_h u$, a_h and ℓ_h; see Theorem 10 below. As a by-product, our analysis provides a clear definition of such a consistency for a wide range of methods, including many for which this notion was never clearly highlighted. The abstract error estimate also enables us to write, in this generic setting, the well-known principle in finite differences:

$$\text{stability} + \text{consistency} \implies \text{convergence}. \tag{2}$$

In Theorem 13 below we also establish, under a consistency assumption of the solution to the continuous dual problem, an estimate in a weaker norm than that of V, which mimics classical improved error estimates (e.g., in L^2 norm when the energy space of the problem is H^1) for Finite Elements, DG, etc.

The abstract analysis framework is then used to derive error estimates for a variety of methods for the discretisation of a variable diffusion problem. The first application is to conforming and non-conforming VEM, for which we derive an energy error estimate; see Theorem 19 below. This estimate is similar to [13, Theorem 6.2], but two additional features deserve to be highlighted: it is obtained as a consequence of a general abstract framework, and the dependencies with respect to the diffusion tensor are accurately tracked. In particular, this estimate reveals that the multiplicative constant in the right-hand side is independent of the heterogeneity of the diffusion field, but depends on the square root of the (local) anisotropy ratio, a behaviour already documented for Hybrid High-Order (HHO) methods; see, e.g., [21]. A unified L^2 error estimate covering both conforming and non-conforming VEM is provided in Theorem 22. Establishing the VEM error estimates requires optimal approximation properties of the oblique elliptic projector. These properties, that are also of interest for other high-order methods (e.g. the HHO method), are the purpose of Sect. 3.2.1; their proof relies on the classical Dupont–Scott approximation theory [12,32].

In the second application, we consider finite volume (FV) methods, both cell-centred and cell- and face-centred. The notion of consistency for such methods has been discussed in various references (see e.g. [38, Section 2.1] or [26, Remark 1.3]), but never directly related to error estimates. We show that the abstract analysis framework yields such estimates in terms of the consistency error purely based on the fluxes. As

in the case of VEM, this error estimate is established in a diffusion-dependent discrete norm, which enables us to explicitly track the local dependencies with respect to the diffusion tensor. As an important application, we obtain the first error estimate for Multi-Point Flux Approximation L and G methods. Several papers have tackled the question of designing a uniform convergence analysis framework for finite volume element methods [14,15,36,41], which are specific forms of finite volume methods on triangles/tetrahedra obtained by writing a balance of fluxes of a conforming or non-conforming $\mathbb{P}_1$ finite element function over a dual mesh. In these references, error estimates are obtained by writing these methods under a Petrov–Galerkin formulation (4). These estimates do not come from consistency errors of the fluxes but, in [14] for example, from consistency errors involving $a - a_h$ and $\ell - \ell_h$; additionally, they are obtained under a global Lipschitz assumption on the diffusion tensor, and do not track the local dependency on its anisotropy ratio. Estimates in terms of flux consistencies (as in Theorems 27 and 29 below) seem natural in the FV setting, and enable us to encompass all finite volume methods, including important practical ones such as MPFA schemes (that are not finite volume element methods).

The rest of the paper is organised as follows. In Sect. 2 we present the abstract analysis framework. The main error estimates are stated in Theorems 10 (energy norm) and 13 (weaker norm). Applications of the abstract analysis framework to VEM and FV methods are considered in Sect. 3. Finally, some conclusions are drawn in Sect. 4.

2 Abstract analysis framework

2.1 Setting

We consider here a setting where the continuous and discrete problems are both written under variational formulations. For the continuous problem, we take

- A Hilbert space H,
- A continuous bilinear form $a : H \times H \to \mathbb{R}$,
- A continuous linear form $\ell : H \to \mathbb{R}$.

The problem we aim at approximating is

$$\text{Find } u \in H \text{ such that } \quad a(u, v) = \ell(v) \quad \forall v \in H. \tag{3}$$

In what follows, problem (3) is named the *continuous problem* in reference to the fact that the space H is usually infinite dimensional.

Remark 1 (*Existence of a continuous solution*) We assume the existence of a solution to problem (3); this existence follows for example from the Lax-Milgram or Babuška-Brezzi lemmas if a is coercive or satisfies an inf–sup condition.

Our approximation is written in fully discrete Petrov–Galerkin form, using trial and test spaces that are possibly different from each other, and not necessarily spaces of functions. In particular, they are not necessarily embedded in any natural space in which H is also embedded. We consider thus

- Two vector spaces X_h and Y_h, with respective norms $\|\cdot\|_{X_h}$ and $\|\cdot\|_{Y_h}$.
- A bilinear form $a_h : X_h \times Y_h \to \mathbb{R}$.
- A linear form $\ell_h : Y_h \to \mathbb{R}$.

Remark 2 (*Discrete spaces*) The spaces X_h and Y_h are always finite-dimensional in applications, but this is not required in our analysis. The index h represents a discretisation parameter (e.g., the meshsize) which characterises these spaces, and such that convergence of the method (in a sense to be made precise) is expected when $h \to 0$. Likewise, the continuity of a_h or ℓ_h is not directly used, but is always verified in practice, and of course usually required to ensure the existence of a solution.

The approximation of problem (3) is

$$\text{Find } u_h \in X_h \text{ such that } \quad a_h(u_h, v_h) = \ell_h(v_h) \quad \forall v_h \in Y_h. \tag{4}$$

In what follows, (4) is named the *discrete problem*, in reference to the fact that the spaces X_h and Y_h are usually finite dimensional. We intend to compare the solutions to (3) and (4) by estimating $u_h - I_h u$, where $I_h u$ is an element of X_h representative of the solution u to (3); see Remark 9.

Remark 3 (*Equivalent Galerkin formulation*) When the spaces X_h and Y_h are finite-dimensional, their dimensions must coincide in order for the discrete problem (4) to be well-posed. In this case, there exists an isomorphism $\mathfrak{I}_h : X_h \to Y_h$, and an equivalent Galerkin formulation can be written based on the linear and bilinear forms $\widetilde{\ell}_h : X_h \to \mathbb{R}$ and $\widetilde{a}_h : X_h \times X_h \to \mathbb{R}$ such that $\widetilde{\ell}_h(v_h) = \ell_h(\mathfrak{I}_h v_h)$ and $\widetilde{a}_h(u_h, v_h) = a_h(u_h, \mathfrak{I}_h v_h)$ for all $u_h, v_h \in X_h$.

2.2 Error estimate in energy norm

We now describe a notion of stability of a_h that yields a bound on the solutions to (4)

Definition 4 (*Inf–sup stability*) The bilinear form a_h is inf–sup stable for $(\|\cdot\|_{X_h}, \|\cdot\|_{Y_h})$ if

$$\exists \gamma > 0 \text{ such that } \gamma \|u_h\|_{X_h} \leq \sup_{v_h \in Y_h \setminus \{0\}} \frac{a_h(u_h, v_h)}{\|v_h\|_{Y_h}} \quad \forall u_h \in X_h. \tag{5}$$

Remark 5 (*Uniform inf–sup stability*) In practice, one typically requires that the real number γ is independent of discretization parameters such as the meshsize. Hence, condition (5) should be verified uniformly with respect to h. This is needed to have optimal error estimates.

Remark 6 (*Coercivity*) The inf–sup stability is of course satisfied if $X_h = Y_h$ and a_h is coercive in the sense that $a_h(v_h, v_h) \geq \gamma \|v_h\|^2_{X_h}$ for all $v_h \in X_h$, where γ does not depend on v_h.

We next prove an a priori bound on the discrete solution. To this end, we recall that, if Z is a Banach space with norm $\|\cdot\|_Z$, the dual norm of a linear form $\mu : Z \to \mathbb{R}$ is classically defined by

$$\|\mu\|_{Z^\star} = \sup_{z \in Z\setminus\{0\}} \frac{|\mu(z)|}{\|z\|_Z}. \tag{6}$$

Proposition 7 (A priori bound on the discrete solution) *If a_h is inf–sup stable in the sense of Definition 4, $m_h : Y_h \to \mathbb{R}$ is linear, and w_h satisfies*

$$a_h(w_h, v_h) = m_h(v_h) \quad \forall v_h \in Y_h,$$

then

$$\|w_h\|_{X_h} \le \gamma^{-1} \|m_h\|_{Y_h^\star}.$$

Proof Take $v_h \in Y_h\setminus\{0\}$ and write, by definition of $\|\cdot\|_{Y_h^\star}$,

$$\frac{a_h(w_h, v_h)}{\|v_h\|_{Y_h}} = \frac{m_h(v_h)}{\|v_h\|_{Y_h}} \le \|m_h\|_{Y_h^\star}.$$

The proof is completed by taking the supremum over such v_h and using (5). □

We then define the key notion of consistency which, in combination with the inf–sup stability, provides the estimate on $u_h - I_h u$ in the X_h norm.

Definition 8 (*Consistency error and consistency*) Let u be the solution to the continuous problem (3) and take $I_h u \in X_h$. The *variational consistency error* is the linear form $\mathcal{E}_h(u;\cdot) : Y_h \to \mathbb{R}$ defined by

$$\mathcal{E}_h(u;\cdot) = \ell_h(\cdot) - a_h(I_h u, \cdot). \tag{7}$$

Let now a family $(X_h, a_h, \ell_h)_{h\to 0}$ of spaces and forms be given, and consider the corresponding family of discrete problems (4). We say that *consistency* holds if

$$\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star} \to 0 \text{ as } h \to 0.$$

Remark 9 (*Choice of $I_h u$*) No particular property is required here on $I_h u$; it could actually be any element of X_h. However, for the estimates that follow to be meaningful, it is expected that $I_h u$ is computed from u, not necessarily in a linear way but such that information on $I_h u$ encodes meaningful information on u itself.

The first main result of the paper, an estimate on $\|u_h - I_h u\|_{X_h}$, is stated in the following theorem. As explained in the introduction, this theorem can be considered as a "third Strang lemma". In passing, it also shows that (2) holds.

Theorem 10 (Abstract error estimate and convergence in energy norm) *Assume that a_h is inf–sup stable in the sense of Definition 4. Let u be a solution to (3), $I_h u \in X_h$, and recall the definition (7) of the variational consistency error $\mathcal{E}_h(u;\cdot)$. If u_h is a solution to (4) then*

$$\|u_h - I_h u\|_{X_h} \leq \gamma^{-1} \|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star}. \tag{8}$$

As a consequence, letting a family $(X_h, a_h, \ell_h)_{h\to 0}$ of spaces and forms be given, if consistency holds and γ does not depend on h, then we have convergence *in the following sense:*

$$\|u_h - I_h u\|_{X_h} \to 0 \text{ as } h \to 0.$$

Proof For any $v_h \in Y_h$, the scheme (4) yields

$$a_h(u_h - I_h u, v_h) = a_h(u_h, v_h) - a_h(I_h u, v_h) = \ell_h(v_h) - a_h(I_h u, v_h).$$

Recalling the definition of the consistency error, we then infer that the error $u_h - I_h u$ can be characterised as the solution to the following *error equation*:

$$a_h(u_h - I_h u, v_h) = \mathcal{E}_h(u; v_h) \quad \forall v_h \in Y_h. \tag{9}$$

The proof is completed by applying Proposition 7 to $m_h = \mathcal{E}_h(u;\cdot)$ and $w_h = u_h - I_h u$. □

Remark 11 (*Quasi-optimality of the error estimate*) Let

$$\|a_h\|_{X_h\times Y_h} := \sup_{w_h\in X_h\setminus\{0\}, v_h\in Y_h\setminus\{0\}} \frac{|a_h(w_h, v_h)|}{\|w_h\|_{X_h}\|v_h\|_{Y_h}}$$

be the standard norm of the bilinear form a_h. The error equation (9) shows that

$$\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star} \leq \|a_h\|_{X_h\times Y_h} \|u_h - I_h u\|_{X_h}.$$

Hence, if $\|a_h\|_{X_h\times Y_h}$ (and γ, see Remark 5) remains bounded with respect to h as $h \to 0$, which is always the case in practice, the estimate (8) is quasi-optimal in the sense that, for some C not depending on h, it holds that

$$C^{-1} \|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star} \leq \|u_h - I_h u\|_{X_h} \leq C \|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star}.$$

2.3 Improved error estimate in a weaker norm

Assume now that H is continuously embedded in a Banach space L, with norm denoted by $\|\cdot\|_L$, and that there exists a linear reconstruction operator

$$r_h : X_h \to L. \tag{10}$$

If r_h is continuous, with norm bounded above by C, then (8) readily gives

$$\|r_h(u_h - I_h u)\|_L \le C\gamma^{-1} \|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star}. \tag{11}$$

Our aim here is to improve this estimate by using an Aubin–Nitsche trick. To this purpose, we assume that, for all $g \in L^\star$ (the space of continuous linear forms $L \to \mathbb{R}$), there exists a solution to the continuous dual problem:

$$\text{Find } z_g \in H \text{ such that } \quad a(w, z_g) = g(w) \quad \forall w \in H. \tag{12}$$

Definition 12 (*Dual consistency error*) Under Assumption (10), let $g \in L^\star$, z_g be a solution to the dual problem (12), and $J_h z_g \in Y_h$. The dual consistency error of z_g is the linear form $\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot) : X_h \to \mathbb{R}$ defined by

$$\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot) = g \circ r_h - a_h(\cdot, J_h z_g). \tag{13}$$

Theorem 13 (Improved estimate in L-norm) *Assume* (10) *and that the dual problem* (12) *has a solution z_g for any $g \in L^\star$. Let $B_{L^\star} = \{g \in L^\star : \|g\|_{L^\star} \le 1\}$ be the unit ball in $L^\star$. Let u and u_h be the solutions to* (3) *and* (4)*, respectively, and take $I_h u \in X_h$ and, for $g \in B_{L^\star}$, $J_h z_g \in Y_h$. Then,*

$$\|r_h(u_h - I_h u)\|_L \le \|u_h - I_h u\|_{X_h} \sup_{g \in B_{L^\star}} \left\|\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)\right\|_{X_h^\star} + \sup_{g \in B_{L^\star}} \mathcal{E}_h(u; J_h z_g). \tag{14}$$

Remark 14 (*Primal-dual consistency error*) The quantity $\mathcal{E}_h(u; J_h z_g) = \ell_h(J_h z_g) - a_h(I_h u, J_h z_g)$ is a measure of consistency of the discrete primal problem (4) that also involves the solution z_g to the continuous dual problem (12). For this reason, we will call $\mathcal{E}_h(u; J_h z_g)$ the *primal-dual consistency error*.

Proof Let $g \in B_{L^\star}$. By definition (13) of $\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)$, it holds for any $w_h \in X_h$,

$$g(r_h w_h) = \mathcal{E}_h^{\mathrm{d}}(z_g; w_h) + a_h(w_h, J_h z_g).$$

Letting $w_h = u_h - I_h u$ and recalling the error equation (9), this gives

$$g(r_h(u_h - I_h u)) = \mathcal{E}_h^{\mathrm{d}}(z_g; u_h - I_h u) + \mathcal{E}_h(u; J_h z_g).$$

Taking the supremum over $g \in B_{L^\star}$, and recalling that $\sup_{g \in B_{L^\star}} g(w) = \|w\|_L$ for all $w \in L$, we infer

$$\|r_h(u_h - I_h u)\|_L \le \sup_{g \in B_{L^\star}} \mathcal{E}_h^{\mathrm{d}}(z_g; u_h - I_h u) + \sup_{g \in B_{L^\star}} \mathcal{E}_h(u; J_h z_g). \tag{15}$$

To conclude, recall the definition (6) of the dual norm to write

$$\mathcal{E}_h^{\mathrm{d}}(z_g; u_h - I_h u) \le \|u_h - I_h u\|_{X_h} \left\|\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)\right\|_{X_h^\star}.$$

□

Remark 15 (*Alternative L-error bound*) The estimate (15) appears slightly sharper than (14). In the statement of Theorem 13, however, we have preferred a formulation which emphasises a general property of the dual consistency error $\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)$ rather than its evaluation at a specific argument; indeed, unlike $\mathcal{E}_h(u;J_h z_g)$ (see Sect. 2.4.3), it does not seem possible in general to have a better bound on $\mathcal{E}_h^{\mathrm{d}}(z_g;u_h-I_h u)$ than the one provided by $\|u_h-I_h u\|_{X_h}\left\|\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)\right\|_{X_h^\star}$.

2.4 Comments

A few comments are in order.

2.4.1 Recovering continuous estimates

For a number of methods, the interpolation operator I_h is naturally defined as part of the method, and there exists some continuous linear reconstruction operator $R_h : X_h \to E$, where E is a space of functions (which might or might not be a subspace of H).

For example, in conforming FE methods, I_h is the nodal interpolant and R_h the reconstruction of functions in the FE space from their nodal values. In HHO or non-conforming VEM methods, I_h corresponds to L^2 projections on local (face- and cell-) polynomial spaces, and R_h is a local potential reconstruction related to the elliptic projector; alternatively, in the VEM setting, $R_h v_h$ can give the unique (but not explicitly known) function in the VEM space that has the degrees of freedom encoded in the vector v_h.

If the norm of R_h is bounded by C, the energy estimate (8) gives

$$\|R_h u_h - R_h I_h u\|_E \le C\gamma^{-1}\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star}.$$

The triangle inequality then leads to the following continuous estimate between the reconstructed function $R_h u_h$ and the solution u to the continuous problem:

$$\|R_h u_h - u\|_E \le C\gamma^{-1}\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star} + \|R_h I_h u_h - u\|_E.$$

The last term is usually estimated by means of approximation properties of the space X_h and of the operators I_h and R_h attached to the scheme (they do not depend on the continuous equation (3) or its discretisation (4)). Hence, even though Theorem 10 states an estimate in a purely discrete setting, from this a continuous estimate can often be easily recovered.

2.4.2 Link between primal, dual and primal-dual consistency errors

If the discrete bilinear form a_h is symmetric (which requires $X_h = Y_h$) and $\ell_h = \ell \circ r_h$, then the primal and dual consistency errors are identical, and thus estimating $\left\|\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)\right\|_{X_h^\star}$ in (14) does not require any additional work than the one done for estimating $\|\mathcal{E}_h(u;\cdot)\|_{X_h^\star}$ in (8).

Even if a_h is not symmetric or $\ell_h \neq \ell \circ r_h$, the dual problem (12) often has a similar structure as the primal problem, with different parameters; this is expected to be reflected in $\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)$, which might simply be $\mathcal{E}_h(z_g;\cdot)$ with different parameters. In this case, the estimate done on the primal consistency error might directly apply, with easy substitutions, to the dual consistency error. For example, the dual problem to the advection–diffusion–reaction model

$$-\boldsymbol{\nabla}\cdot(\boldsymbol{K}\boldsymbol{\nabla}u)+\boldsymbol{\nabla}\cdot(\boldsymbol{\beta}u)+\mu u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{16}$$

is the same problem with $\boldsymbol{\beta}$ replaced with $-\boldsymbol{\beta}$ and μ replaced with $\mu+\boldsymbol{\nabla}\cdot\boldsymbol{\beta}$.

Estimating $\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star}$ in (8) requires to estimate $\mathcal{E}_h(u;v_h)$ for all $v_h \in Y_h$. Some of the steps performed in this estimate can often be directly used to estimate the primal-dual consistency error $\mathcal{E}_h(u;J_hz)$ in (14). One simply needs to be cautious and draw on the additional information available in this latter term: the primal consistency error is not tested on an arbitrary $v_h \in Y_h$, but on the specific vector J_hz_g; taking advantage of that specificity can lead to improved rates of convergence (see next section). This idea is illustrated in the proof of Theorem 22 below.

We also notice that, when multiple terms are present as in problem (16), the consistency error involves one component per term, and these components can be estimated independently. The benefit is twofold: on the one hand, proceeding this way simplifies the analysis; on the other hand, it makes it possible to re-use the consistency results proved individually for each operator.

2.4.3 Integer and fractional rates of convergence

Under some regularity assumptions on the solution u, it is possible to obtain rates of convergence for the quantity $\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star}$ that appears in the right-hand side of the energy error estimate (8). Typically, for second order elliptic problems, one will assume that $u \in H^r(\Omega)\cap H_0^1(\Omega)$ and establish that

$$\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star} \le Ch^{\omega(r)}\|u\|_{H^r(\Omega)}, \tag{17}$$

where $\omega(r)$ is an appropriate power depending on r and on the considered scheme. This estimate is usually easier to establish for integer r, but once this is done it also holds for fractional r, by basic interpolation result on the mapping $u \mapsto \mathcal{E}_h(u;\cdot) \in Y_h^\star$. The same considerations holds for the dual consistency error.

Let us now examine the improved L-error estimate (14) and consider, for example, elliptic problems. Under optimal elliptic regularity assumptions, it is expected that $z_g \in H^2(\Omega)$. This regularity result will translate into a specific rate of convergence of $\|\mathcal{E}_h^{\mathrm{d}}(z_g;\cdot)\|_{X_h^\star}$, say $\mathcal{O}(h)$. The first term in (14) is then one (or more) orders of magnitude less that $\|u_h - I_hu\|_{X_h}$. The regularity of z_g also translates into constraints on the vector J_hz_g, which cannot vary as freely as any $v_h \in Y_h$; because of that, it is expected that the primal-dual consistency error $\mathcal{E}_h(u;J_hz_g)$ is also one or more orders of magnitude less that $\|\mathcal{E}_h(u;\cdot)\|_{Y_h^\star}$. Hence, the right-hand side of (14) should converge at a faster rate than the right-hand side of (8) as $h \to 0$, showing that

Theorem 13 is indeed an improvement over the basic estimate (11) coming from Theorem 10.

2.4.4 Range of applications

Let us explicitly remark that, even though we only consider, for questions of length, second order elliptic problems in Sect. 3, the framework and estimates described in this section cover a wide range of equations and schemes. For example, elliptic equations of order four, such as the ones encountered in the theory of thin plates, also fit into the setting of Sect. 2.1. Several popular numerical tricks are also covered by the present framework, such as the weak enforcement (*à la* Nitsche) of boundary conditions in the discrete formulation (4).

Finally, we also note that, even though this is the classical example we might have in mind for second order elliptic problems, the space L in Sect. 2.3 does not need to be $L^2(\Omega)$. It could for example be $H^s(\Omega)$ for some $s \in (0, 1)$, leading to optimal rates of convergence in H^s norm instead of L^2 norm.

3 Applications

In this section we showcase applications of the discrete analysis framework to a variety of numerical methods.

3.1 Setting

For the sake of simplicity, we focus on a pure diffusion model problem. Denote by $\Omega \subset \mathbb{R}^d$, $d \geq 1$, an open bounded connected polytopal domain with boundary $\partial\Omega$. In what follows, for any measured set X, we denote by $(\cdot,\cdot)_X$ the usual inner product of $L^2(X)$ or $L^2(X)^d$ according to the context, by $\|\cdot\|_X$ the corresponding norm, and we adopt the convention that the subscript is omitted whenever $X = \Omega$.

Let $\boldsymbol{K} : \Omega \to \mathbb{R}^{d\times d}$ denote a symmetric, uniformly elliptic diffusion field, which we additionally assume piecewise constant on a finite partition $P_\Omega = \{\Omega_i \,:\, 1 \leq i \leq N_\Omega\}$ of Ω into polytopes. For a given source term $f : \Omega \to \mathbb{R}$, our model problem reads: Find $u : \Omega \to \mathbb{R}$ such that

$$-\nabla\cdot(\boldsymbol{K}\nabla u) = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{18}$$

Assuming $f \in L^2(\Omega)$, a weak formulation of this problem is: Find $u \in H^1_0(\Omega)$ such that

$$a_{\boldsymbol{K}}(u, v) := (\boldsymbol{K}\nabla u, \nabla v) = (f, v) \quad \forall v \in H^1_0(\Omega). \tag{19}$$

We denote by $\mathcal{M}_h = (\mathcal{T}_h, \mathcal{F}_h)$ a mesh of the domain, where $\mathcal{T}_h$ collects the mesh elements, or cells, and $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^{\mathrm{b}}$ the hyperplanar mesh faces, with $\mathcal{F}_h^{\mathrm{i}}$ and $\mathcal{F}_h^{\mathrm{b}}$ denoting, respectively, the sets of internal and boundary faces. For any mesh element

$T \in \mathcal{T}_h$, h_T is the diameter of T and $\mathcal{F}_T$ is the set of faces that lie on its boundary ∂T. Symmetrically, for any mesh face $F \in \mathcal{F}_h$, we denote by h_F the diameter of F and by $\mathcal{T}_F$ the set collecting the one (if F is a boundary face) or two (if F is an internal face) mesh elements that share F. For any $T \in \mathcal{T}_h$ and any $F \in \mathcal{F}_T$, $\boldsymbol{n}_{TF}$ is the unit vector normal to F and pointing out of T.

The meshes we consider are always part of a regular family $(\mathcal{M}_h)_{h\in\mathcal{H}}$ in the sense of [24, Definition 3.3]. Unless otherwise specified, the notation $a \lesssim b$ means $a \le Cb$ with constant C possibly depending on the regularity factor of that family, but not depending on $\boldsymbol{K}$ or h and, for local inequalities, on the mesh element or face. Additional regularity assumptions on the meshes depend on the considered method and will be given when necessary. We however always assume that the mesh is compliant with the partition P_Ω, i.e., for all $T \in \mathcal{T}_h$, there exists a unique Ω_i, $1 \le i \le N_\Omega$, such that $T \subset \Omega_i$. For all $T \in \mathcal{T}_h$, we denote by $\boldsymbol{K}_T := \boldsymbol{K}_{|T}$ the constant value of $\boldsymbol{K}$ inside T and we introduce the local anisotropy ratio

$$\alpha_T := \frac{\overline{\lambda}_T}{\underline{\lambda}_T} \tag{20}$$

where $\underline{\lambda}_T$ and $\overline{\lambda}_T$ denote, respectively, the smallest and largest eigenvalues of $\boldsymbol{K}_T$.

For a given integer $\mathsf{l} \ge 0$ and X mesh element or face, we denote by $\mathbb{P}^{\mathsf{l}}(X)$ the space spanned by the restriction to X of d-variate polynomials of total degree $\le \mathsf{l}$. The L^2-projector $\pi_X^{0,\mathsf{l}} : L^2(X) \to \mathbb{P}^{\mathsf{l}}(X)$ is defined by: For all $w \in L^2(X)$, $\pi_X^{0,\mathsf{l}} w$ is the unique element in $\mathbb{P}^{\mathsf{l}}(X)$ such that $(\pi_X^{0,\mathsf{l}} w, q)_X = (w, q)_X$ for all $q \in \mathbb{P}^{\mathsf{l}}(X)$. In the discussion, we will need the following approximation results, which are a special case of [18, Lemmas 3.4 and 3.6]: Let an integer $s \in \{0, \ldots, \mathsf{l}+1\}$ be given. Then, for any mesh element $T \in \mathcal{T}_h$, any function $v \in H^s(T)$, and any exponent $m \in \{0, \ldots, s\}$, it holds that

$$\left|v - \pi_T^{0,\mathsf{l}} v\right|_{H^m(T)} \lesssim h_T^{s-m} \, |v|_{H^s(T)} \,. \tag{21}$$

Moreover, if $s \ge 1$ and $m \le s - 1$,

$$\left|v - \pi_T^{0,\mathsf{l}} v\right|_{H^m(\mathcal{F}_T)} \lesssim h_T^{s-m-\frac{1}{2}} \, |v|_{H^s(T)} \,, \tag{22}$$

where $H^m(\mathcal{F}_T) := \left\{ v \in L^2(\partial T) \,:\, v_{|F} \in H^m(F) \text{ for all } F \in \mathcal{F}_T \right\}$ is the broken Sobolev space on $\mathcal{F}_T$ and $|\cdot|_{H^m(\mathcal{F}_T)}$ the corresponding broken seminorm.

The space of broken polynomials of total degree $\le \mathsf{l}$ on $\mathcal{T}_h$ is denoted by $\mathbb{P}^{\mathsf{l}}(\mathcal{T}_h)$, i.e.,

$$\mathbb{P}^{\mathsf{l}}(\mathcal{T}_h) := \left\{ v \in L^2(\Omega) \,:\, v_{|T} \in \mathbb{P}^{\mathsf{l}}(T) \quad \forall T \in \mathcal{T}_h \right\}.$$

For a given exponent $s \in \mathbb{N}$, we define the broken Sobolev space

$$H^s(\mathcal{T}_h) := \left\{ v \in L^2(\Omega) : v_{|T} \in H^s(T) \quad \forall T \in \mathcal{T}_h \right\}.$$

On $H^1(\mathcal{T}_h)$ we define the broken gradient operator ∇_h such that, for all $v \in H^1(\mathcal{T}_h)$, $(\nabla_h v)_{|T} := \nabla v_{|T}$.

3.2 Virtual element methods

The first application we consider is to VEM. The main novelty of this section is the derivation of a unified energy error estimate for both conforming and non-conforming VEM where the dependence on the diffusion field is accurately tracked. Our results show full robustness with respect to the heterogeneity of the diffusion field, and a mild dependence on the square root of the local anisotropy ratio.

3.2.1 The oblique elliptic projector

Like several other arbitrary-order discretisation methods for problem (19) (such as HHO [23], Weak Galerkin [46] methods, and Mimetic Finite Differences [40]), VEM are based on local projectors which possibly embed a dependence on the diffusion field inside T. Let $k \geq 1$ be a natural number. Fixing $T \in \mathcal{T}_h$, we focus here on VEM formulations based on the (oblique) elliptic projector $\pi_{\boldsymbol{K},T}^{1,k} : H^1(T) \to \mathbb{P}^k(T)$ defined by: For $v \in H^1(T)$,

$$(\boldsymbol{K}_T \nabla \pi_{\boldsymbol{K},T}^{1,k} v, \nabla w)_T = (\boldsymbol{K}_T \nabla v, \nabla w)_T \quad \forall w \in \mathbb{P}^k(T), \tag{23a}$$

$$\int_T \pi_{\boldsymbol{K},T}^{1,k} v = \int_T v. \tag{23b}$$

By the Riesz representation theorem in $\nabla\mathbb{P}^k(T)$, (23a) defines a unique element of $\nabla\mathbb{P}^k(T)$, and the closure equation (23b) fixes the corresponding $\pi_{\boldsymbol{K},T}^{1,k} v \in \mathbb{P}^k(T)$. It can be easily checked that $\pi_{\boldsymbol{K},T}^{1,k}$ is a projector, i.e., it is linear and idempotent. As a result, it maps polynomials of total degree $\leq k$ onto themselves. Optimal approximation properties for $\pi_{\boldsymbol{K},T}^{1,k}$ in diffusion-dependent seminorms are studied in the following theorem, where the dependence of the multiplicative constants on the local diffusion tensor $\boldsymbol{K}_T$ is carefully tracked.

Theorem 16 (Approximation properties of the oblique elliptic projector in diffusion-weighted seminorms) *Assume the setting described in Sect. 3.1. For a given polynomial degree $k \geq 0$, let an integer $s \in \{1, \ldots, k+1\}$ be given. Then, recalling the definition (23) of the oblique elliptic projector, for all $v \in H^s(T)$ and all $m \in \{0, \ldots, s-1\}$,*

$$\left| \boldsymbol{K}_T^{\frac{1}{2}} \nabla (v - \pi_{\boldsymbol{K},T}^{1,k} v) \right|_{H^m(T)^d} \lesssim \overline{\lambda}_T^{\frac{1}{2}} h_T^{s-m-1} |v|_{H^s(T)}. \tag{24}$$

If, additionally, $m \leq s-2$ (which enforces $s \geq 2$), then

$$h_T^{\frac{1}{2}} \left| \boldsymbol{K}_T^{\frac{1}{2}} \nabla (v - \pi_{\boldsymbol{K},T}^{1,k} v) \right|_{H^m(\mathcal{F}_T)^d} \lesssim \overline{\lambda}_T^{\frac{1}{2}} h_T^{s-m-1} |v|_{H^s(T)}, \tag{25}$$

where $H^m(\mathcal{F}_T)^d$ is the broken Sobolev space on $\mathcal{F}_T$ defined component-wise as in (22), and $|\cdot|_{H^m(\mathcal{F}_T)^d}$ is the corresponding seminorm.

Proof We consider the following representation of v:

$$v = Q^s v + R^s v, \tag{26}$$

where $Q^s v \in \mathbb{P}^{s-1}(T) \subset \mathbb{P}^k(T)$ is the averaged Taylor polynomial, while the remainder $R^s v$ satisfies, for all $r \in \{0, \ldots, s\}$ (cf. [12, Lemma 4.3.8]),

$$\left|R^s v\right|_{H^r(T)} \lesssim h_T^{s-r} \, |v|_{H^s(T)} \,. \tag{27}$$

We next notice that, using the definition (23) of the oblique elliptic projector, it holds for any $\phi \in H^1(T)$,

$$\left\| \boldsymbol{K}_T^{\frac{1}{2}} \nabla \pi_{\boldsymbol{K},T}^{1,k} \phi \right\|_T \leq \left\| \boldsymbol{K}_T^{\frac{1}{2}} \nabla \phi \right\|_T , \tag{28}$$

as can be inferred selecting $w = \pi_{\boldsymbol{K},T}^{1,k} \phi$ as a function test in (23a) and using the Cauchy–Schwarz inequality. Taking the projection of (26), and using the fact that $\pi_{\boldsymbol{K},T}^{1,k}$ maps polynomials of total degree $\leq k$ onto themselves to write $\pi_{\boldsymbol{K},T}^{1,k} Q^s v = Q^s v$, it is inferred that $\pi_{\boldsymbol{K},T}^{1,k} v = Q^s v + \pi_{\boldsymbol{K},T}^{1,k}(R^s v)$. Subtracting this equation from (26), we obtain $v - \pi_{\boldsymbol{K},T}^{1,k} v = R^s v - \pi_{\boldsymbol{K},T}^{1,k}(R^s v)$. Applying the operator $\boldsymbol{K}_T^{1/2}\nabla$ to this expression, passing to the seminorm, and using the triangle inequality, we arrive at

$$\begin{aligned} \left| \boldsymbol{K}_T^{\frac{1}{2}} \nabla (v - \pi_{\boldsymbol{K},T}^{1,k} v) \right|_{H^m(T)^d} &\leq \left| \boldsymbol{K}_T^{\frac{1}{2}} \nabla R^s v \right|_{H^m(T)^d} + \left| \boldsymbol{K}_T^{\frac{1}{2}} \nabla \pi_{\boldsymbol{K},T}^{1,k} (R^s v) \right|_{H^m(T)^d} \\ &=: \mathfrak{T}_1 + \mathfrak{T}_2. \end{aligned} \tag{29}$$

For the first term, it is readily inferred that $\mathfrak{T}_1 \lesssim \underline{\lambda}_T^{\frac{1}{2}} |R^s v|_{H^{m+1}(T)}$ which, combined with (27) for $r = m + 1$, gives

$$\mathfrak{T}_1 \lesssim \underline{\lambda}_T^{\frac{1}{2}} h_T^{s-m-1} \, |v|_{H^s(T)} \,. \tag{30}$$

For the second term, on the other hand, we can proceed as follows:

$$\begin{aligned} \mathfrak{T}_2 &\lesssim h_T^{-m} \left\| \boldsymbol{K}_T^{\frac{1}{2}} \nabla (\pi_{\boldsymbol{K},T}^{1,k} R^s v) \right\|_T \lesssim h_T^{-m} \left\| \boldsymbol{K}_T^{\frac{1}{2}} \nabla R^s v \right\|_T \\ &\lesssim \underline{\lambda}_T^{\frac{1}{2}} h_T^{-m} \left| R^s v \right|_{H^1(T)} \lesssim \underline{\lambda}_T^{\frac{1}{2}} h_T^{s-m-1} \, |v|_{H^s(T)} \,, \end{aligned} \tag{31}$$

where we have used the local inverse Sobolev embeddings of [18, Remark A.2] in the first bound, (28) with $\phi = R^s v$ in the second bound, the definition of the H^1-seminorm in the third bound, and (27) with $r = 1$ to conclude. Plugging (30) and (31) into (29),

(24) follows. To prove (25), it suffices to combine (24) with a local continuous trace inequality (see [22, Lemma 1.49], where a slightly different notion of face is used which, however, does not affect the final result). □

Remark 17 (*Case of varying $\boldsymbol{K}_T$*) Theorem 16 also holds for a diffusion that varies inside T, provided that $\boldsymbol{K}_T^{1/2}$ belongs to $\mathbb{P}^r(T)^{d\times d}$ for some integer r. Under this assumption, and letting $\overline{\lambda}_T$ be the maximum over $\mathbf{x}\in T$ of the largest eigenvalue of $\boldsymbol{K}_T(\mathbf{x})$, (30) remains valid owing to (27) and inverse inequalities (that show that $\left|\boldsymbol{K}_T^{1/2}\right|_{W^{s,\infty}(T)}\lesssim h_T^{-s}\left\|\boldsymbol{K}_T^{1/2}\right\|_{L^\infty(T)}\lesssim h_T^{-s}\overline{\lambda}_T^{\frac12}$), and the local inverse Sobolev embeddings of [18, Remark A.2] can be invoked to establish (31). In this case, the hidden multiplicative constants in (24)–(25) additionally depend on r.

If $\boldsymbol{K}_T^{1/2}$ is not in $\mathbb{P}^r(T)^{d\times d}$ then, following the proof above, the right-hand sides of (24) and (25) have to be multiplied by $\left\|\boldsymbol{K}_T^{1/2}\right\|_{W^{m,\infty}(T)^{d\times d}}$.

3.2.2 An abstract virtual element method

To define a VEM scheme, one first needs to choose a finite dimensional subspace $V_h^k\subset H^1(\mathcal{T}_h)$ that locally contains polynomials and satisfies some continuity requirements:

$$\mathbb{P}^k(T)\subset V_T^k:=\left\{(v_h)_{|T}\ :\ v_h\in V_h^k\right\}\quad\forall T\in\mathcal{T}_h, \tag{32}$$

and, for all $v_h\in V_h^k$,

$$\pi_F^{0,k-1}(v_h)_{|T}+\pi_F^{0,k-1}(v_h)_{|T'}=0\quad\forall F\in\mathcal{F}_h^{\mathrm{i}}\text{ with }\mathcal{T}_F=\{T,T'\}. \tag{33}$$

A subspace $X_h=V_{h,0}^k$ of V_h^k is then considered to account for the homogeneous Dirichlet boundary conditions, and such that (at least) the following condition holds: For all $v_h\in V_{h,0}^k$,

$$\pi_F^{0,k-1}v_h=0\quad\forall F\in\mathcal{F}_h^{\mathrm{b}}. \tag{34}$$

Different choices of spaces lead to different methods, such as conforming [7] or nonconforming [5] VEM. Our analysis here does not require a complete description of the space $V_{h,0}^k$. We merely need the two following properties of the interpolant $I_h: H_0^1(\Omega)\cap C(\overline{\Omega})\to V_{h,0}^k$:

(I1) *Locality and boundedness.* For all $T\in\mathcal{T}_h$, there is a linear mapping $I_T: H^1(T)\cap C(\overline{T})\to V_T^k$ such that $(I_hw)_{|T}=I_T(w_{|T})$ for all $w\in H_0^1(\Omega)\cap C(\overline{\Omega})$, and

$$\left\|\boldsymbol{K}_T^{\frac12}\nabla(I_T\phi)\right\|_T\lesssim\overline{\lambda}_T^{\frac12}\|\nabla\phi\|_T\quad\forall\phi\in H^1(T)\cap C(\overline{T}). \tag{35}$$

(I2) *Preservation of polynomials.* For all $T\in\mathcal{T}_h$ and $v\in\mathbb{P}^k(T)$, $I_Tv=v$.

Remark 18 (*On the DOFs for VEM*) The degrees of freedom (DOFs) of the VEM, that is the unisolvent family of linear forms $(\lambda_i)_{i\in I}$ on $V_{h,0}^k$, must be chosen to enable the computation of the oblique elliptic projector (23) for functions in $V_{h,0}^k$. Given (23b), this means, in particular, that these DOFs should enable the computation, for all $T \in \mathcal{T}_h$, of $\pi_T^{0,0}$ on $V_{h,0}^k$. In the case $k = 1$, we therefore implicitly consider enhanced VEM spaces [13]. Otherwise, the closure equation (23b) should be modified, see e.g. [11].

Let $V_{h,0}^k$ be a chosen VEM space, and

$$l = 0 \text{ if } k = 1, \quad l \in \{0, 1\} \text{ if } k = 2, \quad l = k - 2 \text{ if } k \geq 3.$$

We assume that the DOFs of $V_{h,0}^k$ enable the computation of $(\pi_T^{0,l})_{T\in\mathcal{T}_h}$ on $V_{h,0}^k$ (in the cases $k = 1$ or $(k, l) = (2, 1)$, this supposes using enhanced spaces). A VEM scheme for (19) is then obtained by writing (4) with

$$\ell_h(v_h) = \sum_{T\in\mathcal{T}_h} \left(f, \pi_T^{0,l} v_h\right)_T \quad \forall v_h \in V_{h,0}^k \tag{36}$$

and

$$\begin{aligned} &a_h(v_h, w_h) = \sum_{T\in\mathcal{T}_h} a_T(v_h, w_h) \quad \forall v_h, w_h \in V_{h,0}^k, \\ &\text{where} \quad a_T(v_h, w_h) = \left(\boldsymbol{K}_T \nabla \pi_{\boldsymbol{K},T}^{1,k} v_h, \nabla \pi_{\boldsymbol{K},T}^{1,k} w_h\right)_T \\ &\quad + s_T\left((I - \pi_{\boldsymbol{K},T}^{1,k}) v_h, (I - \pi_{\boldsymbol{K},T}^{1,k}) w_h\right). \end{aligned} \tag{37}$$

Here, s_T is a symmetric positive semi-definite bilinear form on V_T^k (computable from the degrees of freedom) such that

$$(\boldsymbol{K}_T \nabla w, \nabla w)_T \lesssim s_T(w, w) \lesssim (\boldsymbol{K}_T \nabla w, \nabla w)_T \quad \forall w \in V_T^k \text{ such that } \pi_{\boldsymbol{K},T}^{1,k} w = 0. \tag{38}$$

The definition of a_h implies that if $a_h(v_h, v_h) = 0$ then v_h is constant in each cell, and (33)–(34) then show that $v_h = 0$. Hence, a_h is symmetric positive definite on $V_{h,0}^k$. The norm considered on $X_h = V_{h,0}^k$ is the one induced by a_h, that is,

$$\|v_h\|_{X_h} := \sqrt{a_h(v_h, v_h)} \quad \forall v_h \in X_h. \tag{39}$$

3.2.3 Error estimate in energy norm

With this setting in place, Theorem 10 yields the following estimates.

Theorem 19 (Energy estimates for VEM schemes) *Let $1 \le r \le k$ and assume that the solution $u \in H_0^1(\Omega) \cap C(\overline{\Omega})$ to (19) belongs to $H^{r+1}(\mathcal{T}_h)$. Let u_h be the solution of the VEM scheme (that is, (4) with the choices (36) and (37)). Then the following estimates hold:*

$$\|u_h - I_h u\|_{X_h} \lesssim \left(\sum_{T \in \mathcal{T}_h} \alpha_T \overline{\lambda}_T h_T^{2r} \, |u|^2_{H^{r+1}(T)} \right)^{\frac{1}{2}} \tag{40}$$

and

$$\left\| \boldsymbol{K}^{\frac{1}{2}} (\nabla_h \pi^{1,k}_{\boldsymbol{K},h} u_h - \nabla u) \right\| \lesssim \left(\sum_{T \in \mathcal{T}_h} \alpha_T \overline{\lambda}_T h_T^{2r} \, |u|^2_{H^{r+1}(T)} \right)^{\frac{1}{2}}, \tag{41}$$

where $\pi^{1,k}_{\boldsymbol{K},h}$ is the patched elliptic projector such that, for all $T \in \mathcal{T}_h$ and $w \in H^1(\mathcal{T}_h)$, $(\pi^{1,k}_{\boldsymbol{K},h} w)_{|T} := \pi^{1,k}_{\boldsymbol{K},T}(w_{|T})$.

Remark 20 (*Diffusion varying in each cell*) If $\boldsymbol{K}$ is not piecewise-constant in the cells, the construction of VEM methods has to be adjusted by using the L^2-orthogonal projection of the gradient of virtual functions; see, e.g., [8,13]. In the context of HHO methods (strongly related to non-conforming VEM), similar ideas have been used in [18] for the approximation of fully non-linear models; see also the discussion in [20, Section 4]. A different approach in the context of HHO methods, valid for linear problems, consists in incorporating the variable diffusion coefficient into the local reconstruction; see [25].

It is worth noting that Theorem 19 remains of interest even for a piecewise constant diffusion tensor. Even though the $H^{r+1}(\mathcal{T}_h)$ regularity of the solution cannot always be ascertained if $\boldsymbol{K}$ is discontinuous (counter-examples to the $H^2(\mathcal{T}_h)$ regularity can be constructed for some piecewise-constant diffusions and smooth source terms [45]), one can easily find many situations in which $\boldsymbol{K}$ is piecewise-constant and the solution belongs to $H^{r+1}(\mathcal{T}_h)$, situations for which Theorem 19 yields a meaningful estimate. On the contrary, an estimate based on the $H^{r+1}(\Omega)$-norm would essentially impose a smooth $\boldsymbol{K}$ over the entire domain—as any discontinuity of the diffusion essentially prevents the solution from being in $H^2(\Omega)$ or more regular.

Proof By choice of the norm on $V^k_{h,0}$, the bilinear form a_h is coercive with constant 1. Hence, (40) follows from (8) (with $\gamma = 1$) if we estimate the norm of the consistency error appropriately. Since $f = -\nabla\cdot(\boldsymbol{K}\nabla u)$ we have, for all $v_h \in V^k_{h,0}$,

$$\begin{aligned} \mathcal{E}_h(u; v_h) = & \sum_{T \in \mathcal{T}_h} (-\nabla\cdot(\boldsymbol{K}_T \nabla u), \pi^{0,l}_T v_h)_T - \sum_{T \in \mathcal{T}_h} (\boldsymbol{K}_T \nabla \pi^{1,k}_{\boldsymbol{K},T} I_h u, \nabla \pi^{1,k}_{\boldsymbol{K},T} v_h)_T \\ & - \sum_{T \in \mathcal{T}_h} s_T \left((I - \pi^{1,k}_{\boldsymbol{K},T}) I_h u, (I - \pi^{1,k}_{\boldsymbol{K},T}) v_h \right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{T\in\mathcal{T}_h} (-\nabla\cdot(\boldsymbol{K}_T\nabla u), \pi_T^{0,l} v_h)_T - \sum_{T\in\mathcal{T}_h} (\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u, \nabla\pi_{\boldsymbol{K},T}^{1,k} v_h)_T \\
&\quad - \sum_{T\in\mathcal{T}_h} (\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k}(I_h u - u), \nabla\pi_{\boldsymbol{K},T}^{1,k} v_h)_T \\
&\quad - \sum_{T\in\mathcal{T}_h} s_T\left((I-\pi_{\boldsymbol{K},T}^{1,k}) I_h u, (I-\pi_{\boldsymbol{K},T}^{1,k}) v_h\right) \\
&=: \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3 + \mathfrak{T}_4. \qquad (42)
\end{aligned}
$$

(i) *Term* $\mathfrak{T}_1 + \mathfrak{T}_2$. Performing element-wise integrations-by-parts, we write

$$
\begin{aligned}
\mathfrak{T}_1 &= \sum_{T\in\mathcal{T}_h} (\boldsymbol{K}_T\nabla u, \nabla\pi_T^{0,l} v_h)_T - \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} (\boldsymbol{K}_T\nabla u\cdot\boldsymbol{n}_{TF}, \pi_T^{0,l} v_h)_F \\
&= \sum_{T\in\mathcal{T}_h} (\boldsymbol{K}_T\nabla u, \nabla\pi_T^{0,l} v_h)_T - \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} (\boldsymbol{K}_T\nabla u\cdot\boldsymbol{n}_{TF}, \pi_T^{0,l} v_h - \pi_F^{0,k-1} v_h)_F, \qquad (43)
\end{aligned}
$$

where the introduction of the term $\pi_F^{0,k-1} v_h$ is justified by the conservativity property $\boldsymbol{K}_T\nabla u\cdot\boldsymbol{n}_{TF} + \boldsymbol{K}_{T'}\nabla u\cdot\boldsymbol{n}_{T'F} = 0$ for all $F \in \mathcal{F}_h^{\mathrm{i}}$ with $\mathcal{T}_F = \{T, T'\}$, and the continuity property and boundary conditions expressed by (33)–(34). Setting $\mathfrak{T}_2 = \sum_{T\in\mathcal{T}_h} \mathfrak{T}_{2,T}$, we write

$$
\begin{aligned}
\mathfrak{T}_{2,T} &= -(\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u, \nabla v_h)_T \\
&= (\nabla\cdot(\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u), \pi_T^{0,l} v_h)_T - \sum_{F\in\mathcal{F}_T} (\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u\cdot\boldsymbol{n}_{TF}, \pi_F^{0,k-1} v_h)_F \\
&= -(\boldsymbol{K}_T\nabla u, \nabla\pi_T^{0,l} v_h)_T - \sum_{F\in\mathcal{F}_T} (\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u\cdot\boldsymbol{n}_{TF}, \pi_F^{0,k-1} v_h - \pi_T^{0,l} v_h)_F, \qquad (44)
\end{aligned}
$$

where the first line follows from the definition (23a) of the oblique elliptic projector with $v = v_h$ and $w = \pi_{\boldsymbol{K},T}^{1,k} u$, the second line is obtained by performing an integration-by-parts and using $\nabla\cdot(\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u) \in \mathbb{P}^l(T)$ (since $l \geq k-2$) and $\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u\cdot\boldsymbol{n}_{TF} \in \mathbb{P}^{k-1}(F)$ to replace v_h by its projections on local element- and face-polynomial spaces, and the third line is a consequence of another integration-by-parts and of the definition (23a) of the oblique elliptic projector which, applied to $v = u$ and $w = \pi_T^{0,l} v_h$ (note that $l \leq k$), gives $(\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} u, \nabla\pi_T^{0,l} v_h)_T = (\boldsymbol{K}_T\nabla u, \nabla\pi_T^{0,l} v_h)_T$. Hence, summing (44) over $T \in \mathcal{T}_h$ and gathering with (43) yields

$$
\mathfrak{T}_1 + \mathfrak{T}_2 = \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} (\boldsymbol{K}_T(\nabla\pi_{\boldsymbol{K},T}^{1,k} u - \nabla u)\cdot\boldsymbol{n}_{TF}, \pi_T^{0,l} v_h - \pi_F^{0,k-1} v_h)_F.
$$

We then estimate $\mathfrak{T}_1+\mathfrak{T}_2$ by using the Cauchy–Schwarz inequality:

$$|\mathfrak{T}_1+\mathfrak{T}_2| \le \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\left\|\boldsymbol{K}_T^{\frac12}(\nabla\pi_{\boldsymbol{K},T}^{1,k}u-\nabla u)\right\|_F \overline{\lambda}_T^{\frac12}\left\|\pi_F^{0,k-1}(\pi_T^{0,l}v_h-(v_h)_{|T})\right\|_F \tag{45}$$

$$\lesssim \sum_{T\in\mathcal{T}_h}\overline{\lambda}_T^{\frac12}h_T^{r-\frac12}\,|u|_{H^{r+1}(T)}\,\overline{\lambda}_T^{\frac12}\left\|\pi_T^{0,l}v_h-(v_h)_{|T}\right\|_F$$

$$\lesssim \sum_{T\in\mathcal{T}_h}\overline{\lambda}_T^{\frac12}h_T^{r-\frac12}\,|u|_{H^{r+1}(T)}\,\overline{\lambda}_T^{\frac12}h_T^{\frac12}\,\|\nabla v_h\|_T\,, \tag{46}$$

where we have used $l\le k-1$ together with the linearity and idempotency of $\pi_F^{0,k-1}$ in the first line to write $\pi_T^{0,l}v_h-\pi_F^{0,k-1}v_h=\pi_T^{0,l}v_h-\pi_F^{0,k-1}(v_h)_{|T}=\pi_F^{0,k-1}(\pi_T^{0,l}v_h-(v_h)_{|T})$, we passed to the second line by using the trace approximation properties (25) of $\pi_{\boldsymbol{K},T}^{1,k}$ (with $s=r+1$ and $m=0$) and the $L^2(F)$-boundedness property of $\pi_F^{0,k-1}$, and we concluded by invoking the trace approximation property (22) of $\pi_T^{0,l}$ with $m=0$ and $s=1$. Recalling the definition (20) of α_T, we have

$$\overline{\lambda}_T^{\frac12}\,\|\nabla v_h\|_T \le \overline{\lambda}_T^{\frac12}\underline{\lambda}_T^{-\frac12}\left\|\boldsymbol{K}_T^{\frac12}\nabla v_h\right\|_T = \alpha_T^{\frac12}\left\|\boldsymbol{K}_T^{\frac12}\nabla v_h\right\|_T \lesssim \alpha_T^{\frac12}a_T(v_h,v_h)^{\frac12}, \tag{47}$$

where the last inequality is obtained introducing $\pm\boldsymbol{K}_T^{\frac12}\nabla\pi_{\boldsymbol{K},T}^{1,k}v_h$ into the norm, using the triangle inequality, invoking the property (38) of s_T with $w_h=v_h-\pi_{\boldsymbol{K},T}^{1,k}v_h$, and recalling the definition (37) of a_T. Thus, using a Cauchy–Schwarz inequality on the sum over $T\in\mathcal{T}_h$ and recalling the definition (39) of $\|\cdot\|_{X_h}$, we conclude that

$$|\mathfrak{T}_1+\mathfrak{T}_2| \lesssim \left(\sum_{T\in\mathcal{T}_h}\alpha_T\overline{\lambda}_T h_T^{2r}\,|u|^2_{H^{r+1}(T)}\right)^{\frac12}\|v_h\|_{X_h}\,. \tag{48}$$

(ii) *Term* $\mathfrak{T}_3$. Apply (35) with $\phi=u-\pi_{\boldsymbol{K},T}^{1,k}u$, which satisfies $I_T\phi=I_Tu-\pi_{\boldsymbol{K},T}^{1,k}u$ by the linearity of I_T together with **(I2)**, to write

$$\left\|\boldsymbol{K}_T^{\frac12}\nabla(I_Tu-\pi_{\boldsymbol{K},T}^{1,k}u)\right\|_T \lesssim \overline{\lambda}_T^{\frac12}\left\|\nabla(u-\pi_{\boldsymbol{K},T}^{1,k}u)\right\|_T. \tag{49}$$

A triangle inequality (introducing $\pm\boldsymbol{K}_T^{\frac12}\nabla\pi_{\boldsymbol{K},T}^{1,k}u$ into the left-hand side) followed by (49), the definition (20) of α_T, and the approximation properties (24) of $\pi_{\boldsymbol{K},T}^{1,k}$ with $s=r+1$ and $m=0$ yield

$$\left\|\boldsymbol{K}_T^{\frac12}\nabla(I_Tu-u)\right\|_T \le \left\|\boldsymbol{K}_T^{\frac12}\nabla(I_Tu-\pi_{\boldsymbol{K},T}^{1,k}u)\right\|_T + \left\|\boldsymbol{K}_T^{\frac12}\nabla(\pi_{\boldsymbol{K},T}^{1,k}u-u)\right\|_T$$

$$\lesssim \overline{\lambda}_T^{\frac12} \left\| \nabla(u - \pi_{\boldsymbol{K},T}^{1,k} u) \right\|_T \lesssim \alpha_T^{\frac12} \left\| \boldsymbol{K}_T^{\frac12} \nabla(u - \pi_{\boldsymbol{K},T}^{1,k} u) \right\|_T$$
$$\lesssim \alpha_T^{\frac12} \overline{\lambda}_T^{\frac12} h_T^r \, |u|_{H^{r+1}(T)} . \tag{50}$$

Applying the boundedness property (28) of $\pi_{\boldsymbol{K},T}^{1,k}$ to $\phi = I_T u - u$ and using (50) then leads to

$$\left\| \boldsymbol{K}_T^{\frac12} \nabla \pi_{\boldsymbol{K},T}^{1,k} (I_T u - u) \right\|_T \lesssim \alpha_T^{\frac12} \overline{\lambda}_T^{\frac12} h_T^r \, |u|_{H^{r+1}(T)} . \tag{51}$$

Using the Cauchy–Schwarz inequality, (51) along with (28) for $\phi = v_h$, again a Cauchy–Schwarz inequality this time on the sum over $T \in \mathcal{T}_h$, and (47) followed by the definition (39) of the norm $\|\cdot\|_{X_h}$, we finally infer for the third term

$$|\mathfrak{T}_3| \le \sum_{T\in\mathcal{T}_h} \left\| \boldsymbol{K}_T^{\frac12} \nabla \pi_{\boldsymbol{K},T}^{1,k} (I_h u - u) \right\|_T \left\| \boldsymbol{K}_T^{\frac12} \nabla \pi_{\boldsymbol{K},T}^{1,k} v_h \right\|_T$$
$$\lesssim \sum_{T\in\mathcal{T}_h} \alpha_T^{\frac12} \overline{\lambda}_T^{\frac12} h_T^r \, |u|_{H^{r+1}(T)} \left\| \boldsymbol{K}_T^{\frac12} \nabla v_h \right\|_T \lesssim \left(\sum_{T\in\mathcal{T}_h} \alpha_T \overline{\lambda}_T h_T^{2r} \, |u|_{H^{r+1}(T)}^2 \right)^{\frac12} \|v_h\|_{X_h} . \tag{52}$$

(iii) *Term* $\mathfrak{T}_4$. We have

$$\left| s_T\Big((I - \pi_{\boldsymbol{K},T}^{1,k}) I_h u, (I - \pi_{\boldsymbol{K},T}^{1,k}) v_h \Big) \right|$$
$$\lesssim s_T\Big((I - \pi_{\boldsymbol{K},T}^{1,k}) I_h u, (I - \pi_{\boldsymbol{K},T}^{1,k}) I_h u \Big)^{\frac12} s_T\Big((I - \pi_{\boldsymbol{K},T}^{1,k}) v_h, (I - \pi_{\boldsymbol{K},T}^{1,k}) v_h \Big)^{\frac12}$$
$$\lesssim \left\| \boldsymbol{K}_T^{\frac12} (\nabla (I - \pi_{\boldsymbol{K},T}^{1,k}) I_h u) \right\|_T \left\| \boldsymbol{K}_T^{\frac12} \nabla (v_h - \pi_{\boldsymbol{K},T}^{1,k} v_h) \right\|_T \tag{53}$$
$$\lesssim \left\| \boldsymbol{K}_T^{\frac12} (\nabla (I - \pi_{\boldsymbol{K},T}^{1,k}) I_h u) \right\|_T \left\| \boldsymbol{K}_T^{\frac12} \nabla v_h \right\|_T , \tag{54}$$

where the first line follows from a Cauchy–Schwarz inequality, the second line is a consequence of (38), and the third line is obtaind using the boundedness property (28) of $\pi_{\boldsymbol{K},T}^{1,k}$. Introducing $\pm \boldsymbol{K}_T^{\frac12} \nabla (u - \pi_{\boldsymbol{K},T}^{1,k} u)$ into the norm and using triangle inequalities, the first factor in the right-hand side of (54) is estimated by

$$\left\| \boldsymbol{K}_T^{\frac12} \nabla (I_T u - \pi_{\boldsymbol{K},T}^{1,k} I_T u) \right\|_T \le \left\| \boldsymbol{K}_T^{\frac12} \nabla (I_T u - u) \right\|_T + \left\| \boldsymbol{K}_T^{\frac12} \nabla (u - \pi_{\boldsymbol{K},T}^{1,k} u) \right\|_T$$
$$+ \left\| \boldsymbol{K}_T^{\frac12} \nabla \pi_{\boldsymbol{K},T}^{1,k} (u - I_T u) \right\|_T$$
$$\lesssim \alpha_T^{\frac12} \overline{\lambda}_T^{\frac12} h_T^r \, |u|_{H^{r+1}(T)} , \tag{55}$$

where the conclusion follows from (50), the approximation properties (24) of $\pi_{\boldsymbol{K},T}^{1,k}$, and (51). Plugging this estimate into (54), summing over $T \in \mathcal{T}_h$, using a Cauchy–Schwarz inequality on the sum over $T \in \mathcal{T}_h$, and invoking (47) together with the definition (39) of $\|\cdot\|_{X_h}$, this yields

$$|\mathfrak{T}_4| \lesssim \left(\sum_{T \in \mathcal{T}_h} \alpha_T \overline{\lambda}_T h_T^{2r} \, |u|^2_{H^{r+1}(T)} \right)^{\frac{1}{2}} \|v_h\|_{X_h} . \tag{56}$$

(iv) *Conclusion*. Plugging (48), (52), and (56) into (42) shows that $\|\mathcal{E}_h(u;\cdot)\|_{X_h^\star}$ is bounded (up to a multiplicative constant) by the right-hand side of (40), which concludes the proof of this inequality.

To establish (41), we notice that, by definitions (39) of the norm on $V_{h,0}^k$ and (37) of a_h,

$$\left\| \boldsymbol{K}^{\frac{1}{2}} (\nabla_h \pi_{\boldsymbol{K},h}^{1,k} u_h - \nabla_h \pi_{\boldsymbol{K},h}^{1,k} I_h u) \right\| \leq \|u_h - I_h u\|_{X_h} \lesssim \left(\sum_{T \in \mathcal{T}_h} \alpha_T \overline{\lambda}_T h_T^{2r} \, |u|^2_{H^{r+1}(T)} \right)^{\frac{1}{2}} .$$

The estimate (41) follows by introducing $\pm \boldsymbol{K}^{\frac{1}{2}} \nabla_h (\pi_{\boldsymbol{K},h}^{1,k} I_h u - \pi_{\boldsymbol{K},h}^{1,k} u)$ in its left-hand side, and by invoking (51) and the optimal approximation properties (24) of the oblique elliptic projector, in a similar way as in (55). □

Remark 21 (*Unified analysis of conforming and non-conforming VEM*) A unified analysis of conforming and non-conforming VEM based on an adaptation of the second Strang lemma has been recently proposed in [13] in the context of more general second-order elliptic problems.

A first difference with the present work is that, therein, the error is measured as $u - u_h$, the difference between the continuous and the virtual solutions. Thus, compared to Theorem 10, several additional terms have to be estimated in order to deduce an order of convergence from [13, Theorem 2]. These measure, in an appropriate way: the approximation properties of the virtual space $V_{h,0}^k$, those of the broken polynomial space $\mathbb{P}^k(\mathcal{T}_h)$, and the nonconformity of the method.

A second difference with respect to the present work is that the dependence of the constants on the problem data is not specifically tracked. In the context of HHO methods, error estimates robust with respect to the problem data for second-order elliptic problems similar to the ones considered in [13] have been recently proposed in [19]; for a study of the links between HHO and non-conforming VEM we refer the reader to [10,17,20].

3.2.4 Improved error estimate in the L^2 norm

Theorem 22 (L^2 estimates for VEM schemes) *Under the hypotheses of Theorem 19, assume moreover that $k \geq 2$, that $l = 1$ if $k = 2$, that elliptic regularity holds for (18), and that*

$$\pi_T^{0,k-2} I_T \phi = \pi_T^{0,k-2} \phi \quad \forall T \in \mathcal{T}_h, \quad \forall \phi \in H^1(T). \tag{57}$$

Then, it holds that

$$\left\| \pi_{\boldsymbol{K},h}^{1,k} u_h - u \right\| \lesssim h^{r+1} \, |u|_{H^{r+1}(\mathcal{T}_h)}, \tag{58}$$

where the multiplicative constant additionally depends on $\boldsymbol{K}$.

Remark 23 (*Assumption* (57)) Assumption (57) holds for both conforming and non-conforming VEM methods, as the moments of degree $k-2$ in the cells are part of the DOFs of the methods, and $I_T \phi$ is defined as the element of V_T^k that has the same DOFs as ϕ.

Remark 24 (*Dependency on the diffusion field in* L^2 *estimates*) Elliptic regularity for problem (18) is only known if Ω is convex and $\boldsymbol{K}$ is Lipschitz continuous. Combined with the assumption that $\boldsymbol{K}$ is piecewise constant, this imposes $\boldsymbol{K}$ constant over the entire domain, which means that we can treat anisotropic but not heterogeneous diffusion. For this reason, we make no attempt whatsoever to track the dependence on the diffusion field in the L^2 error estimate.

Proof The elliptic regularity shows that, for all $g \in L^2(\Omega)$, $z_g \in H^2(\Omega)$ and $\|z_g\|_{H^2(\Omega)} \lesssim \|g\|$. Estimate (58) therefore follows from (40), Theorem 13 with the choice $r_h := \pi_{\boldsymbol{K},h}^{1,k}$, and (57) (which shows that $\pi_{\boldsymbol{K},T}^{1,k} I_T u - u = \pi_{\boldsymbol{K},T}^{1,k} I_T u - u - \pi_T^{0,0}(\pi_{\boldsymbol{K},T}^{1,k} I_T u - u)$, whose $L^2(T)$-norm can be estimated using (21) and (51)), if we can prove that (with, as in the theorem, hidden constants in $\lesssim$ possibly depending on $\boldsymbol{K}$)

$$\left\| \mathcal{E}_h^{\mathrm{d}}(z_g; \cdot) \right\|_{X_h^\star} \lesssim h \, \|z_g\|_{H^2(\Omega)} \tag{59}$$

$$|\mathcal{E}_h(u; I_h z_g)| \lesssim h^{r+1} \, |u|_{H^{r+1}(\mathcal{T}_h)} \, \|z_g\|_{H^2(\Omega)}. \tag{60}$$

(i) *Dual consistency*. With our choice of r_h, we have $\mathcal{E}_h^{\mathrm{d}}(z_g; v_h) = (g, \pi_{\boldsymbol{K},h}^{1,k} v_h) - a_h(v_h, I_h z_g)$. Since a_h is symmetric, we see that $\mathcal{E}_h^{\mathrm{d}}(z_g; v_h)$ is equal to $\mathcal{E}_h(z_g; v_h)$ in which the source term $(f, \pi_h^{0,l} v_h)$ has been replaced with $(g, \pi_{\boldsymbol{K},h}^{1,k} v_h)$. The estimate obtained in the proof of Theorem 19 on the primal consistency error can therefore be used, with $r = 1$ and $z_g \in H^2(\Omega) \subset H^{1+r}(\mathcal{T}_h)$ instead of u, and yields (59), provided we examine the impact of changing $\pi_h^{0,l} v_h$ into $\pi_{\boldsymbol{K},h}^{1,k} v_h$.

The main difference between these two polynomials is that $\pi_{\boldsymbol{K},h}^{1,k} v_h$ is a polynomial of degree $\le k$, whereas $\pi_h^{0,l} v_h$ is a polynomial of degree $l \le k-1$. An inspection of the estimate of the primal consistency error shows that the only place where we used $\pi_h^{0,l} v_h \in \mathbb{P}^{k-1}(\mathcal{T}_h)$ is in (45), when estimating $\left\| \pi_T^{0,l} v_h - \pi_F^{0,k-1} v_h \right\|_F$. Here, we therefore have to establish that, for $F \in \mathcal{F}_T$,

$$\left\| \pi_{\boldsymbol{K},T}^{1,k} v_h - \pi_F^{0,k-1} v_h \right\|_F \lesssim h_T^{\frac{1}{2}} \, \|\nabla v_h\|_T. \tag{61}$$

We introduce $\pm\pi_T^{0,k-1}v_h$ and write

$$\left\|\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_F^{0,k-1}v_h\right\|_F \lesssim \left\|\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h\right\|_F+\left\|\pi_T^{0,k-1}v_h-\pi_F^{0,k-1}v_h\right\|_F \lesssim h_T^{-\frac{1}{2}}\left\|\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h\right\|_T+h_T^{\frac{1}{2}}\left\|\nabla v_h\right\|_T, \tag{62}$$

where the first line is a triangle inequality, and the second line follows from a discrete trace inequality in $\mathbb{P}^k(T)$ (which can be proved along the lines of [22, Lemma 1.46]) together with the arguments deployed after (45). By (23b) and since $\pi_T^{0,k-1}v_h$ has the same average value over T as v_h, we have $\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h=\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h-\pi_T^{0,0}(\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h)$. The approximation properties (21) of $\pi_T^{0,0}$ with $s=1$, $m=0$ and $v=\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h$ then yield

$$\left\|\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h\right\|_T\lesssim h_T\left\|\nabla(\pi_{\boldsymbol{K},T}^{1,k}v_h-\pi_T^{0,k-1}v_h)\right\|_T\lesssim h_T\left\|\nabla v_h\right\|_T,$$

where we have used the boundedness (28) of $\pi_{\boldsymbol{K},T}^{1,k}$, and the estimate $\left\|\nabla\pi_T^{0,k-1}v_h\right\|_T\lesssim\left\|\nabla v_h\right\|_T$ (which follows from (21) with $s=m=1$). Estimate (61) is then a consequence of (62).

(ii) *Primal-dual consistency*. Following the discussion in Sect. 2.4.2, we re-visit the estimates on $\mathfrak{T}_1,\ldots,\mathfrak{T}_4$ in the proof of Theorem 19, and show that an additional $\mathcal{O}(h)$ factor can be obtained when $v_h=I_hz_g$.

Let us start with the estimate (46) on $\mathfrak{T}_1+\mathfrak{T}_2$. Recalling that $(v_h)_{|T}=I_Tz_g$, introducing $\pm(\pi_T^{0,l}z_g-z_g)$ into the norm, and using a triangle inequality, we can write

$$\left\|\pi_T^{0,l}v_h-(v_h)_{|T}\right\|_F\le\left\|\pi_T^{0,l}(I_Tz_g-z_g)-(I_Tz_g-z_g)\right\|_F+\left\|\pi_T^{0,l}z_g-z_g\right\|_F \lesssim h_T^{\frac{1}{2}}\left|I_Tz_g-z_g\right|_{H^1(T)}+h_T^{\frac{3}{2}}\left|z_g\right|_{H^2(T)},$$

where the last line follows by applying (22) with, for the first term, $v=I_Tz_g-z_g$, $s=1\le l+1$ and $m=0$ and, for the second term, $v=z_g$, $s=2\le l+1$ and $m=0$. Invoking then (50) with z_g instead of u and $r=1$, we infer $\left\|\pi_T^{0,l}v_h-v_h\right\|_F\lesssim h_T^{\frac{3}{2}}\left|z_g\right|_{H^2(T)}$. Plugged into (46), this yields, thanks to a Cauchy–Schwarz inequality on the sum over $T\in\mathcal{T}_h$,

$$|\mathfrak{T}_1+\mathfrak{T}_2|\lesssim\sum_{T\in\mathcal{T}_h}h_T^{r+1}\left|u\right|_{H^{r+1}(T)}\left|z_g\right|_{H^2(T)}\le h^{r+1}\left|u\right|_{H^{r+1}(\mathcal{T}_h)}\left|z_g\right|_{H^2(\Omega)}. \tag{63}$$

The term $\mathfrak{T}_4$ is estimated starting from (53). Each term in the right-hand side of this estimate can be estimated by using (55) on u for the first factor, and with z_g instead

of u and $r = 1$ for the second factor. Summing the resulting estimates over $T \in \mathcal{T}_h$ and using a Cauchy–Schwarz inequality on the sum over $T \in \mathcal{T}_h$ shows that

$$|\mathfrak{T}_4| \lesssim \sum_{T\in\mathcal{T}_h} h_T^r \, |u|_{H^{r+1}(T)} \, h_T \left|z_g\right|_{H^2(T)} \leq h^{r+1} \, |u|_{H^{r+1}(\mathcal{T}_h)} \left|z_g\right|_{H^2(\Omega)}. \tag{64}$$

We now turn to $\mathfrak{T}_3$. Coming back to its definition in (42), we have $\mathfrak{T}_3 = \sum_{T\in\mathcal{T}_h} \mathfrak{T}_{3,T}$ with

$$\begin{aligned}
\mathfrak{T}_{3,T} &= -\left(\boldsymbol{K}_T\nabla(I_T u - u), \nabla\pi_{\boldsymbol{K},T}^{1,k} I_T z_g\right)_T \\
&= \left(I_T u - u, \nabla\cdot(\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} I_T z_g)\right)_T - \sum_{F\in\mathcal{F}_T}\left(I_T u - u, \boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} I_T z_g\cdot\boldsymbol{n}_{TF}\right)_F \\
&= \left(\pi_T^{0,k-2}(I_T u - u), \nabla\cdot(\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} I_T z_g)\right)_T \\
&\quad - \sum_{F\in\mathcal{F}_T}\left(\pi_F^{0,k-1}(I_T u - u), \boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} I_T z_g\cdot\boldsymbol{n}_{TF}\right)_F,
\end{aligned}$$

where we have used the definition (23a) of $\pi_{\boldsymbol{K},T}^{1,k}(I_T u - u)$ with $w = \pi_{\boldsymbol{K},T}^{1,k} I_T z_g$ in the first line, an integration by parts in the second line, and the fact that $\nabla\cdot(\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} I_T z_g) \in \mathbb{P}^{k-2}(T)$ and $\boldsymbol{K}_T\nabla\pi_{\boldsymbol{K},T}^{1,k} I_T z_g\cdot\boldsymbol{n}_{TF} \in \mathbb{P}^{k-1}(F)$ to introduce the L^2-projections of $I_T u - u$ in the third line. By Assumption (57), the first term in the right-hand side vanishes, and thus

$$\mathfrak{T}_3 = -\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}(\pi_F^{0,k-1}(I_T u - u), \boldsymbol{K}_T\nabla(\pi_{\boldsymbol{K},T}^{1,k} I_T z_g - z_g)\cdot\boldsymbol{n}_{TF})_F,$$

where we have used the continuity property (33) and the boundary condition (34) on the functions in $V_{h,0}^k$, together with the continuity of the normal component of the flux $\boldsymbol{K}\nabla z_g$, to subtract

$$\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}(\pi_F^{0,k-1}(I_T u - u), \boldsymbol{K}_T\nabla z_g\cdot\boldsymbol{n}_{TF})_T = 0.$$

A Cauchy–Schwarz inequality then gives

$$|\mathfrak{T}_3| \leq \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\left\|\pi_F^{0,k-1}(I_T u - u)\right\|_F \left\|\boldsymbol{K}_T\nabla(\pi_{\boldsymbol{K},T}^{1,k} I_T z_g - z_g)\cdot\boldsymbol{n}_{TF}\right\|_F. \tag{65}$$

We next bound the factors inside the summation. Assumption (57) shows that $\pi_T^{0,0}(I_T u - u) = 0$ and thus, by (22) with $s = 1$ and $m = 0$, we have for the first factor

$$\left\|\pi_F^{0,k-1}(I_T u - u)\right\|_F \le \|I_T u - u\|_F = \left\|(I_T u - u) - \pi_T^{0,0}(I_T u - u)\right\|_F \lesssim h_T^{\frac{1}{2}} \|\nabla(I_T u - u)\|_T \lesssim h_T^{\frac{1}{2}+r} |u|_{H^{r+1}(T)}, \tag{66}$$

the conclusion following from (50). Introducing $\pm \boldsymbol{K}_T \nabla \pi_{\boldsymbol{K},T}^{1,k} z_g$ and using a triangle inequality, we get for the second factor

$$\left\|\boldsymbol{K}_T \nabla(\pi_{\boldsymbol{K},T}^{1,k} I_T z_g - z_g)\cdot \boldsymbol{n}_{TF}\right\|_F \lesssim \left\|\nabla \pi_{\boldsymbol{K},T}^{1,k}(I_T z_g - z_g)\right\|_F + \left\|\nabla(\pi_{\boldsymbol{K},T}^{1,k} z_g - z_g)\right\|_F \lesssim h_T^{-\frac{1}{2}} \left\|\nabla \pi_{\boldsymbol{K},T}^{1,k}(I_T z_g - z_g)\right\|_T + h_T^{\frac{1}{2}} |z_g|_{H^2(T)} \lesssim h_T^{\frac{1}{2}} |z_g|_{H^2(T)}, \tag{67}$$

where we have used a discrete trace inequality in $\mathbb{P}^{k-1}(T)$ and (25) (with $v = z_g$, $s = 2 \le k+1$ and $m = 0$) to pass to the second line, and we have concluded by invoking (51) with z_g instead of u and $r = 1$. Plugging (67) and (66) into (65), we obtain

$$|\mathfrak{T}_3| \lesssim \sum_{T\in\mathcal{T}_h} h_T^{r+\frac{1}{2}} |u|_{H^{r+1}(T)} h_T^{\frac{1}{2}} |z_g|_{H^2(T)} \lesssim h^{r+1} |u|_{H^{r+1}(\mathcal{T}_h)} |z_g|_{H^2(\Omega)}.$$

Together with (63) and (64), this establishes (60) and concludes the proof. □

Remark 25 (Simplifications) The proofs of Theorems 19 and 22 have been made in a unified setting that covers both conforming and non-conforming VEM. Simplifications are possible when those methods are considered individually.

For non-conforming VEM [5], significant simplifications stem from the following preservation properties of the interpolant: for all $T \in \mathcal{T}_h$ and $\phi \in H^1(T)$,

$$\pi_T^{0,k-2} I_T \phi = \pi_T^{0,k-2} \phi \quad \text{and} \quad \pi_F^{0,k-1} I_T \phi = \pi_F^{0,k-1} \phi \quad \forall F \in \mathcal{F}_T.$$

Performing integrations-by-parts on the definition (23a) of $\pi_{\boldsymbol{K},T}^{1,k}$, it can easily be seen that these properties imply $\pi_{\boldsymbol{K},T}^{1,k} I_T \phi = \pi_{\boldsymbol{K},T}^{1,k} \phi$. As a consequence, the term $\mathfrak{T}_3$ entirely vanishes, and a few other estimates are shorter (e.g., (55) is a direct consequence of (49) and (24), etc.). Note that $\mathfrak{T}_3$ is by far the most troublesome term to estimate in the proof of Theorem 22.

In the context of conforming VEM, on the other hand, a slightly simpler argument can be invoked working in a more standard setting corresponding to the classical first Strang lemma; see, e.g., [11, Lemma 3.11]. In this case, the source term for the dual problem is the error $u - u_h$ measured as the difference between the continuous and virtual solutions.

We close this remark by noticing that, unlike [13, Theorem 6] and [11, Theorem 3.14], our L^2-error estimate stems from an application of the abstract result of Theorem 13, which is not problem-specific.

Remark 26 (*The lowest-order case*) As is often the case with mixed and non-conforming methods for diffusion equations on generic grids, L^2 error estimates for the lowest degree(s) require specific work and, possibly, additional regularity on the source term; see, e.g., [21,23] for primal and mixed HHO methods, [5,6] for conforming and non-conforming VEM, [10, Remark 8.5] for further insight into this topic, and [40, Section 2.7] for a fix in the context of high-order Mimetic Finite Difference methods. In Theorem 13, additional work would be required on the primal-dual consistency error. The details can be evinced from the above references, and are omitted here for the sake of brevity.

3.3 Finite volume methods

The second application of the abstract analysis framework of Sect. 2 considered here is to Finite Volume (FV) methods. In this context, several novelties are present. First of all, the analysis is carried out under general assumptions on the numerical fluxes, which enables the simultaneous treatment of several (cell-centred or cell- and face-centred) schemes. Second, we provide a clear definition of consistency also for FV schemes for which this notion hadn't been clearly highlighted in the literature. Third, to the best of our knowledge, we write the first error estimates, for FV methods, in which the dependency on the diffusion field is finely tracked.

3.3.1 General theory

The discrete unknowns of Finite Volume methods are usually values at points. We consider here methods with cell- and face-unknowns (see Sect. 3.3.3 for cell-centred methods). A mesh $\mathcal{M}_h = (\mathcal{T}_h, \mathcal{F}_h)$ being chosen, we therefore take one point $\mathbf{x}_T$ in each cell $T \in \mathcal{T}_h$ and one point $\mathbf{x}_F$ on each face $F \in \mathcal{F}_h$; note that these points may not be the centres of mass of the corresponding geometrical objects, and may need to satisfy specific geometric properties. The $(d-1)$-dimensional measure of a face $F \in \mathcal{F}_h$ is denoted by $|F|$ and, if $T \in \mathcal{T}_F$, d_{TF} is the orthogonal distance between $\mathbf{x}_T$ and F.

The space of unknowns is

$$X_h := \left\{ v_h = ((v_T)_{T\in\mathcal{T}_h}, (v_F)_{F\in\mathcal{F}_h}) \,:\, v_T \in \mathbb{R}\ \forall T \in \mathcal{T}_h, \quad v_F \in \mathbb{R}\ \forall F \in \mathcal{F}_h^{\mathrm{i}}, \right.$$
$$\left. v_F = 0\ \forall F \in \mathcal{F}_h^{\mathrm{b}} \right\},$$

which is equipped with the following discrete equivalent of the H_0^1-norm:

$$\|v_h\|_{1,\mathcal{T}_h} := \left(\sum_{T\in\mathcal{T}_h} \underline{\lambda}_T\, |v_h|_{1,T}^2 \right)^{\frac{1}{2}} \quad \text{with} \quad |v_h|_{1,T}^2 := \sum_{F\in\mathcal{F}_T} |F| d_{TF} \left(\frac{v_T - v_F}{d_{TF}} \right)^2, \tag{68}$$

(see, e.g., [29, Section 7.1]—note that, contrary to this reference, we explicitly account for the diffusion coefficient here). For $u \in C(\overline{\Omega})$ with $u_{|\partial\Omega} = 0$, an interpolant $I_h u \in X_h$ is defined by setting

$$I_h u = \Big((u(\mathbf{x}_T))_{T\in\mathcal{T}_h}, (u(\mathbf{x}_F))_{F\in\mathcal{F}_h} \Big).$$

Note that, in dimensions ≤ 3, the solution u to (19) is (Hölder) continuous on $\overline{\Omega}$ [42].

FV methods are characterised by flux conservativity and balance equations. Following the presentation in [26], a generic FV method for (18) is written: Find $u_h \in X_h$ such that

$$\mathfrak{F}_{T,F}(u_h) + \mathfrak{F}_{T'F}(u_h) = 0 \quad \forall F \in \mathcal{F}_h^{\mathrm{i}} \text{ with } \mathcal{T}_F = \{T, T'\}, \tag{69a}$$

$$\sum_{F\in\mathcal{F}_T} \mathfrak{F}_{T,F}(u_h) = \int_T f \quad \forall T \in \mathcal{T}_h. \tag{69b}$$

Here, for $T \in \mathcal{T}_h$ and $F \in \mathcal{F}_T$, $\mathfrak{F}_{T,F} : X_h \to \mathbb{R}$ is a linear numerical flux such that $\mathfrak{F}_{T,F}(I_h u)$ approximates $-\int_F \boldsymbol{K}\boldsymbol{\nabla} u \cdot \boldsymbol{n}_{TF}$.

The following general estimate is a direct consequence of Theorem 10.

Theorem 27 (Energy estimate for FV methods) *Assume that the fluxes $(\mathfrak{F}_{T,F})_{T\in\mathcal{T}_h,\, F\in\mathcal{F}_T}$ satisfy the following coercivity property, for some $\gamma > 0$: For all $v_h \in X_h$,*

$$\sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} \mathfrak{F}_{T,F}(v_h)(v_T - v_F) \geq \gamma \, \|v_h\|_{1,\mathcal{T}_h}^2 . \tag{70}$$

Then, if the solution u to (19) belongs to $C(\overline{\Omega}) \cap H^2(\mathcal{T}_h)$, denoting by u_h the solution to the FV scheme (69), it holds

$$\|u_h - I_h u\|_{1,\mathcal{T}_h} \leq \gamma^{-1} \left(\sum_{T\in\mathcal{T}_h} \underline{\lambda}_T^{-1} \sum_{F\in\mathcal{F}_T} \frac{d_{TF}}{|F|} \left[\int_F \boldsymbol{K}_T \boldsymbol{\nabla} u_{|T} \cdot \boldsymbol{n}_{TF} + \mathfrak{F}_{T,F}(I_h u) \right]^2 \right)^{\frac{1}{2}} . \tag{71}$$

Remark 28 (*Consistency of the fluxes*) Estimate (71) highlights the following well-known fact (see [38, Example 3.1] or [26, Remark 1.3]): in FV methods, the appropriate consistency is that of the *fluxes*, not of the discrete second order differential operator as in Finite Difference methods .

Proof We first recast problem (69) under a discrete weak form. For an arbitrary vector $v_h = ((v_T)_{T\in\mathcal{T}_h}, (v_F)_{F\in\mathcal{F}_h}) \in X_h$, notice that, by the flux conservativity (69a) and the boundary condition on v_h,

$$\begin{aligned}\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\mathfrak{F}_{T,F}(u_h)v_F = &\sum_{F\in\mathcal{F}_h^{\mathrm{i}},\,\mathcal{T}_F=\{T,T'\}}\left(\mathfrak{F}_{T,F}(u_h)+\mathfrak{F}_{T'F}(u_h)\right)v_F\\ &+\sum_{F\in\mathcal{F}_h^{\mathrm{b}},\,\mathcal{T}_F=\{T\}}\mathfrak{F}_{T,F}(u_h)v_F = 0,\end{aligned}\tag{72}$$

where the first equality comes from a re-arrangement of the sum over the faces. Hence, multiplying (69b) by v_T, summing over $T\in\mathcal{T}_h$ and using the above relation, we see that u_h satisfies

$$\underbrace{\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\mathfrak{F}_{T,F}(u_h)(v_T-v_F)}_{a_h(u_h,v_h)} = \underbrace{\sum_{T\in\mathcal{T}_h}\int_T f v_T}_{\ell_h(v_h)}\quad\forall v_h\in X_h.\tag{73}$$

This problem has the form (4) with $X_h = Y_h$. The coercivity assumption (70) shows that a_h is coercive on X_h, with coercivity constant γ. Hence, Theorem 10 yields

$$\|u_h - I_h u\|_{1,\mathcal{T}_h} \le \gamma^{-1}\,\|\mathcal{E}_h(I_h u;\cdot)\|_{X_h^\star}.\tag{74}$$

To estimate the primal consistency error, notice first that the relation $f = -\nabla\cdot(\boldsymbol{K}\nabla u)$ and the divergence formula in each cell give

$$\begin{aligned}\ell_h(v_h) = \sum_{T\in\mathcal{T}_h}\left(\int_T -\nabla\cdot(\boldsymbol{K}\nabla u)\right)v_T &= -\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\left(\int_F \boldsymbol{K}_T\nabla u_{|T}\cdot\boldsymbol{n}_{TF}\right)v_T\\ &= -\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\left(\int_F \boldsymbol{K}_T\nabla u_{|T}\cdot\boldsymbol{n}_{TF}\right)(v_T-v_F)\end{aligned}$$

where we have used (72) with $\mathfrak{F}_{T,F}(u_h)$ replaced with $\int_F \boldsymbol{K}_T\nabla u_{|T}\cdot\boldsymbol{n}_{TF}$ (these exact fluxes also satisfy the conservativity relation (69a) since $\nabla\cdot(\boldsymbol{K}\nabla u)\in L^2(\Omega)$). Hence,

$$\mathcal{E}_h(I_h u; v_h) = -\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\left[\int_F \boldsymbol{K}_T\nabla u_{|T}\cdot\boldsymbol{n}_{TF} + \mathfrak{F}_{T,F}(I_h u)\right](v_T - v_F).$$

A Cauchy–Schwarz inequality and the definition (68) of the norm on X_h shows that $\|\mathcal{E}_h(I_h u;\cdot)\|_{X_h^\star}$ is bounded above by the bracketed term in the right-hand side of (71). Plugging into (74) this bound of the primal consistency error concludes the proof. □

3.3.2 Stable and linearly exact fluxes

The estimate (71) enables us to identify simple local properties on the fluxes, under which an $\mathcal{O}(h)$ energy estimate can be established: local dependency, linear exactness and boudedness. Similar properties were proposed in [28], but without the concept of local dependency, which is essential for establishing a proper error estimate. Additionally, the analysis in [28] was only sketched, and did not track the dependency of the estimates on the diffusion tensor $\boldsymbol{K}$.

In this section, for $T \in \mathcal{T}_h$ we let $X_T := \{v = (v_T, (v_F)_{F\in\mathcal{F}_T}) : v_T \in \mathbb{R},\ v_F \in \mathbb{R}\ \forall F \in \mathcal{F}_T\}$ be the local space of unknowns and, for $\phi \in C(\overline{T})$, $I_T\phi = (\phi(\mathbf{x}_T), (\phi(\mathbf{x}_F))_{F\in\mathcal{F}_T}) \in X_T$ defines the local interpolant of ϕ.

Theorem 29 (Energy error estimate for linearly exact FV methods) *Assume that the family of numerical fluxes $(\mathfrak{F}_{T,F})_{T\in\mathcal{T}_h,\,F\in\mathcal{F}_T}$ satisfies the coercivity property (70), as well as the following properties:*

(i) Local dependency and linear exactness. *For all $v_h \in X_h$, $T \in \mathcal{T}_h$ and $F \in \mathcal{F}_T$, $\mathfrak{F}_{T,F}(v_h)$ depends only on the values $(v_T, (v_F)_{F\in\mathcal{F}_T}) \in X_T$, and if L is an affine function on $\overline{T}$ then $\mathfrak{F}_{T,F}(I_T L) = -\int_F \boldsymbol{K}_T\nabla L\cdot\boldsymbol{n}_{TF}$.*

(ii) Boundedness. *There is $C_b \geq 0$ such that, for all $T \in \mathcal{T}_h$ and $v \in X_T$,*

$$\sum_{F\in\mathcal{F}_T} \frac{d_{TF}}{|F|}|\mathfrak{F}_{T,F}(v)|^2 \leq C_b\overline{\lambda}_T^2\,|v|_{1,T}^2\,. \tag{75}$$

Let

$$\theta \geq \max_{T\in\mathcal{T}_h}\left(\max_{F\in\mathcal{F}_T}\frac{h_T}{d_{TF}} + \mathrm{Card}(\mathcal{F}_T)\right).$$

Then, if the solution u to (19) belongs to $C(\overline{\Omega})\cap H^2(\mathcal{T}_h)$, denoting by u_h the solution of (69),

$$\|u_h - I_h u\|_{1,\mathcal{T}_h} \lesssim \gamma^{-1}\left(\sum_{T\in\mathcal{T}_h}\alpha_T\overline{\lambda}_T h_T^2\,|u|_{H^2(T)}^2\right)^{\frac{1}{2}}, \tag{76}$$

with hidden constant independent on $\boldsymbol{K}$ and h, but depending on θ and C_b.

Proof Fix $T \in \mathcal{T}_h$ and notice that, by definition of θ and [29, Lemma B.1], there is a ball of radius $\gtrsim h_T$ such that T is star-shaped with respect to all points in this ball. Hence, [29, Lemma 7.61] yields the existence of a linear function L_T such that, setting $R_T = u_{|T} - L_T$,

$$\sup_{\overline{T}}|R_T| \lesssim h_T^{2-\frac{d}{2}}\,|u|_{H^2(T)} \quad\text{and}\quad \|\nabla R_T\|_T \lesssim h_T\,|u|_{H^2(T)}\,. \tag{77}$$

Subtracting L_T and using the linear exactness of the fluxes, we have

$$\mathfrak{T}_{TF} := \left| \int_F \boldsymbol{K}_T \nabla u_{|T} \cdot \boldsymbol{n}_{TF} + \mathfrak{F}_{T,F}(I_T u_{|T}) \right| = \left| \int_F \boldsymbol{K}_T \nabla R_T \cdot \boldsymbol{n}_{TF} + \mathfrak{F}_{T,F}(I_T R_T) \right| \leq \overline{\lambda}_T \int_F |\nabla R_T| + |\mathfrak{F}_{T,F}(I_T R_T)|.$$

Hence, by boundedness of the fluxes,

$$\sum_{F \in \mathcal{F}_T} \frac{d_{TF}}{|F|} \mathfrak{T}_{TF}^2 \leq 2\overline{\lambda}_T^2 \sum_{F \in \mathcal{F}_T} d_{TF}|F| \left(\frac{1}{|F|} \int_F |\nabla R_T| \right)^2 + 2C_b \overline{\lambda}_T^2 |I_T R_T|_{1,T}^2 =: \mathfrak{T}_T^{(1)} + \mathfrak{T}_T^{(2)}.$$

The definition of $\mathfrak{T}_{TF}$ and Theorem 27 show that

$$\|u_h - I_h u\|_{1,\mathcal{T}_h} \leq \gamma^{-1} \left(\sum_{T \in \mathcal{T}_h} \underline{\lambda}_T^{-1} (\mathfrak{T}_T^{(1)} + \mathfrak{T}_T^{(2)}) \right)^{\frac{1}{2}}. \tag{78}$$

To estimate $\mathfrak{T}_T^{(1)}$, we apply [29, Lemma B.6] to $|\nabla R_T| \in H^1(T)$ to see that

$$\left(\frac{1}{|F|} \int_F |\nabla R_T| \right)^2 \lesssim \left(\frac{1}{|T|} \int_T |\nabla R_T| \right)^2 + \frac{h_T}{|F|} |R_T|_{H^2(T)}^2.$$

Since L_T is linear, $|R_T|_{H^2(T)} = |u - L_T|_{H^2(T)} = |u|_{H^2(T)}$. Hence, the Jensen inequality on the first term in the right-hand side and (77) yield

$$\left(\frac{1}{|F|} \int_F |\nabla R_T| \right)^2 \lesssim \left(\frac{h_T^2}{|T|} + \frac{h_T}{|F|} \right) |u|_{H^2(T)}^2.$$

Plugging this bound into the definition of $\mathfrak{T}_T^{(1)}$, using $d_{TF} \leq h_T$, and using $\sum_{F \in \mathcal{F}_T} d_{TF}|F| = d|T|$ (see [29, Lemma B.2]), we infer

$$\mathfrak{T}_T^{(1)} \lesssim \overline{\lambda}_T^2 h_T^2 |u|_{H^2(T)}^2. \tag{79}$$

For $\mathfrak{T}_T^{(2)}$, we recall the definition of $|\cdot|_{1,T}$, use the first bound in (77), and the estimates $\frac{1}{d_{TF}} \leq \frac{\theta}{h_T}$ and $|F| \lesssim h_T^{d-1}$ to write

$$\mathfrak{T}_T^{(2)} \lesssim \overline{\lambda}_T^2 \sum_{F \in \mathcal{F}_T} \frac{|F|}{d_{TF}} (|R_T(\mathbf{x}_T)|^2 + |R_T(\mathbf{x}_F)|^2) \lesssim \overline{\lambda}_T^2 h_T^2 |u|_{H^2(T)}^2.$$

Using this estimate together with (79) into (78) concludes the proof. □

We now give two classical examples of FV methods that satisfy the coercivity, linear exactness and stability properties, and to which Theorem 29 thus applies. Error estimates for these two methods can be found in the literature (see e.g. [30,38]) but, to our best knowledge, contrary to (76), none of the currently available estimate has explicit dependency on the local anisotropy ratio and diffusion magnitude.

Example 30 (Two-Point Flux Approximation (TPFA) method) The TPFA scheme [38] requires meshes with a specific geometric property: the points $(\mathbf{x}_T)_{T\in\mathcal{T}_h}$ and $(\mathbf{x}_F)_{F\in\mathcal{F}_h}$ must be chosen such that, for any $T\in\mathcal{T}_h$ and $F\in\mathcal{F}_T$, $\mathbf{x}_T\mathbf{x}_F$ is parallel to $\boldsymbol{K}_T\boldsymbol{n}_{TF}$. The fluxes are then defined by: for $v_h\in X_h$,

$$\mathfrak{F}_{T,F}(v_h) = |F|\,|\boldsymbol{K}_T\boldsymbol{n}_{TF}|\frac{v_T - v_F}{|\mathbf{x}_T-\mathbf{x}_F|}. \tag{80}$$

The assumption on the points show that $\mathbf{x}_F-\mathbf{x}_T = \alpha_{TF}\boldsymbol{K}_T\boldsymbol{n}_{TF}$ with $\alpha_{TF}>0$ (because $\boldsymbol{K}_T$ is symmetric positive definite and $(\mathbf{x}_F-\mathbf{x}_T)\cdot\boldsymbol{n}_{TF}>0$). Taking the norm on both sides yields $\alpha_{TF}=\frac{|\mathbf{x}_T-\mathbf{x}_F|}{|\boldsymbol{K}_T\boldsymbol{n}_{TF}|}$. Hence, if L is a linear function,

$$L(\mathbf{x}_T)-L(\mathbf{x}_F) = \nabla L\cdot(\mathbf{x}_T-\mathbf{x}_F) = -\frac{|\mathbf{x}_T-\mathbf{x}_F|}{|\boldsymbol{K}_T\boldsymbol{n}_{TF}|}\nabla L\cdot\boldsymbol{K}_T\boldsymbol{n}_{TF}$$

and thus $\mathfrak{F}_{T,F}(I_T L) = -|F|\boldsymbol{K}_T\nabla L\cdot\boldsymbol{n}_{TF}$, showing that the flux is linearly exact. Fixing $C_b \geq \max_{T\in\mathcal{T}_h}\max_{F\in\mathcal{F}_T}\frac{d_{TF}^2}{|\mathbf{x}_T-\mathbf{x}_F|^2}$, the boundedness property (75) is a straightforward consequence of (80). Since $|\boldsymbol{K}_T\boldsymbol{n}_{TF}|\geq\underline{\lambda}_T$, the coercivity (70) also easily follows from (80), provided that $\gamma>0$ is chosen such that $\gamma\leq\min_{T\in\mathcal{T}_h}\min_{F\in\mathcal{F}_T}\frac{d_{TF}}{|\mathbf{x}_T-\mathbf{x}_F|}$.

Example 31 (Mixed Finite Volume (MFV) method) The MFV method is the FV presentation of the Hybrid Mimetic Mixed (HMM) method [30]. Here, $(\mathbf{x}_T)_{T\in\mathcal{T}_h}$ can be any points in the cells, but $(\mathbf{x}_F)_{F\in\mathcal{F}_h}=(\overline{\mathbf{x}}_F)_{F\in\mathcal{F}_h}$ are taken as the centers of mass of the faces. To construct the MFV method [27,30], we start by reconstructing, from known fluxes, a local gradient. For $T\in\mathcal{T}_h$, if $\mathfrak{f}_T=(\mathfrak{f}_{TF})_{F\in\mathcal{F}_T}$ is a family of real numbers (representing fluxes through the faces of T), define the following discrete gradient and boundary residuals:

$$\mathbf{G}_T(\mathfrak{f}_T) := -\frac{1}{|T|}\boldsymbol{K}_T^{-1}\sum_{F\in\mathcal{F}_T}\mathfrak{f}_{TF}(\overline{\mathbf{x}}_F-\mathbf{x}_T),$$
$$\mathcal{R}_{TF}(\mathfrak{f}_T) := \mathfrak{f}_{TF}+|F|\boldsymbol{K}_T\mathbf{G}_T(\mathfrak{f}_T)\cdot\boldsymbol{n}_{TF}\quad\forall F\in\mathcal{F}_T.$$

Then, fixing a symmetric positive definite matrix $\mathbb{B}^T=(\mathbb{B}^T_{FF'})_{F,F'\in\mathcal{F}_T}\in\mathbb{R}^{\mathcal{F}_T\times\mathcal{F}_T}$, the MFV fluxes $(\mathfrak{F}_{T,F}(v_h))_{F\in\mathcal{F}_T}$ are defined, for $v_h\in X_h$, as the unique solution of the following problem

$$\forall \mathfrak{f}_T = (\mathfrak{f}_{TF})_{F\in\mathcal{F}_T} \in \mathbb{R}^{\mathcal{F}_T},$$
$$|T|\boldsymbol{K}_T\mathbf{G}_T(\mathfrak{F}_{T,F}(v_h))\cdot\mathbf{G}_T(\mathfrak{f}_T) + \sum_{F,F'\in\mathcal{F}_T} \mathbb{B}^T_{FF'}\mathcal{R}_{TF}(\mathfrak{F}_{T,F}(v_h))\mathcal{R}_{TF}(\mathfrak{f}_T)$$
$$= \sum_{F\in\mathcal{F}_T} (v_T - v_F)\mathfrak{f}_{TF}. \tag{81}$$

Assume that L is a linear map and that $v_h = I_h L$. Let $\mathfrak{g}_T = (-|F|\boldsymbol{K}_T\nabla L\cdot\boldsymbol{n}_{TF})_{F\in\mathcal{F}_T}$ be the exact fluxes of L. The divergence theorem shows that $\mathbf{G}_T(\mathfrak{g}_T) = \nabla L$ and thus $\mathcal{R}_{TF}(\mathfrak{g}_T) = 0$. Moreover, for all $\mathfrak{f}_T \in \mathbb{R}^{\mathcal{F}_T}$,

$$\sum_{F\in\mathcal{F}_T} (v_T - v_F)\mathfrak{f}_{TF} = \sum_{F\in\mathcal{F}_T} \nabla L\cdot(\mathbf{x}_T - \overline{\mathbf{x}}_F)\mathfrak{f}_{TF} = \nabla L\cdot|T|\boldsymbol{K}_T\mathbf{G}_T(\mathfrak{f}_T).$$

Hence, (81) holds with $\mathfrak{g}_T$ instead of $(\mathfrak{F}_{T,F}(v_h))_{F\in\mathcal{F}_T}$, which shows that these two families of fluxes are equal, and thus that the fluxes are linearly exact. The stability and coercivity of the method follow easily from (81), under natural assumption on the matrices $\mathbb{B}^T$, see [30, Section 4.1] or [29, Chapter 13].

Remark 32 (L^2 *estimates and super-convergence*) Define $r_h : X_h \to L^2(\Omega)$ by $(r_h v_h)_{|T} = v_T$ for all $v_h \in X_h$ and $T \in \mathcal{T}_h$. A discrete Poincaré inequality [29, Remark B.16] yields $\|r_h v_h\| \le C\|v_h\|_{1,\mathcal{T}_h}$, with C depending only on $\eta \ge \max_{F\in\mathcal{F}_h^{\mathrm{i}},\,\mathcal{T}_F=\{T,T'\}}\left(\frac{d_{TF}}{d_{T'F}} + \frac{d_{T'F}}{d_{TF}}\right)$. Hence, (71) and (76) directly give estimates on $\left\|r_h u_h - u_{\mathcal{T}_h}\right\|$, where $u_{\mathcal{T}_h} = r_h I_h u$ is the piecewise constant function defined by $(u_{\mathcal{T}_h})_{|T} = u(\mathbf{x}_T)$ for all $T \in \mathcal{T}_h$.

One can naturally wonder whether Theorem 13 could yield better error estimates on this L^2-norm. The answer is no in general. Numerical test 2 in [31] shows that, for the MFV scheme, the L^2-norm error can, in some cases, converge at the same rate as the discrete energy error (that is, $\mathcal{O}(h)$). Actually, for the MFV and TPFA schemes at least, the super-convergence properties in L^2-norm seem to be related to the proximity, locally and on average, of the interpolation points $(\mathbf{x}_T)_{T\in\mathcal{T}_h}$ and the centers of mass of the cells [31, Theorem 5.3].

3.3.3 Cell-centred methods, application to multi-point flux approximations

The theory in Sect. 3.3.1 can easily be adapted to purely cell-centred methods. For such methods, the space of unknowns is

$$X_h^c := \mathbb{P}^0(\mathcal{T}_h) = \left\{v_h = (v_T)_{T\in\mathcal{T}_h} \,:\, v_T \in \mathbb{R}\right\},$$

with discrete H_0^1 norm defined by

$$\|v_h\|_{1,\mathcal{T}_h,c} := \left(\sum_{F\in\mathcal{F}_h} \underline{\lambda}_F|F|d_F\left(\frac{v_T - v_{T'}}{d_F}\right)^2\right)^{\frac{1}{2}},$$

with the notations

$$\forall F \in \mathcal{F}_h^{\mathrm{i}} \,:\, \underline{\lambda}_F = \min(\underline{\lambda}_T, \underline{\lambda}_{T'}) \text{ and } d_F = d_{TF} + d_{T'F}\,, \quad \text{where } \{T, T'\} = \mathcal{T}_F,$$
$$\forall F \in \mathcal{F}_h^{\mathrm{b}} \,:\, \underline{\lambda}_F = \underline{\lambda}_T\,,\ d_F = d_{TF} \text{ and } v_{T'} = 0, \quad \text{where } \{T\} = \mathcal{T}_F.$$

The interpolant of a continuous function u is $I_h^c u := (u(\mathbf{x}_T))_{T\in\mathcal{T}_h} \in X_h^c$. To write a cell-centred FV method, linear fluxes $\mathfrak{F}_{T,F}^c : X_h^c \to \mathbb{R}$ are first chosen such that

$$\mathfrak{F}_{T,F}^c + \mathfrak{F}_{T'F}^c = 0 \text{ on } X_h^c, \quad \forall F \in \mathcal{F}_h^{\mathrm{i}} \text{ with } \mathcal{T}_F = \{T, T'\}. \tag{82}$$

Then the FV scheme reads: Find $u_h \in X_h^c$ such that

$$\sum_{F\in\mathcal{F}_T} \mathfrak{F}_{T,F}^c(u_h) = \int_T f \quad \forall T \in \mathcal{T}_h. \tag{83}$$

The following result is the equivalent for cell-centred methods of Theorem 27.

Theorem 33 (Energy estimate for cell-centred FV methods) *Assume that the fluxes $(\mathfrak{F}_{T,F}^c)_{T\in\mathcal{T}_h,\, F\in\mathcal{F}_T}$ satisfy the following coercivity property, for some $\gamma > 0$: For all $v_h \in X_h^c$,*

$$\sum_{F\in\mathcal{F}_h^{\mathrm{i}},\, \mathcal{T}_F=\{T,T'\}} \mathfrak{F}_{T,F}^c(v_h)(v_T - v_{T'}) + \sum_{F\in\mathcal{F}_h^{\mathrm{b}},\, \mathcal{T}_F=\{T\}} \mathfrak{F}_{T,F}^c(v_h) v_T \geq \gamma \left\| v_h \right\|_{1,\mathcal{T}_h,c}^2. \tag{84}$$

Then, if the solution u to (19) belongs to $C(\overline{\Omega}) \cap H^2(\mathcal{T}_h)$, denoting by u_h the solution to the FV scheme (83), it holds

$$\left\| u_h - I_h^c u \right\|_{1,\mathcal{T}_h,c} \leq \gamma^{-1} \left(\sum_{F\in\mathcal{F}_h} \underline{\lambda}_F^{-1} \frac{d_F}{|F|} \left[\int_F \boldsymbol{K}_T \nabla u_{|T} \cdot \boldsymbol{n}_{TF} + \mathfrak{F}_{T,F}^c(I_h^c u) \right]^2 \right)^{\frac{1}{2}} \tag{85}$$

where, for $F \in \mathcal{F}_h$, T is an arbitrary cell in $\mathcal{T}_F$.

Proof For all $v_h \in X_h^c$, gathering the sum by faces and using the flux conservativity (82) shows that

$$\begin{aligned}
\ell_h(v_h) := \sum_{T\in\mathcal{T}_h} \int_T f v_T &= \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} \mathfrak{F}_{T,F}^c(u_h) v_T \\
&= \sum_{F\in\mathcal{F}_h^{\mathrm{i}},\, \mathcal{T}_F=\{T,T'\}} \mathfrak{F}_{T,F}^c(u_h) v_T + \mathfrak{F}_{T'F}^c(u_h) v_{T'} \\
&\quad + \sum_{F\in\mathcal{F}_h^{\mathrm{b}},\, \mathcal{T}_F=\{T\}} \mathfrak{F}_{T,F}^c(u_h) v_T
\end{aligned}$$

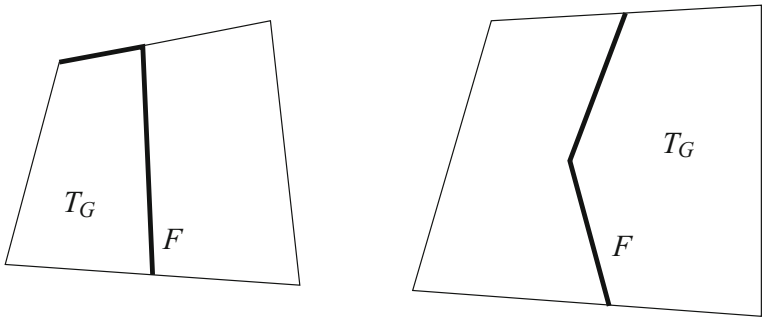

Fig. 1 Two examples of groups of face (in bold) containing one particular face F. Left: unique choice for T_G; right: two possible choices for T_G, one has been arbitrarily made

$$
\begin{aligned}
&= \sum_{F\in\mathcal{F}_h^{\mathrm{i}},\, \mathcal{T}_F=\{T,T'\}} \mathfrak{F}_{T,F}^{c}(u_h)(v_T - v_{T'}) \\
&\quad + \sum_{F\in\mathcal{F}_h^{\mathrm{b}},\, \mathcal{T}_F=\{T\}} \mathfrak{F}_{T,F}^{c}(u_h) v_T \\
&=: a_h(u_h, v_h).
\end{aligned}
$$

Hence, the cell-centred Finite Volume scheme (83) has been recast in the framework of Sect. 2. The error estimate (85) then follows Theorem 10, in a similar way as the error estimate (71) for cell- and face-centred schemes. □

As for cell- and face-centred FV methods, we could deduce from this theorem an error estimate for schemes with local, bounded and linearly exact fluxes. However, some important FV methods are not linearly exact if the diffusion tensor $\boldsymbol{K}$ is discontinuous. This is the case, for example, of Multi-Point Flux Approximation (MPFA) methods [1,2,33]. To properly account for the diffusion jump, the fluxes are constructed to be exact on interpolants of *piecewise* linear functions that have continuous fluxes (and, thus, usually discontinuous gradients to compensate for the discontinuity of the diffusion tensor involved in the fluxes). Theorem 33 however still yields energy error estimate for such methods. To illustrate this, we consider here the case of two Multi-Point Flux Approximation methods: the MPFA-L and MPFA-G methods.

Let us first briefly present these two schemes (see [3,4] for the details). Here, $(\mathbf{x}_T)_{T\in\mathcal{T}_h}$ are still free points in the cells, but $(\mathbf{x}_F)_{F\in\mathcal{F}_h}$ are the centers of mass of the faces. A *group of faces* is any set of d faces that belong to the same cell and share the same vertex, see Fig. 1. For each such group G we fix a cell T_G whose boundary contains all the faces in G; in most cases there is actually only one such cell, but for some non-convex cells there might situations with two possible choices for T_G—in which case we arbitrarily fix one choice (see Fig. 1, right).

The fluxes of $v_h \in X_h^c$ are constructed via the notion of *group gradients*. For a given group of faces G, let $\mathcal{T}_G$ be the set of cells that have at least one face in G. The group gradients $\left\{(\nabla_{\mathcal{D}} v_h)_T^{G,F} \ : \ T \in \mathcal{T}_G \,,\ F \in \mathcal{F}_T \cap G\right\} \subset \mathbb{R}^d$ are constructed by imposing the continuity of values and fluxes on all faces $F \in G$ of the piecewise linear functions

having these gradients in each corresponding cells, and taking the values $(v_T)_{T\in\mathcal{T}_G}$ at $(\mathbf{x}_T)_{T\in\mathcal{T}_G}$. Additionally, for the cell T_G previously selected, it is imposed that all group gradients are independent of the corresponding faces: $(\nabla_{\mathcal{D}} v_h)_{T_G}^{G,F} = (\nabla_{\mathcal{D}} v_h)_{T_G}^{G,F'}$ for all $F, F' \in \mathcal{F}_T \cap G$. We denote by $(\nabla_{\mathcal{D}} v_h)_{T_G}^{G}$ the common value of all these group gradients associated with T_G, and it can be proved that this vector is a solution of the following linear system:

$$\mathcal{A}_G(\nabla_{\mathcal{D}} v_h)_{T_G}^{G} = \mathcal{B}_G(v_h), \tag{86}$$

where $\mathcal{A}_G \in \mathbb{R}^{d\times d}$ is defined row-wise by

$$\mathcal{A}_G = \begin{bmatrix} \left(\frac{\boldsymbol{K}_T \boldsymbol{n}_{TF}\cdot\boldsymbol{n}_{TF}}{d_{T,F}}(\mathbf{x}_T - \mathbf{x}_{T_G}) + \boldsymbol{K}_{T_G}\boldsymbol{n}_{T_GF} + \boldsymbol{K}_T\boldsymbol{n}_{TF}\right)^t_{F\in G\cap\mathcal{F}_h^{\mathrm{i}}} \\ \left(\frac{\boldsymbol{K}_{T_G}\boldsymbol{n}_{T_GF}\cdot\boldsymbol{n}_{T_GF}}{d_{T_G,F}}(\mathbf{x}_F - \mathbf{x}_{T_G})\right)^t_{F\in G\cap\mathcal{F}_h^{\mathrm{b}}} \end{bmatrix},$$

with T the cell on the other side of T_G with respect to F, and $\mathcal{B}_G(v) \in \mathbb{R}^d$ is defined as

$$\mathcal{B}_G(v_h) = \begin{bmatrix} \left(\frac{\boldsymbol{K}_T\boldsymbol{n}_{TF}\cdot\boldsymbol{n}_{TF}}{d_{T,F}}(v_T - v_{T_G})\right)_{F\in G\cap\mathcal{F}_h^{\mathrm{i}}} \\ \left(\frac{\boldsymbol{K}_{T_G}\boldsymbol{n}_{T_GF}\cdot\boldsymbol{n}_{T_GF}}{d_{T_G,F}}(-v_{T_G})\right)_{F\in G\cap\mathcal{F}_h^{\mathrm{b}}} \end{bmatrix}.$$

For a given face F, we denote by $\mathcal{G}_F$ the set of groups G containing F and such that $\mathcal{A}_G$ is invertible (it is assumed that $\mathcal{G}_F \neq \emptyset$ for all $F \in \mathcal{F}_h$). The numerical fluxes are then defined as a convex combination of the fluxes corresponding to the group gradients: for $T \in \mathcal{T}_h$, $F \in \mathcal{F}_T$ and $v_h \in X_h^c$,

$$\mathfrak{F}_{T,F}^c(v_h) := \sum_{G\in\mathcal{G}_F} \theta_F^G \mathfrak{F}_{T,F}^{c,G}(v_h) \quad \text{with} \quad \mathfrak{F}_{T,F}^{c,G}(v_h) := -|F|\boldsymbol{K}_T(\nabla_{\mathcal{D}} v_h)_T^{G,F}\cdot\boldsymbol{n}_{TF}, \tag{87}$$

where $(\theta_F^G)_{G\in\mathcal{G}_F}$ are the coefficients of the convex combination. The L-scheme corresponds to the case where, for each face, this convex combination has only one non-zero coefficient, chosen to maximise the monotonicity properties of the scheme. The G-scheme corresponds to a choice of coefficients that maximise the coercivity properties of the resulting scheme.

We now show that Theorem 33 yields the following error estimate. This estimate seems to be the first one for the MPFA-L and MPFA-G methods in the case of discontinuous permeability tensors; all previous estimates available in the literature have been derived under the assumption that $\boldsymbol{K} \in C^1(\overline{\Omega})^{d\times d}$, see [26] and reference therein.

Theorem 34 (Error estimate for the MPFA-L/G methods) *Let η be such that*

$$\eta \geq \max_{T\in\mathcal{T}_h,\, F\in\mathcal{F}_T} \frac{h_T}{d_{TF}}, \quad \eta \geq \max_{F\in\mathcal{F}_h^{\mathrm{i}},\, \mathcal{T}_F=\{T,T'\}} \frac{d_{TF}}{d_{T'F}} \quad \textit{and} \quad \eta \geq \max_{F\in\mathcal{F}_h} \sum_{G\in\mathcal{G}_F} \theta_F^G |\mathcal{A}_G^{-1}|,$$

where $|\mathcal{A}_G^{-1}|$ *is the induced Euclidean norm of* $\mathcal{A}_G^{-1}$. *Assume that the solution* u *to* (19) *belongs to* $C(\overline{\Omega})$ *and that* $u_{|\Omega_i} \in C^2(\overline{\Omega_i})$ *for each* $i \in \{1, \ldots, N_\Omega\}$. *Assume that the fluxes* (87) *satisfy the coercivity property* (70), *and let* u_h *be the solution to the MPFA-L/G scheme (that is,* (83) *with these fluxes). Then*

$$\left\| u_h - I_h^c u \right\|_{1,\mathcal{T}_h,c} \lesssim \gamma^{-1} \|u\|_{C^2} h,$$

where $\|u\|_{C^2} := \max_{i=1,\ldots,N_\Omega} \|u\|_{C^2(\overline{\Omega_i})}$ *and the hidden constant in* $\lesssim$ *depends only on* Ω, η *and* $\boldsymbol{K}$.

Remark 35 (*About the coercivity*) In general, the coercivity of MPFA methods is not known, and numerical tests indicate that it might actually fail for MPFA-O scheme on some very distorted meshes [26, Section 3.3]. However, for the MPFA-L/G schemes, an indicator can be designed that only requires to compute the eigenvalues of small systems, and that provides a sufficient condition for the methods to be coercive [4, Lemma 3.4].

Proof [4, Lemma 3.3] shows that, for all $T \in \mathcal{T}_h$, $F \in \mathcal{F}_T$ and $G \in \mathcal{G}_F$, $|(\nabla_{\mathcal{D}} I_h^c u)_T^{G,F} - \nabla u(\mathbf{x}_T)| \lesssim \|u\|_{C^2} (1 + |\mathcal{A}_G^{-1}|)h$ (in this lemma, the quantity $\|u\|_{C^2}$ does not explicitly appear but is hidden in a constant 'C_5'; the proof however clearly shows that this constant depends linearly on $\|u\|_{C^2}$). By C^2 regularity of u in the sub-domain Ω_i that contains T, we infer that

$$\sup_{\mathbf{x} \in F} |(\nabla_{\mathcal{D}} I_h^c u)_T^{G,F} - \nabla u(\mathbf{x})| \lesssim \|u\|_{C^2} (1 + |\mathcal{A}_G^{-1}|)h.$$

Hence,

$$\left| -|F| \boldsymbol{K}_T (\nabla_{\mathcal{D}} I_h^c u)_T^{G,F} \cdot \boldsymbol{n}_{TF} + \int_F \boldsymbol{K}_T \nabla u_{|T} \cdot \boldsymbol{n}_{TF} \right| \lesssim |F| \, \|u\|_{C^2} (1 + |\mathcal{A}_G^{-1}|)h.$$

Taking the convex combination weighted by $(\theta_F^G)_{G \in \mathcal{G}_F}$ of this inequality and recalling the definition of η and (87), we infer that

$$\left| \mathfrak{F}_{T,F}^c (I_h^c u) + \int_F \boldsymbol{K}_T \nabla u_{|T} \cdot \boldsymbol{n}_{TF} \right| \lesssim |F| \, \|u\|_{C^2} h.$$

The proof is completed by plugging this estimate into (85) and by noticing that $d_F |F| \lesssim |T| + |T'|$ if $F \in \mathcal{F}_h^{\mathrm{i}}$ with $\mathcal{T}_F = \{T, T'\}$, and $d_F |F| \lesssim |T|$ if $F \in \mathcal{F}_h^{\mathrm{b}}$ with $\mathcal{T}_F = \{T\}$, so that $\sum_{F \in \mathcal{F}_h} d_F |F| \lesssim \sum_{T \in \mathcal{T}_h} |T| = |\Omega|$. □

4 Conclusion

We developed an abstract analysis framework, in the spirit of Strang's second lemma, for approximations of linear partial differential equations (PDEs) in weak form. Contrary to Strang's lemma, the approximations can be written in fully discrete form,

with test and trial spaces that are not spaces of functions—and thus not manipulable together with the continuous test and trial spaces. The framework identifies a general consistency error that bounds, under an inf–sup condition, the discrete norm of the difference between the approximation solution and an interpolant of the continuous solution. We also established improved estimates in a weaker norm, using the Aubin–Nitsche trick.

This abstract framework was applied to two popular families of numerical methods for diffusion equations: conforming and non-conforming VEM, and cell-centred or cell- and face-centred Finite Volume methods. For each of these methods, we obtained energy error estimates that accurately track the local dependencies on the diffusion tensor, through local anisotropy ratios and diffusion magnitude. In both cases, such estimates seem to be entirely new. Optimal L^2 error estimates were also established for VEM in a unified setting.

To analyse the VEM schemes for the anisotropic diffusion model, optimal approximation properties of the oblique elliptic projector on local polynomial spaces were established. These properties are of general interest to several high-order methods for diffusion equations on polytopal meshes.

The range of models and numerical techniques covered by the analysis framework goes beyond the examples above. Actually, an inspection of error bounds in some previous works show that they are based on estimations of terms that are (components of) the consistency error of our abstract setting. For example, in [19, Theorem 10], robust error estimates for the HHO method applied to an advection–diffusion–reaction model are established by bounding terms $\mathfrak{T}_1$, $\mathfrak{T}_2$ and $\mathfrak{T}_3$ that respectively correspond to the consistency errors of the diffusion component of the model, of the advection–reaction component, and of the weakly enforced (*à la* Nitsche) boundary conditions. The analysis in [19] was however carried out in an *ad-hoc* setting, and not identified as part of a wider theory as done in this paper.

Acknowledgements The work of the first author was supported by Agence Nationale de la Recherche Grants HHOMM (ANR-15-CE40-0005) and fast4hho (ANR-17-CE23-0019). The work of the second author was partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (Project Number DP170100605). Fruitful discussions with Simon Lemaire (INRIA Lille - Nord Europe) are gratefully acknowledged.

References

1. Aavatsmark, I.: An introduction to multipoint flux approximations for quadrilateral grids. Comput. Geosci. **6**(3–4), 405–432 (2002). https://doi.org/10.1023/A:10212911
2. Aavatsmark, I., Barkve, T., Bøe, O., Mannseth, T.: Discretization on unstructured grids for inhomogeneous, anisotropic media. I. Derivation of the methods. SIAM J. Sci. Comput. **19**(5), 1700–1716 (1998). https://doi.org/10.1137/S1064827595293582
3. Aavatsmark, I., Eigestad, G.T., Mallison, B.T., Nordbotten, J.M.: A compact multipoint flux approximation method with improved robustness. Numer. Methods Partial Differ. Equ. **24**(5), 1329–1360 (2008). https://doi.org/10.1002/num.20320
4. Agélas, L., Di Pietro, D.A., Droniou, J.: The G method for heterogeneous anisotropic diffusion on general meshes. ESAIM Math. Model. Numer. Anal. **44**(4), 597–625 (2010). https://doi.org/10.1051/m2an/2010021

5. Ayuso de Dios, B., Lipnikov, K., Manzini, G.: The nonconforming virtual element method. ESAIM Math. Model. Numer. Anal. **50**(3), 879–904 (2016). https://doi.org/10.1051/m2an./2015090
6. Beirão da Veiga, L., Brezzi, F., Cangiani, A., Manzini, G., Marini, L.D., Russo, A.: Basic principles of virtual element methods. Math. Models Methods Appl. Sci. (M3AS) **199**(23), 199–214 (2013). https://doi.org/10.1142/S0218202512500492
7. Beirão da Veiga, L., Brezzi, F., Marini, L.D., Russo, A.: The hitchhiker's guide to the virtual element method. Math. Models Methods Appl. Sci. **24**(8), 1541–1573 (2014). https://doi.org/10.1142/S021820251440003X
8. Beirão da Veiga, L., Brezzi, F., Marini, L.D., Russo, A.: Virtual element method for general second order elliptic problems on polygonal meshes. Math. Models Methods Appl. Sci. **26**(4), 729–750 (2016). https://doi.org/10.1142/S0218202516500160
9. Beirão da Veiga, L., Lipnikov, K., Manzini, G.: The Mimetic Finite Difference Method for Elliptic Problems. Modeling, Simulation and Applications, vol. 11. Springer, Berlin (2014). https://doi.org/10.1007/978-3-319-02663-3
10. Boffi, D., Di Pietro, D.A.: Unified formulation and analysis of mixed and primal discontinuous skeletal methods on polytopal meshes. ESAIM Math. Model. Numer. Anal. **52**(1), 1–28 (2018). https://doi.org/10.1051/m2an/2017036
11. Brenner, S.C., Guan, Q., Sung, L.-Y.: Some estimates for virtual element methods. Comput. Methods Appl. Math. **17**(4), 553–574 (2017). https://doi.org/10.1515/cmam-2017-0008
12. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, vol. 15, 3rd edn, p. xviii++397. Springer, New York (2008). https://doi.org/10.1007/978-0-387-75934-0. ISBN: 978-0-387-75933-3
13. Cangiani, A., Manzini, G., Sutton, O.J.: Conforming and nonconforming virtual element methods for elliptic problems. IMA J. Numer. Anal. **37**(3), 1317–1354 (2017). https://doi.org/10.1093/imanum/drw036
14. Chatzipantelidis, P.: Finite volume methods for elliptic PDE's: a newapproach. M2AN Math. Model. Numer. Anal. **36**(2), 307–324 (2002). https://doi.org/10.1051/m2an:2002014
15. Chou, S.-H., Li, Q.: Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: a unified approach. Math. Comput. **69**(229), 103–120 (2000). https://doi.org/10.1090/S0025-5718-99-01192-8
16. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics, vol. 40. Reprint of the 1978 Original [North-Holland, Amsterdam; MR0520174 (58 #25001)], p. xxviii+530. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2002). ISBN: 0-89871-514-8
17. Cockburn, B., Di Pietro, D.A., Ern, A.: Bridging the hybrid high-order and hybridizable discontinuous Galerkin methods. ESAIM Math. Model. Numer. Anal. **50**(3), 635–650 (2016). https://doi.org/10.1051/m2an/2015051
18. Di Pietro, D.A., Droniou, J.: A hybrid high-order method for Leray–Lions elliptic equations on general meshes. Math. Comput. **86**(307), 2159–2191 (2017). https://doi.org/10.1142/S0218202517500191
19. Di Pietro, D.A., Droniou, J., Ern, A.: A discontinuous-skeletal method for advection–diffusion–reaction on general meshes. SIAM J. Numer. Anal. **53**(5), 2135–2157 (2015). https://doi.org/10.1137/140993971
20. Di Pietro, D.A., Droniou, J., Manzini, G.: Discontinuous skeletal gradient discretisation methods on polytopalmeshes. J. Comput. Phys. **355**, 397–425 (2018). https://doi.org/10.1016/j.jcp.2017.11.018
21. Di Pietro, D.A., Ern, A.: Arbitrary-order mixed methods for heterogeneous anisotropic diffusion on general meshes. IMA J. Numer. Anal. **37**(1), 40–63 (2017). https://doi.org/10.1093/imanum/drw003
22. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Mathématiques and Applications, vol. 69. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-22980-0
23. Di Pietro, D.A., Ern, A., Lemaire, S.: An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. Comput. Methods Appl. Math. **14**(4), 461–472 (2014). https://doi.org/10.1007/978-3-319-41640-3
24. Di Pietro, D. A., Tittarelli, R.: An introduction to Hybrid High-Order methods. In: Di Pietro, D. A., Ern, A., Formaggia, L. (eds.) Numerical Methods for PDEs: State of the Art Techniques. Springer (2018). ISBN: 978-3-319-94675
25. Di Pietro, D.A., Ern, A.: Hybrid high-order methods for variable-diffusion problems on general meshes. C. R. Math. Acad. Sci. Paris **353**(1), 31–34 (2015). https://doi.org/10.1016/j.crma.2014.10.013

26. Droniou, J.: Finite volume schemes for diffusion equations: introduction to and review of modern methods. Math. Models Methods Appl. Sci. **24**(8), 1575–1619 (2014). https://doi.org/10.1142/S0218202514400041
27. Droniou, J., Eymard, R.: A mixed finite volume scheme for anisotropic diffusion problems on any grid. Numer. Math. **105**, 35–71 (2006). https://doi.org/10.1007/s00211-006-0034-1
28. Droniou, J., Eymard, R.: The asymmetric gradient discretisation method. In: Finite Volumes for Complex Applications VIII-Methods and Theoretical Aspects, vol. 199. Springer Proceedings in Mathematics and Statistics. Springer, Cham, pp. 311–319 (2017)
29. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: The Gradient Discretisation Method. Mathematics and Applications, vol. 82. Springer, p. 511 (2018). ISBN: 978-3-319-79041-1 (Softcover) 978- 3-319-79042-8 (eBook). https://doi.org/10.1007/978-3-319-79042-8. https://hal.archives-ouvertes.fr/hal-01382358
30. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. Math. Models Methods Appl. Sci. (M3AS) **20**(2), 1–31 (2010). https://doi.org/10.1142/S0218202510004222
31. Droniou, J., Nataraj, N.: Improved L2 estimate for gradient schemes and super-convergence of the TPFAfinite volume scheme. IMA J. Numer. Anal. **38**(3), 1254–1293 (2018). https://doi.org/10.1093/imanum/drx028. arxiv: 1602.07359
32. Dupont, T., Scott, R.: Polynomial approximation of functions in Sobolev spaces. Math. Comput. **34**(150), 441–463 (1980)
33. Edwards, M.G., Rogers, C.F.: A flux continuous scheme for the full tensor pressure equation. In: Proceedings of the 4th European Conference on the Mathematics of Oil Recovery, Vol. D. Røros, Norway (1994)
34. Ern, A., Guermond, J.-L.: Abstract nonconforming error estimates and application to boundary penalty methods for diffusion equations and time-harmonic Maxwell's equations. Comput. Methods Appl. Math. (2018). https://doi.org/10.1515/cmam-2017-0058
35. Ern, A., Guermond, J.-L.: Theory and Practice of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, New York (2004)
36. Ewing, R., Lazarov, R., Lin, Y.: Finite volume element approximations of nonlocal reactive flows in porous media. Numer. Methods Partial Differ. Equ. **16**(3), 285–311 (2000). https://doi.org/10.1002/(SICI)1098-2426(200005)16:3<285::AID-NUM2>3.0.CO;2-3
37. Eymard, R., Gallouët, T., Herbin, R.: Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes. SUSHI: a scheme using stabilization and hybrid interfaces. IMA J. Numer. Anal. **30**(4), 1009–1043 (2010). https://doi.org/10.1093/imanum/drn084
38. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., Lions, J.-L. (eds.) Handbook of Numerical Analysis, VII. Techniques of Scientific Computing, Part III, pp. 713–1020. North- Holland, Amsterdam (2000)
39. Gudi, T.: A new error analysis for discontinuous finite element methods for linear elliptic problems. Math. Comput. **79**(272), 2169–2189 (2010). https://doi.org/10.1090/S0025-5718-10-02360-4
40. Lipnikov, K., Manzini, G.: A high-order mimetic method on unstructured polyhedral meshes for the diffusion equation. J. Comput. Phys. **272**, 360–385 (2014). https://doi.org/10.1016/j.jcp.2014.04.021
41. Mishev, I.D.: Finite volume element methods for non-definite problems. Numer. Math. **83**(1), 161–175 (1999). https://doi.org/10.1007/s002110050443
42. Stampacchia, G.: Le probléme de Dirichlet pour les èquations elliptiques du second ordre à coefficients discontinus. Ann. Inst. Fourier (Grenoble) **15**(fasc. 1), 189–258 (1965)
43. Strang, G.: Variational crimes in the finite element method. In: The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, pp. 689–710 (Proceedings of Symposia, University Maryland, Baltimore, MD, 1972). Academic Press, New York (1972)
44. Strang, G., Fix, G.: An Analysis of the Finite Element Method, 2nd edn, p. x+402. Wellesley-Cambridge Press, Wellesley (2008)
45. Tartar, L.: Personal Communication. Dec. 26 (2015)
46. Wang, J., Ye, X.: A weak Galerkin element method for second-order elliptic problems. J. Comput. Appl. Math. **241**, 103–115 (2013). https://doi.org/10.1016/j.cam.2012.10.003