A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems

Assyr Abdulle, Gilles Vilmart

Numerische Mathematik manuscript No.
(will be inserted by the editor)

A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems

Assyr Abdulle · **Gilles Vilmart**

November 10, 2011. Received: date / Accepted: date

Abstract The effect of numerical quadrature in finite element methods for solving quasilinear elliptic problems of nonmonotone type is studied. Under similar assumption on the quadrature formula as for linear problems, optimal error estimates in the L^2 and the H^1 norms are proved. The numerical solution obtained from the finite element method with quadrature formula is shown to be unique for a sufficiently fine mesh. The analysis is valid for both simplicial and rectangular finite elements of arbitrary order. Numerical experiments corroborate the theoretical convergence rates.

Keywords nonmonotone quasilinear elliptic problem · a priori error estimates · numerical quadrature · variational crime · finite elements.

Mathematics Subject Classification (2000) 65N30 · 65M60 · 65D30

1 Introduction

The use of numerical quadrature for the practical implementation of finite element methods (FEMs), when discretizing boundary value problems, is usually required. Indeed, except in very special cases, the inner product involved in the FEM cannot be evaluated exactly and must be approximated. This introduces additional errors in the numerical method, which rates of decay have to be estimated. The control of the effects introduced by numerical quadrature is important for almost all applications of FEMs to problem in engineering and the sciences. Compared to the huge literature concerned with the analysis of FEM, the effect of numerical quadrature has only be treated in a few papers. Such results have been derived by Ciarlet and Raviart [12] and Strang [30] for second order linear elliptic equation, by Raviart [27] for

A. Abdulle, G. Vilmart
Section de Mathématiques, École Polytechnique Fédérale de Lausanne, Station 8, CH-1015 Lausanne, Switzerland
E-mail: Assyr.Abdulle@epfl.ch, Gilles.Vilmart@bretagne.ens-cachan.fr
Present address of G. Vilmart: École Normale Supérieure de Cachan, Antenne de Bretagne, av. Robert Schuman, F-35170 Bruz, France

parabolic equations and by Baker and Dougalis for second order hyperbolic equations [7]. In our paper we derive optimal a priori convergence rates in the H^1 and L^2 norm for FEMs with numerical quadrature applied to quasilinear elliptic problems of nonmonotone type. The analysis is valid for dimensions $d \leq 3$ and for simplicial or quadrilateral FEs of arbitrary order. We also show the uniqueness of the numerical solutions for a sufficiently fine FE mesh. Both the a priori convergence rates and the uniqueness results are new.

We first mention that quasilinear problems as considered in this paper are used in many applications [5]. For example, the stationary state of the Richards problems [8] used to model infiltration processes in porous media is the solution of a nonlinear nonmonotone quasilinear problem as considered in this paper (see Section 5 for a numerical example). Second, our results are also of interest in connection to the recent development of numerical homogenization methods (see for example [15, 16, 1, 2, 19] and the references therein). Indeed, such methods are based on a macroscopic solver whose bilinear form is obtained by numerical quadrature, with data recovered by microscopic solvers defined on sampling domains at the quadrature nodes [1, 2, 15]. Convergence rates for FEMs with numerical quadrature are thus essential in the analysis of numerical homogenization methods and the a priori error bounds derived in this paper allow to use an approach similar to the linear case for the analysis of nonlinear homogenization problems [3, 4].

We briefly review the literature for FEM applied to quasilinear elliptic problems of nonmonotone type. In the absence of numerical quadrature, optimal a priori error estimates in the H^1 and L^2 norms where first given by Douglas and Dupont [13]. This paper contains many ideas useful for our analysis. We also mention that Nitsche derived in [25] an error estimate for the L^∞ norm (without numerical quadrature). The analysis of FEMs with numerical quadrature for quasilinear problems started with Feistauer and Ženíšek [18], where *monotone problems* have been considered. The analysis (for piecewise linear triangular FEs) does not apply for nonmonotone problems that we consider. Nonmonotone problems have been considered by Feistauer et al. in [17], where the convergence of a FEM with numerical quadrature has been established for piecewise linear FEs. Convergence rates have not been derived in the aforementioned paper and the question of the uniqueness of a numerical solution has not been addressed. This will be discussed in the present paper for simplicial or quadrilateral FEs of arbitrary order (see Theorem 5). We note that in [17], it is also discussed the approximation problem introduced by using a curved boundary of the domain for the dimension $d = 2$; this was generalized for $d = 3$ in [23].

The paper is organized as follows. In section 2 we introduce the model problem together with the FEM based on numerical quadrature. We also state our main results. In section 3 we collect and prove several preliminary results as a preparation for the analysis of the numerical method given in section 4. Numerical examples are given in section 5. They corroborate our theoretical convergence rates and illustrate the application of the numerical method to the (stationary) Richards equation. Finally, an appendix contains the proof of technical lemmas used to derive the a priori convergence rates.

Notations Let $\Omega \subset \mathbb{R}^d$ be open and denote by $W^{s,p}(\Omega)$ the standard Sobolev spaces. We use the standard Sobolev norms $\|\cdot\|_{H^s(\Omega)}$ and $\|\cdot\|_{W^{s,p}(\Omega)}$. For $p=2$ we use the notation $H^s(\Omega)$, and $H^1_0(\Omega)$ denotes the closure in $H^1(\Omega)$ of $C^\infty_0(\Omega)$ (the space of functions of class C^∞ with compact support in Ω). Let $(\cdot,\cdot)$ denote the scalar product in $L^2(\Omega)$ or the duality between $H^{-1}(\Omega)$ and $H^1_0(\Omega)$. For a domain $K \subset \Omega$, $|K|$ denotes the measure of K. For a smooth function $a(x,u)$, we will sometimes use the notations $\partial_u a, \partial^2_u a$ or alternatively a_u, a_{uu} for the partial derivatives $\frac{\partial}{\partial u}a, \frac{\partial^2}{\partial u^2}a$.

2 Model problem and FEM with numerical quadrature

2.1 Model problem

Let Ω be a bounded polyhedron in $\mathbb{R}^d$ where $d \le 3$. We consider quasilinear elliptic problems of the form

$$\begin{aligned} -\nabla\cdot(a(x,u(x))\nabla u(x)) &= f(x) \ \text{ in } \Omega, \\ u(x) &= 0 \ \text{ on } \partial\Omega. \end{aligned} \tag{1}$$

We make the following assumptions on the tensor $a(x,s) = (a_{mn}(x,s))_{1\le m,n\le d}$

– the coefficients $a_{mn}(x,s)$ are continuous functions on $\overline{\Omega}\times\mathbb{R}$ which are uniformly Lipschitz continuous with respect to s, i.e., there exist $\Lambda_1 > 0$ such that

$$|a_{mn}(x,s_1) - a_{mn}(x,s_2)| \le \Lambda_1|s_1-s_2|, \ \forall x\in\overline{\Omega}, \forall s_1,s_2\in\mathbb{R}, \forall\ 1\le m,n\le d. \tag{2}$$

– $a(x,s)$ is uniformly elliptic and bounded, i.e., there exist $\lambda,\Lambda_0 > 0$ such that

$$\lambda\|\xi\|^2 \le a(x,s)\xi\cdot\xi, \qquad \|a(x,s)\xi\| \le \Lambda_0\|\xi\|, \qquad \forall\xi\in\mathbb{R}^d, \forall x\in\overline{\Omega}, \forall s\in\mathbb{R}. \tag{3}$$

We also assume that $f \in H^{-1}(\Omega)$. Consider the forms

$$A(z;v,w) := \int_\Omega a(x,z(x))\nabla v(x)\cdot\nabla w(x)dx, \quad \forall z,v,w\in H^1_0(\Omega), \tag{4}$$

and

$$F(w) := (f,w), \quad \forall w\in H^1_0(\Omega). \tag{5}$$

From (3), it can be shown that the bilinear form $A(z;\cdot,\cdot)$ is elliptic and bounded in $H^1_0(\Omega)$, i.e., there exist $\lambda,\Lambda_0 > 0$ such that

$$\lambda\|v\|^2_{H^1(\Omega)} \le A(z;v,v), \ \forall z,v\in H^1_0(\Omega), \tag{6}$$

$$A(z;v,w) \le \Lambda_0\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}, \ \forall z,v,w\in H^1_0(\Omega). \tag{7}$$

We can then state the weak formulation of problem (1) which reads: find $u\in H^1_0(\Omega)$ such that

$$A(u;u,w) = F(w), \qquad \forall w\in H^1_0(\Omega). \tag{8}$$

Theorem 1 *[14,22,10]. Assume* (2), (3) *and* $f\in H^{-1}(\Omega)$. *Then Problem* (8) *has a unique solution* $u\in H^1_0(\Omega)$.

Remark 1 The existence of a solution u of the weak formulation (8) of problem (1) was first shown in [13, p. 693], using a compactness argument. We refer to [10, Thm. 11.6] for a short proof of the uniqueness of the solution. In [22], the existence and the uniqueness of a weak solution of Problem (1) are shown for $f \in L^2(\Omega)$, with more general mixed Dirichlet-Neumann boundary conditions, on a bounded domain with a Lipschitz boundary. For the proof of the uniqueness, the divergence form of the differential operator is an essential ingredient. In the case of a domain Ω with a smooth boundary $\partial\Omega$, assuming the α-Hölder continuity of the right-hand side f on $\overline{\Omega}$ and $a \in C^2(\overline{\Omega} \times \mathbb{R})$, it is shown in [13] that the solution has regularity $u \in C^{2+\alpha}(\overline{\Omega})$ and that it is unique (using results from [14]).

Remark 2 Since the tensor $a(x,s)$ depends on x, and also is not proportional in general to the identity I, the classical Kirchhoff transformation (see for instance [26]) cannot be used in our study.

A comment about monotonicity A (nonlinear) form $M(\cdot,\cdot)$ defined on $H^1(\Omega) \times H^1(\Omega)$ is called a $H^1(\Omega)$-monotone if it satisfies

$$M(v, v-w) - M(w, v-w) \geq 0, \quad \forall v, w \in H^1(\Omega).$$

Notice that the form $(v,w) \mapsto A(v;v,w)$ in (4) *is not monotone* in general, so the results in [18] do not apply in our study. For instance, it is non-monotone for the tensor $a(x,u) := b(u)I$ with a differentiable scalar function b satisfying $s_0 b'(s_0) + b(s_0) < 0$ for some real s_0.

2.2 FEM with quadrature formula

In this section we present the FEM with numerical quadrature that will be used throughout the paper. We shall often use the following broken norms for scalar or vector functions v^h that are piecewise polynomial with respect to the triangulation $\mathscr{T}_h$,

$$\|v^h\|_{\bar{W}^{s,p}(\Omega)} := \Big(\sum_{K \in \mathscr{T}_h} \|v^h\|^p_{W^{s,p}(K)} \Big)^{1/p},$$
$$\|v^h\|_{\bar{H}^{s}(\Omega)} := \Big(\sum_{K \in \mathscr{T}_h} \|v^h\|^2_{H^{s}(K)} \Big)^{1/2},$$
$$\|v^h\|_{\bar{W}^{s,\infty}(\Omega)} := \max_{K \in \mathscr{T}_h} \|v^h\|^\infty_{W^{s,\infty}(K)},$$

for all $s \geq 0$ and all $1 \leq p < \infty$.

Let $\mathscr{T}_h$ be a family of partition of Ω in simplicial or quadrilateral elements K of diameter h_K and denote $h := \max_{K \in \mathscr{T}_h} h_K$. We assume that the family of triangulations is conformal and shape regular. For some results (where indicated), we will need in addition the following inverse assumption

$$\frac{h}{h_K} \leq C \ \text{ for all } K \in \mathscr{T}_h \text{ and all } \mathscr{T}_h \text{ of the family of triangulations.} \tag{9}$$

We consider the following FE spaces

$$S_0^\ell(\Omega,\mathscr{T}_h)=\{v^h\in H_0^1(\Omega);\ v^h|_K\in\mathscr{R}^\ell(K),\ \forall K\in\mathscr{T}_h\},\tag{10}$$

where $\mathscr{R}^\ell(K)$ is the space $\mathscr{P}^\ell(K)$ of polynomials on K of total degree at most ℓ if K is a simplicial FE, or the space $\mathscr{Q}^\ell(K)$ of polynomials on K of degree at most ℓ in each variables if K is a quadrilateral FE. We next consider a quadrature formula $\{x_{K_j},\omega_{K_j}\}_{j=1}^J$, where $x_{K_j}\in K$ are integration points and ω_{K_j} quadrature weights. For any element K of the triangulation, we consider a C^1-diffeomorphism F_K such that $K=F_K(\hat{K})$, where $\hat{K}$ is the reference element. For a given quadrature formula on $\hat{K}$, the quadrature weights and integration points on $K\in\mathscr{T}_h$ are given by $\omega_{K_j}=\hat{\omega}_j|\det(\partial F_K)|,\ x_{K_j}=F_K(\hat{x}_j),\ j=1,\ldots,J$. We next state the assumptions that we make on the quadrature formulas.

(Q1) $\hat{\omega}_j>0,\ j=1,\ldots,J,\quad \sum_{j\in J}\hat{\omega}_j|\nabla\hat{p}(\hat{x}_j)|^2\geq\hat{\lambda}\|\nabla\hat{p}\|^2_{L^2(\hat{K})},\ \forall\hat{p}(\hat{x})\in\mathscr{R}^\ell(\hat{K}),\ \hat{\lambda}>0$;

(Q2) $\int_{\hat{K}}\hat{p}(x)dx=\sum_{j=1}^J\hat{\omega}_j\hat{p}(\hat{x}_j),\ \forall\hat{p}(\hat{x})\in\mathscr{R}^\sigma(\hat{K})$, where $\sigma=\max(2\ell-2,\ell)$ if $\hat{K}$ is a simplicial FE, or $\sigma=\max(2\ell-1,\ell+1)$ if $\hat{K}$ is a rectangular FE.

Notice that **(Q1)**,**(Q2)** are the usual assumptions for the case of linear elliptic problems. Based on the above quadrature formulas we define for all $z^h;v^h,w^h\in S_0^\ell(\Omega,\mathscr{T}_h)$,

$$A_h(z^h;v^h,w^h)=\sum_{K\in\mathscr{T}_h}\sum_{j=1}^J\omega_{K_j}a(x_{K_j},z^h(x_{K_j}))\nabla v^h(x_{K_j})\cdot\nabla w^h(x_{K_j}).\tag{11}$$

From (3) and **(Q1)**, it can be shown that the bilinear form $A_h(z^h;\cdot,\cdot)$ is elliptic and bounded in $S_0^\ell(\Omega,\mathscr{T}_h)$, i.e., there exist $\lambda,\Lambda_0>0$ (independent of h) such that

$$\lambda\|v^h\|^2_{H^1(\Omega)}\leq A_h(z^h;v^h,v^h),\ \forall z^h,v^h\in S_0^\ell(\Omega,\mathscr{T}_h)\tag{12}$$

$$A_h(z^h;v^h,w^h)\leq\Lambda_0\|v^h\|_{H^1(\Omega)}\|w^h\|_{H^1(\Omega)},\ \forall z^h,v^h,w^h\in S_0^\ell(\Omega,\mathscr{T}_h).\tag{13}$$

The FE solution of (1) with numerical integration reads: find $u^h\in S_0^\ell(\Omega,\mathscr{T}_h)$ such that

$$A_h(u^h;u^h,w^h)=F_h(w^h)\qquad\forall w^h\in S_0^\ell(\Omega,\mathscr{T}_h),\tag{14}$$

where the linear form $F_h(\cdot)$ is an approximation of (5) obtained for example by using quadrature formulas. If one uses the same quadrature formulas for (5) as used for (11) and if **(Q2)** holds, then for $1\leq q\leq\infty$ with $\ell>d/q$, if $f\in W^{\ell,q}(\Omega)$ we have

$$|F_h(w^h)-F(w^h)|\leq Ch^\ell\|f\|_{W^{\ell,q}(\Omega)}\|w^h\|_{H^1(\Omega)},\ \forall w^h\in S_0^\ell(\Omega,\mathscr{T}_h),\tag{15}$$

and if $f\in W^{\ell+1,q}(\Omega)$, we have

$$|F_h(w^h)-F(w^h)|\leq Ch^{\ell+1}\|f\|_{W^{\ell+1,q}(\Omega)}\|w^h\|_{\bar{H}^2(\Omega)},\ \forall w^h\in S_0^\ell(\Omega,\mathscr{T}_h),\tag{16}$$

where C is independent of h (See [11, Sect. 29]).

The existence of a solution of (14) (summarized in Theorem 2) can be established using the Brouwer fixed point theorem for the nonlinear map $S_h : S_0^\ell(\Omega, \mathscr{T}_h) \to S_0^\ell(\Omega, \mathscr{T}_h)$ defined by

$$A_h(z^h; S_h z^h, w^h) = F_h(w^h),\ \forall w^h \in S_0^\ell(\Omega, \mathscr{T}_h). \tag{17}$$

Details can be found for example in [13] (see also [9]).

Theorem 2 *Assume that the bilinear form $A_h(z^h;\cdot,\cdot)$, $z^h \in S_0^\ell(\Omega, \mathscr{T}_h)$, defined in* (11) *is uniformly elliptic* (12) *and bounded* (13)*. Then, for all $h > 0$, the nonlinear problem* (14) *possesses at least one solution $u^h \in S_0^\ell(\Omega, \mathscr{T}_h)$. A solution u^h is uniformly bounded in $H_0^1(\Omega)$, i.e.*

$$\|u^h\|_{H^1(\Omega)} \le C \|f\|_{W^{1,q}(\Omega)}$$

where C is independent of h.

Remark 3 Notice that there is no smallness asumption on h in Theorem 2.

The uniqueness of a solution of (14) will also be proved along with our convergence rate estimates.

Given a solution u^h of (14) the next task is now to estimates the error $u - u^h$ where u is the unique solution of (8). The convergence $\|u - u^h\|_{H^1(\Omega)} \to 0$ for $h \to 0$ of a numerical solution of problem (12) has been given in [17, Thm. 2.7] for piecewise linear simplicial FEs. We now state in Theorem 3 below the convergence for the L^2 norm for general simplicial and quadrilateral FEs in $S_0^\ell(\Omega, \mathscr{T}_h)$. It will be used to derive our optimal convergence rates in the L^2 or H^1 norms. It can be proved using a compactness argument similar to [17, Thm. 2.6] or [13, p. 893]. For the convenience of the reader we give a short proof in the appendix.

Theorem 3 *Let u^h be a numerical solution of* (14)*. Assume that for any sequences $(v^{h_k})_{k>0}, (w^{h_k})_{k>0}$ in $S_0^\ell(\Omega, \mathscr{T}_h)$ satisfying $\|w^{h_k}\|_{H^1(\Omega)} \le C$ and $\|v^{h_k}\|_{\bar{W}^{2,\infty}(\Omega)} \le C$, where C is independent of k, we have for $h_k \to 0$,*

$$|A(w^{h_k}; w^{h_k}, v^{h_k}) - A_{h_k}(w^{h_k}; w^{h_k}, v^{h_k})| \to 0, \tag{18}$$

$$|F_{h_k}(w^{h_k}) - F(w^{h_k})| \to 0, \tag{19}$$

then $\|u - u^h\|_{L^2(\Omega)} \to 0$ for $h \to 0$.

Remark 4 In the case of linear simplicial FEs, it is shown in [17, Thm. 2.6] that Theorem 3 holds if one considers in the assumptions all sequences $(v^{h_k})_{k>0}$ bounded in $W^{1,p}(\Omega)$ for some p with $d < p \le \infty$. It is sufficient for our study to consider sequences bounded for the broken norm of $W^{2,\infty}(\Omega)$.

2.3 Main results

We can now state our main results: the uniqueness of the numerical solution and optimal a priori error estimates for the H^1 and L^2 norms.

Theorem 4 *Consider u the solution of problem* (1)*, and u^h one solution of* (14)*. Let $\ell \geq 1$. Assume (***Q1***), (***Q2***),* (2)*,* (3) *and*

$$u \in H^{\ell+1}(\Omega), \tag{20}$$

$$a_{mn} \in W^{\ell,\infty}(\Omega \times \mathbb{R}), \qquad \forall m,n = 1 \dots d, \tag{21}$$

$$f \in W^{\ell,q}(\Omega), \qquad \text{where } 1 \leq q \leq \infty, \ell > d/q. \tag{22}$$

Then, there exists a constant C_1 depending only on the domain Ω and family of FE spaces $(S_0^\ell(\Omega, \mathscr{T}_h))_{h>0}$ such that if the exact solution u satisfies

$$C_1 \Lambda_1 \lambda^{-1} \|u\|_{H^2(\Omega)} < 1, \tag{23}$$

where Λ_1, λ are the constants in (2),(3)*, then the following H^1 error estimate holds for all $h > 0$,*

$$\|u - u^h\|_{H^1(\Omega)} \leq Ch^\ell, \tag{24}$$

where C is independent of h. If in addition to the above hypotheses, (9) *holds, then there exists $h_0 > 0$ such that for all $h \leq h_0$, the solution u^h of* (14) *is unique.*

Remark 5 Notice that if the tensor $a(x,s)$ is independent of s, then $\Lambda_1 = 0$ and (23) is automatically satisfied. In that case, we retrieve in Theorem 4 the usual assumptions for linear elliptic problems [11]. Notice that the analysis in [9, Sect. 8.7] also relies on such a smallness assumption on the solution.

Assuming slightly more regularity on the solution and the tensor and (9), we can remove the smallness assumption (23), as illustrated in the following theorem. In addition, we obtain an optimal L^2 error estimate.

Theorem 5 *Consider u the solution of problem* (1)*. Let $\ell \geq 1$. Let $\mu = 0$ or 1. Assume (***Q1***), (***Q2***),* (9)*,* (63) *and*

$$u \in H^{\ell+1}(\Omega) \cap W^{1,\infty}(\Omega),$$

$$a_{mn} \in W^{\ell+\mu,\infty}(\Omega \times \mathbb{R}), \qquad \forall m,n = 1 \dots d,$$

$$f \in W^{\ell+\mu,q}(\Omega), \qquad \text{where } 1 \leq q \leq \infty, \ell > d/q.$$

In addition to (2)*,* (3)*, assume that $\partial_u a_{mn} \in W^{1,\infty}(\Omega \times \mathbb{R})$, and that the coefficients $a_{mn}(x,s)$ are twice differentiable with respect to s, with the first and second order derivatives continuous and bounded on $\overline{\Omega} \times \mathbb{R}$, for all $m,n = 1 \dots d$.*

Then there exists $h_0 > 0$ such that for all $h \leq h_0$, the solution u^h of (14) *is unique and the following H^1 and L^2 error estimates hold:*

$$\|u - u^h\|_{H^1(\Omega)} \leq Ch^\ell, \qquad \text{for } \mu = 0, 1, \tag{25}$$

$$\|u - u^h\|_{L^2(\Omega)} \leq Ch^{\ell+1}, \quad \text{for } \mu = 1. \tag{26}$$

Here, the constants C are independent of h.

Notice that the above rates of convergence in the H^1 and L^2 norms are the same as what is known in the absence of numerical quadrature [13], or for linear elliptic problems with numerical quadrature [11]. The assumption (63) is an hypothesis on the adjoint L^* of the linearized operator corresponding to (1). This hypothesis is also required to use the Aubin-Nitsche duality argument for L^2 estimates in the case of linear problems [12]. Under our assumptions on the coefficients of (1), (63) is for example automatically satisfied if the domain Ω is a convex polyhedron.

3 Preliminaries

3.1 Useful inequalities

Based on the quadrature formulas defined in Section 2.2, we consider, for v, w scalar or vector functions that are piecewise continuous with respect to the partition $\mathscr{T}_h$ of Ω, the semi-definite inner product

$$(v,w)_{\mathscr{T}_h} := \sum_{K\in\mathscr{T}_h}\sum_{j=1}^{J} \omega_{K_j} v(x_{K_j})\cdot w(x_{K_j}).$$

and the semi-norm $\|v\|_{\mathscr{T}_h,2}$ where for all $r\geq 1$ we define

$$\|v\|_{\mathscr{T}_h,r} := \Big(\sum_{K\in\mathscr{T}_h}\sum_{j=1}^{J} \omega_{K_j} (v(x_{K_j}))^r\Big)^{1/r}. \tag{27}$$

We have (Hölder)

$$|(v,w)_{\mathscr{T}_h}| \leq \|v\|_{\mathscr{T}_h,p}\|w\|_{\mathscr{T}_h,q}, \tag{28}$$

where $1/p+1/q=1$.

Notice that for v^h in a piecewise polynomial spaces (as $S_0^\ell(\Omega,\mathscr{T}_h)$), we have for all $r\geq 1$,

$$\|v^h\|_{\mathscr{T}_h,r} \leq C\|v^h\|_{L^r(\Omega)}, \tag{29}$$

where C depends on the degree of the (piecewise) polynomials, on r and the shape regularity but is independent of h. The proof of (29), that can be obtained following the lines of [27, Lemma 5] is based on a scaling argument and the equivalence of norms on a finite-dimensional space.

We shall often use the estimate

$$|(zv,w)| \leq \|z\|_{L^3(\Omega)}\|v\|_{L^6(\Omega)}\|w\|_{L^2(\Omega)},\ \forall z\in L^3(\Omega),\forall v\in L^6(\Omega),\forall w\in L^2(\Omega), \tag{30}$$

which is a consequence of the Cauchy-Schwarz and Hölder inequalities. Using the continuous inclusion $H^1(\Omega)\subset L^6(\Omega)$ for $\dim\Omega\leq 3$, the special case $z=v=w$ in (30) yields the so-called Gagliardo-Nirenberg [24] inequality,

$$\|v\|_{L^3(\Omega)} \leq C\|v\|_{L^2(\Omega)}^{1/2}\|v\|_{H^1(\Omega)}^{1/2}, \quad \forall v\in H^1(\Omega). \tag{31}$$

A discrete version of (30) holds for continuous functions on Ω,

$$|(zv,w)_h| \leq \|z\|_{\mathscr{T}_h,3}\|v\|_{\mathscr{T}_h,6}\|w\|_{\mathscr{T}_h,2} \tag{32}$$

If z^h, v^h, w^h are in piecewise polynomial spaces (as $S_0^\ell(\Omega, \mathscr{T}_h)$), then using (29) we have

$$|(z^h v^h, w^h)_h| \le C\|z^h\|_{L^3(\Omega)}\|v^h\|_{L^6(\Omega)}\|w^h\|_{L^2(\Omega)}, \tag{33}$$

where C depends on the degrees of the (piecewise) polynomials and on the exponent $r = 2, 3, 6$ in (27) (see (29)).

The following results will be often used.

Lemma 1 *Assume* (9). *Let $k \ge 1$ and $v^0 \in H^{k+1}(\Omega)$ and consider a sequence (v^h) in $S_0^\ell(\Omega, \mathscr{T}_h)$ satisfying for all h small enough,*

$$\|v^h - v^0\|_{H^1(\Omega)} \le C_0 h^k.$$

Then, for all h small enough,

$$\begin{aligned} \|v^h\|_{\bar{H}^{k+1}(\Omega)} + \|v^h\|_{\bar{W}^{k,6}(\Omega)} &\le C(\|v^0\|_{H^{k+1}(\Omega)} + C_0), \\ \|v^h\|_{\bar{W}^{k,3}(\Omega)} &\le C\|v^0\|_{H^{k+1}(\Omega)}. \end{aligned}$$

where the constant C depends only on k, the domain Ω and the finite element space $(S_0^\ell(\Omega, \mathscr{T}_h))_{h>0}$.

Proof It follows from the inverse inequality (9) that for all integers $m \ge n \ge 0$ and all $p, q \ge 1$ (see [11, Thm. 17.2])[1]

$$|v^h|_{\bar{W}^{m,q}(\Omega)} \le \frac{C}{h^{\max(d(1/p-1/q),0)+m-n}} |v^h|_{\bar{W}^{n,p}(\Omega)} \quad \forall v^h \in S_0^\ell(\Omega, \mathscr{T}_h), \tag{34}$$

where C depends on m, n, p, q, the dimension d, the domain Ω and the family of finite element spaces $(S_0^\ell(\Omega, \mathscr{T}_h))_{h>0}$. The triangle inequality $\|v^h\|_{\bar{W}^{k,q}(\Omega)} \le \|v^h - \mathscr{I}_h v_0\|_{\bar{W}^{k,q}(\Omega)} + \|\mathscr{I}_h v_0\|_{\bar{W}^{k,q}(\Omega)}$ and the inequality (38) below concludes the proof. □

3.2 Error bounds on $A_h - A$

Let $\ell \ge \ell' \ge 1$. We consider the usual nodal interpolant [11, Sect. 12] $\mathscr{I}_h : C^0(\overline{\Omega}) \to S_0^\ell(\Omega, \mathscr{T}_h)$ onto the FE space $S_0^\ell(\Omega, \mathscr{T}_h)$ defined in (10). Then, we have the following estimates (see [11, Thm. 16.2])

$$\|\mathscr{I}_h z\|_{W^{1,\infty}(\Omega)} \le C\|z\|_{W^{1,\infty}(\Omega)}, \qquad \forall z \in W^{1,\infty}(\Omega), \tag{35}$$

$$\|\mathscr{I}_h z - z\|_{W^{1,\infty}(\Omega)} \le Ch\|z\|_{W^{2,\infty}(\Omega)}, \qquad \forall z \in W^{2,\infty}(\Omega), \tag{36}$$

$$\|\mathscr{I}_h z - z\|_{H^1(\Omega)} \le Ch^{\ell'}\|z\|_{H^{\ell'+1}(\Omega)}, \qquad \forall z \in H^{\ell'+1}(\Omega), \tag{37}$$

$$\begin{aligned} &\|\mathscr{I}_h z\|_{\bar{W}^{\ell'-1,\infty}(\Omega)} + \|\mathscr{I}_h z\|_{\bar{W}^{\ell',6}(\Omega)} + \|\mathscr{I}_h z\|_{\bar{H}^{\ell'+1}(\Omega)} \\ &\le C\|z\|_{H^{\ell'+1}(\Omega)}, \qquad \forall z \in H^{\ell'+1}(\Omega). \end{aligned} \tag{38}$$

[1] Notice that (34) remains valid for $q = \infty$, replacing $1/q$ by 0 in the right-hand side (idem for p).

In our analysis, we need a priori estimates for the difference between the forms (4) and (14) (Propositions 1, 2 below). Consider for all element $K \in \mathscr{T}_h$ the quadrature error functional

$$E_K(\varphi) := \int_K \varphi(x)dx - \sum_{j=1}^{J} \omega_{K_j} \varphi(x_{K_j}), \tag{39}$$

defined for all continuous function φ on K. The next task is to estimate the quantity $|E_K(a(\cdot,z^h)\nabla v^h \cdot \nabla w^h)|$, where $a(\cdot,\cdot)$ is the tensor given in (1). Such error estimates have been derived for the linear case in [11, Thm. 28.2]. In the non-linear case, it is the purpose of the following Propositions 1, 2.

Proposition 1 *Let $\ell \geq 1$. Assume (**Q2**), $u \in H^{\ell+1}(\Omega)$. Then,*

– *for $a \in (W^{\ell,\infty}(\Omega \times \mathbb{R}))^{d\times d}$, we have for all $w^h \in S_0^\ell(\Omega, \mathscr{T}_h)$,*

$$|A_h(\mathscr{I}_h u; \mathscr{I}_h u, w^h) - A(\mathscr{I}_h u; \mathscr{I}_h u, w^h)| \leq Ch^\ell \|w^h\|_{H^1(\Omega)}, \tag{40}$$

where C depends on $\|a\|_{(W^{\ell,\infty}(\Omega\times\mathbb{R}))^{d\times d}}$ and $\|u\|_{H^{\ell+1}(\Omega)}$ but is independent of h.

– *Assume* (9). *For $a \in (W^{\ell+1,\infty}(\Omega \times \mathbb{R}))^{d\times d}$, we have for all $v^h, w^h \in S_0^\ell(\Omega, \mathscr{T}_h)$,*

$$|A_h(\Pi_h u; \Pi_h u, w^h) - A(\Pi_h u; \Pi_h u, w^h)| \leq Ch^{\ell+1}(\|w^h\|_{\bar{H}^2(\Omega)} + \|w^h\|_{W^{1,6}(\Omega)}), \tag{41}$$

where C depends on $\|a\|_{(W^{\ell+1,\infty}(\Omega\times\mathbb{R}))^{d\times d}}$ and $\|u\|_{H^{\ell+1}(\Omega)}$ but is independent of h.

Here, $\mathscr{I}_h u$ denoted the usual nodal interpolant of u on $S_0^\ell(\Omega, \mathscr{T}_h)$, while $\Pi_h u$ denotes the L^2-orthogonal projection of u on $S_0^\ell(\Omega, \mathscr{T}_h)$.

The proof of Proposition 1 relies on the following lemma which gives an estimate on each finite element $K \in S_0^\ell(\Omega, \mathscr{T}_h)$, with the proof postponed to the Appendix.

Lemma 2 *Assume that (**Q2**) holds and $a \in (W^{\ell,\infty}(\Omega \times \mathbb{R}))^{d\times d}$, then, for all $K \in \mathscr{T}_h$, and all $z, v, w \in \mathscr{R}^\ell(K)$,*

$$\begin{aligned} |E_K(a(\cdot,z)\nabla v \cdot \nabla w)| &\leq Ch_K^\ell \|a\|_{(W^{\ell,\infty}(K\times\mathbb{R}))^{d\times d}} \|\nabla w\|_{L^2(K)} \\ &\Big(\|v\|_{H^\gamma(K)}(1 + \|z\|^\ell_{W^{\ell-1,\infty}(K)}) + \|z\|_{W^{\ell,\alpha}(K)} \|\nabla v\|_{L^\beta(K)}\Big), \end{aligned} \tag{42}$$

*Assume that (**Q2**) holds and $a \in (W^{\ell+1,\infty}(\Omega \times \mathbb{R}))^{d\times d}$, then, for all $K \in \mathscr{T}_h$, and all $z, v, w \in \mathscr{R}^\ell(K)$,*

$$\begin{aligned} |E_K(a(\cdot,z)\nabla v \cdot \nabla w)| &\leq Ch_K^{\ell+1} \|a\|_{(W^{\ell+1,\infty}(K\times\mathbb{R}))^{d\times d}} \\ &\Big((1 + \|z\|^{\ell+1}_{W^{\ell-1,\infty}(K)}) \|v\|_{H^\gamma(K)} \|\nabla w\|_{H^1(K)} + \|z\|_{W^{\ell,\alpha}(K)} \|\nabla v\|_{L^\beta(K)} \|\nabla w\|_{H^1(K)} \\ &+ \|z\|_{H^\gamma(K)} \|\nabla v\|_{L^\alpha(K)} \|\nabla w\|_{L^\beta(K)} + \|z\|_{W^{\ell,\alpha}(K)} \|\nabla v\|_{H^1(K)} \|\nabla w\|_{L^\beta(K)}\Big) \end{aligned} \tag{43}$$

Here $\gamma = \ell$ if $v \in \mathscr{P}^\ell(K)$, $\gamma = \ell+1$ if $v \in \mathscr{Q}^\ell(K)$, $1 \leq \alpha, \beta \leq \infty$ with $1/\alpha + 1/\beta = 1/2$. The constants C are independent of h_K and the element K. For the case $\ell = 1$, the term $\|z\|_{W^{\ell-1,\infty}(K)}$ can be omitted in the above estimates.

Proof of Proposition 1. The proof of (40) is a consequence of (42) in Lemma 2 with $\alpha = 3, \beta = 6$. We have

$$\begin{aligned}
&|A_h(z^h; v^h, w^h) - A(z^h; v^h, w^h)| \\
&\leq C \sum_{K \in \mathscr{T}_h} h_K^\ell \|a\|_{(W^{\ell,\infty}(K\times\mathbb{R}))^{d\times d}} \|z^h\|_{W^{\ell,3}(K)} \|\nabla v^h\|_{L^6(K)} \|\nabla w^h\|_{L^2(K)} \\
&+ \sum_{K \in \mathscr{T}_h} h_K^\ell \|a\|_{(W^{\ell,\infty}(K\times\mathbb{R}))^{d\times d}} (1 + \|z^h\|^\ell_{W^{\ell-1,\infty}(K)}) \|v^h\|_{H^{\ell+1}(K)} \|\nabla w^h\|_{L^2(K)} \\
&\leq Ch^\ell \|a\|_{(W^{\ell,\infty}(\Omega\times\mathbb{R}))^{d\times d}} (\|z^h\|_{\bar{W}^{\ell,3}(\Omega)} \|\nabla v^h\|_{L^6(\Omega)} \|\nabla w^h\|_{L^2(\Omega)} \\
&+ (1 + \|z^h\|^\ell_{\bar{W}^{\ell-1,\infty}(\Omega)}) \|v^h\|_{\bar{H}^{\ell+1}(\Omega)} \|\nabla w^h\|_{L^2(\Omega)}).
\end{aligned}$$

for all $z^h, v^h, w^h \in S_0^\ell(\Omega, \mathscr{T}_h)$, where we applied for the first sum the Hölder inequality and for the second sum the Cauchy-Schwarz inequality. Finally, we take $z^h = v^h = \mathscr{I}_h u$, and we use the bound (38) to obtain (40).

The proof of (41) is a consequence of (43) in Lemma 2 and is very similar to that of (40). The main difference is we take $z^h = v^h = \Pi_h u$, where $\Pi_h u$ is the L^2-orthogonal projection of u on $S_0^\ell(\Omega, \mathscr{T}_h)$. We have $\|\Pi_h u - u\|_{L^2(\Omega)} \leq h^{\ell+1}$ and $\|\Pi_h u - u\|_{H^1(\Omega)} \leq Ch^\ell$ and we use Lemma 1. □

We shall also need the following estimate where only the first derivatives of v^h and z^h are involved in the right-hand side of (44). This is crucial for using Proposition 4 in the proof of Lemma 5, and for showing the estimate (18) of Theorem 3 in the proof of Theorem 5. Notice that for piecewise linear simplicial FEs the result follows from [17, Lemma 2.5].

Proposition 2 *Let $\ell \geq 1$. Assume (**Q2**), $a \in (W^{1,\infty}(\Omega \times \mathbb{R}))^{d\times d}$. We have for all $z^h, v^h, w^h \in S_0^\ell(\Omega, \mathscr{T}_h)$,*

$$\begin{aligned}
|A_h(z^h; v^h, w^h) - A(z^h; v^h, w^h)| \leq Ch \|\nabla v^h\|_{L^2(\Omega)} (&\|\nabla w^h\|_{\bar{H}^1(\Omega)} \\
&+ \|\nabla z^h\|_{L^\alpha(\Omega)} \|\nabla w^h\|_{L^\beta(\Omega)}),
\end{aligned} \tag{44}$$

where $1 \leq \alpha, \beta \leq \infty$ with $1/\alpha + 1/\beta = 1/2$ and C is independent of h.

The proof[2] of Proposition 2 relies on the following lemma with proof postponed to Appendix.

Lemma 3 *Let $\ell \geq 1$. If (**Q2**) holds and $a \in (W^{1,\infty}(\Omega \times \mathbb{R}))^{d\times d}$, then, for all $K \in \mathscr{T}_h$, and all $z, v, w \in \mathscr{R}^\ell(K)$,*

$$\begin{aligned}
&|E_K(a(\cdot, z)\nabla v \cdot \nabla w)| \\
&\leq Ch_K \|a\|_{(W^{1,\infty}(K\times\mathbb{R}))^{d\times d}} \|\nabla v\|_{L^2(K)} \big(\|\nabla w\|_{H^1(K)} + \|\nabla z\|_{L^\alpha(K)} \|\nabla w\|_{L^\beta(K)}\big),
\end{aligned}$$

where $1 \leq \alpha, \beta \leq \infty$ with $1/\alpha + 1/\beta = 1/2$.

[2] Notice that we need Proposition 2 for ℓ possibly larger than one. Thus, simply setting $\ell = 1$ in Proposition 1 is not sufficient.

Proof of Proposition 2. Using Lemma 3, we have

$$
\begin{aligned}
&|A_h(z^h;v^h,w^h)-A(z^h;v^h,w^h)|\\
&\le C\sum_{K\in\mathscr{T}_h} h_K\|a\|_{(W^{1,\infty}(K\times\mathbb{R}))^{d\times d}}\|\nabla v^h\|_{L^2(K)}\|\nabla w^h\|_{H^1(K)}\\
&+\sum_{K\in\mathscr{T}_h} h_K\|a\|_{(W^{1,\infty}(K\times\mathbb{R}))^{d\times d}}\|\nabla v^h\|_{L^2(K)}\|\nabla z^h\|_{L^\alpha(K)}\|\nabla w^h\|_{L^\beta(K)}\\
&\le Ch\|a\|_{(W^{1,\infty}(\Omega\times\mathbb{R}))^{d\times d}}\|\nabla v^h\|_{L^2(\Omega)}(\|\nabla w^h\|_{\bar{H}^1(\Omega)}+\|\nabla z^h\|_{L^\alpha(\Omega)}\|\nabla w^h\|_{L^\beta(\Omega)}).
\end{aligned}
$$

where we applied the Cauchy-Schwarz and Hölder inequalities. □

Similarly, we have (see the proof in Appendix)

Proposition 3 *Let $\ell\ge 1$. Assume (**Q2**), $a_u\in(W^{1,\infty}(\Omega\times\mathbb{R}))^{d\times d}$ and $u\in H^2(\Omega)\cap W^{1,\infty}(\Omega)$. Then, for all $v^h\in S_0^\ell(\Omega,\mathscr{T}_h)$, $w\in H^2(\Omega)$,*

$$(a_u(\cdot,\mathscr{I}_h u)\nabla\mathscr{I}_h u\cdot\nabla v^h,\mathscr{I}_h w)_h-(a_u(\cdot,\mathscr{I}_h u)\nabla\mathscr{I}_h u\cdot\nabla v^h,\mathscr{I}_h w)\le Ch\|v^h\|_{H^1(\Omega)}\|w\|_{H^2(\Omega)}$$

and for all $w^h\in S_0^\ell(\Omega,\mathscr{T}_h)$, $v\in H^2(\Omega)$,

$$(a_u(\cdot,\mathscr{I}_h u)\nabla\mathscr{I}_h u\cdot\nabla\mathscr{I}_h v,w^h)_h-(a_u(\cdot,\mathscr{I}_h u)\nabla\mathscr{I}_h u\cdot\nabla\mathscr{I}_h v,w^h)\le Ch\|v\|_{H^2(\Omega)}\|w^h\|_{H^1(\Omega)}$$

where C depends on a,u and is independent of h.

3.3 Finite element method with numerical quadrature for indefinite linear elliptic problems

In this section, we generalize to the case of numerical quadrature a result of Schatz [28,29] for the finite element solution of non-symmetric indefinite linear elliptic problems of the form

$$\mathscr{L}\varphi=f\ \text{ on }\Omega,\quad \varphi=0\ \text{ on }\partial\Omega, \tag{45}$$

where $\mathscr{L}\varphi:=-\nabla\cdot(a(x)\nabla\varphi)+b(x)\cdot\nabla\varphi+c(x)\varphi$, with $a\in(W^{1,\infty}(\Omega))^{d\times d}$, $b\in(L^\infty(\Omega))^d$, $c\in L^\infty(\Omega)$. We assume that the tensor $a(x)$ is uniformly elliptic and bounded, i.e. satisfies (3). We consider the associated bilinear form on $H^1(\Omega)\times H^1(\Omega)$,

$$B(v,w)=(a(x)\nabla v,\nabla w)+(b(x)\cdot\nabla v+c(x)v,w),\ \forall v,w\in H^1(\Omega). \tag{46}$$

Using the Cauchy-Schwarz and Young inequalities, we have that $B(v,w)$ satisfies the so-called Gårding inequality (with $\lambda_1,\lambda_2>0$)

$$\lambda_1\|v\|^2_{H^1(\Omega)}-\lambda_2\|v\|^2_{L^2(\Omega)}\le B(v,v),\quad \forall v\in H^1_0(\Omega), \tag{47}$$

and ($\Lambda_0>0$)

$$|B(v,w)|\le\Lambda_0\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)},\quad \forall v,w\in H^1(\Omega). \tag{48}$$

The proof of the error estimate given in Proposition 4 below for FEM relies on the Aubin-Nitsche duality argument. The use of such duality argument is instrumental in deriving the error estimates (26) (see Lemmas 5, 6).

Proposition 4 *Let $\ell \geq \ell' \geq 1$. Consider $B(\cdot,\cdot)$ defined in* (46) *and a bilinear form $B_h(\cdot,\cdot)$ defined on $S_0^\ell(\Omega,\mathscr{T}_h)\times S_0^\ell(\Omega,\mathscr{T}_h)$, satisfying also a Gårding inequality*

$$\lambda_1\|v^h\|^2_{H^1(\Omega)} - \lambda_2\|v^h\|^2_{L^2(\Omega)} \leq B_h(v^h,v^h), \quad \forall v^h \in S_0^\ell(\Omega,\mathscr{T}_h), \tag{49}$$

and for all $v \in H^{\ell'+1}(\Omega)$, $w^h \in S_0^\ell(\Omega,\mathscr{T}_h)$,

$$\begin{aligned} |B(\mathscr{I}_h v, w^h) - B_h(\mathscr{I}_h v, w^h)| &\leq Ch^{\ell'}\|v\|_{H^{\ell'+1}(\Omega)}\|w^h\|_{H^1(\Omega)}, \\ |B(w^h, \mathscr{I}_h v) - B_h(w^h, \mathscr{I}_h v)| &\leq Ch^{\ell'}\|w^h\|_{H^1(\Omega)}\|v\|_{H^{\ell'+1}(\Omega)}. \end{aligned} \tag{50}$$

Assume that for all $f \in H^{-1}(\Omega)$, the solution $\varphi \in H_0^1(\Omega)$ of problem (45) *is unique. For a fixed f, assume that the solution of* (45) *exists with regularity $\varphi \in H^{\ell'+1}(\Omega)$.*

Then, for all h small enough, the finite element problem

$$B_h(\varphi^h, v^h) = (f, v^h) \quad \forall v^h \in S_0^\ell(\Omega,\mathscr{T}_h) \tag{51}$$

possesses a unique solution $\varphi^h \in S_0^\ell(\Omega,\mathscr{T}_h)$; and φ^h satisfies the estimate

$$\|\varphi^h - \varphi\|_{H^1(\Omega)} \leq Ch^{\ell'}\|\varphi\|_{H^{\ell'+1}(\Omega)} \tag{52}$$

where C is independent of h.

Proof Due to the finite dimension of the linear system (51), to prove the uniqueness of φ^h, it suffices to show that the homogeneous system has a unique solution. This will be proved if we can show the a priori estimate (52).
We define $\xi^h = \varphi^h - \mathscr{I}_h\varphi$ and claim (as proved below) that for all $\eta > 0$ there exists $h_0 > 0$ such that for all $h \leq h_0$, we have[3]

$$\|\xi^h\|_{L^2(\Omega)} \leq \eta\|\xi^h\|_{H^1(\Omega)} + Ch^{\ell'}\|\varphi\|_{H^{\ell'+1}(\Omega)}, \tag{53}$$

where C is independent of h. We choose η such that $\lambda_1 - 2\eta^2\lambda_2 > 0$. Using the Gårding inequality (49) and (53), we obtain

$$\|\xi^h\|^2_{H^1(\Omega)} \leq C(h^{2\ell'}\|\varphi\|^2_{H^{\ell'+1}(\Omega)} + B_h(\xi^h,\xi^h)).$$

Using (48) and (50) we obtain

$$\begin{aligned} B_h(\xi^h,\xi^h) &= B(\varphi - \mathscr{I}_h\varphi, \xi^h) + (B(\mathscr{I}_h\varphi,\xi^h) - B_h(\mathscr{I}_h\varphi,\xi^h)) \\ &\leq Ch^{\ell'}\|\varphi\|_{H^{\ell'+1}(\Omega)}\|\xi^h\|_{H^1(\Omega)} \end{aligned}$$

where we used also (37). Applying the Young inequality, we deduce for all $\mu > 0$,

$$\|\xi^h\|^2_{H^1(\Omega)} \leq C(1+1/\mu)h^{2\ell'}\|\varphi\|^2_{H^{\ell'+1}(\Omega)} + C\mu\|\xi^h\|^2_{H^1(\Omega)}.$$

We choose μ such that $1 - C\mu > 0$, and using the triangular inequality and (37), we deduce (52).

[3] Notice that one cannot simply let the parameter η tend to zero in (53) because h_0 depends on η.

It remains to prove the above claim (53). Since by assumption the kernel of the operator $\mathscr{L} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is zero, using the Gårding inequality (47), it follows from the Fredholm alternative (see [21]) that the adjoint operator $\mathscr{L}^* : H_0^1(\Omega) \to H^{-1}(\Omega)$ is an isomorphism and for all $g \in H^{-1}(\Omega)$, the adjoint problem

$$B(v, \varphi^*) = (g, v), \quad \forall v \in H_0^1(\Omega), \tag{54}$$

has a unique solution $\varphi^* \in H_0^1(\Omega)$. Now, let $Y = \{g \in L^2(\Omega) \ ; \ \|g\|_{L^2(\Omega)} = 1\}$ and recall that

$$\|\xi^h\|_{L^2(\Omega)} = \sup_{g \in Y} (\xi^h, g). \tag{55}$$

For $g \in Y$, we consider $w_g \in H_0^1(\Omega)$ the unique solution of the adjoint problem (54) with right-hand side g. We take in (54) the test function $v = \xi^h$ and using (48), (50), we observe for $\chi \in H^{\ell'+1}(\Omega)$ that

$$\begin{aligned}
(\xi^h, g) &= B(\xi^h, w_g) \\
&= B(\xi^h, w_g - \mathscr{I}_h\chi) + \big(B(\varphi^h, \mathscr{I}_h\chi) - B_h(\varphi^h, \mathscr{I}_h\chi)\big) \\
&+ \big(B_h(\varphi^h, \mathscr{I}_h\chi) - B(\varphi, \mathscr{I}_h\chi)\big) + B(\varphi - \mathscr{I}_h\varphi, \mathscr{I}_h\chi) \\
&\leq C\|\xi^h\|_{H^1(\Omega)}\|w_g - \mathscr{I}_h\chi\|_{H^1(\Omega)} + Ch^{\ell'}\|\varphi^h\|_{H^1(\Omega)}\|\chi\|_{H^{\ell'+1}(\Omega)} \\
&+ C\|\varphi - \mathscr{I}_h\varphi\|_{H^1(\Omega)}\|\mathscr{I}_h\chi\|_{H^1(\Omega)}.
\end{aligned}$$

Using $\|\varphi^h\|_{H^1(\Omega)} \leq \|\xi^h\|_{H^1(\Omega)} + \|\mathscr{I}_h\varphi\|_{H^1(\Omega)}$ and (37), we obtain for all $\chi \in H^{\ell'+1}(\Omega)$,

$$\begin{aligned}
(\xi^h, g) &\leq C\|\xi^h\|_{H^1(\Omega)}(\|w_g - \mathscr{I}_h\chi\|_{H^1(\Omega)} + h^{\ell'}\|\chi\|_{H^{\ell'+1}(\Omega)}) \\
&+ Ch^{\ell'}\|\chi\|_{H^{\ell'+1}(\Omega)}\|\varphi\|_{H^{\ell'+1}(\Omega)}.
\end{aligned} \tag{56}$$

Since the injection $L^2(\Omega) \subset H^{-1}(\Omega)$ is compact, the set Y is compact in $H^{-1}(\Omega)$. Using that $\mathscr{L}^* : H_0^1(\Omega) \to H^{-1}(\Omega)$ is an isomorphism, we obtain that the set

$$Z := \{z \in H_0^1(\Omega) \ ; \ B(v, z) = (g, v), \forall v \in H_0^1(\Omega), \ g \in Y\},$$

is compact in $H^1(\Omega)$. For a fixed $\eta > 0$, the set Z is therefore contained in the union of a finite family of balls with centers $z_i \in Z$ and radius $\eta/3$ for the $H^1(\Omega)$ norm. Taking any $z \in Z$, there exists i_0 such that $\|z - z_{i_0}\|_{H^1(\Omega)} \leq \eta/3$. Since $H^{\ell'+1}(\Omega)$ is dense in $H^1(\Omega)$, for all i there exists $\bar{z}_i \in H^{\ell'+1}(\Omega)$ such that $\|z_i - \bar{z}_i\|_{H^1(\Omega)} \leq \eta/3$. Then, we have

$$\begin{aligned}
\|z - \mathscr{I}_h\bar{z}_{i_0}\|_{H^1(\Omega)} &\leq \|z - z_{i_0}\|_{H^1(\Omega)} + \|z_{i_0} - \bar{z}_{i_0}\|_{H^1(\Omega)} + \|z_{i_0} - \mathscr{I}_h\bar{z}_{i_0}\|_{H^1(\Omega)} \\
&\leq \eta/3 + \eta/3 + C_{i_0}h^{\ell'}\|\bar{z}_{i_0}\|_{H^{\ell'+1}(\Omega)}
\end{aligned}$$

where we use (37). We take $\chi := \bar{z}_{i_0}$. Notice that $\|\chi\|_{H^{\ell'+1}(\Omega)} \leq C(\eta)$ with $C(\eta)$ independent of z, i_0 and h. Taking h small enough so that $C_i h^{\ell'} C(\eta) \leq \eta/3$ for all i,

we obtain that for all $\eta > 0$ there exists $h_0 > 0$ such that for all $h \le h_0$ and for all $z \in Z$,

$$\text{there exists } \chi \in H^{\ell'+1}(\Omega) \text{ such that } \|\chi\|_{H^{\ell'+1}(\Omega)} \le C(\eta),\ \|z - \mathscr{I}_h\chi\|_{H^1(\Omega)} \le \eta. \tag{57}$$

Using (55), (56), and (57) with $z = w_g$, we deduce that (53) holds for all $h \le h_0$. □

Remark 6 In Proposition 4, notice that we did not use neither an assumption of the form (63) on the adjoint $\mathscr{L}^*$ of the operator $\mathscr{L}$ in (45), nor the inequality (9). In fact, we will use Proposition 4 in the proof of Lemma 5 only for the special case $\ell' = 1$. If for the case $\ell' = 1$, we add the regularity assumption (63) on $\mathscr{L}^*$ (or e.g., the assumption that Ω is a convex polyhedron) then the end of the proof of Proposition 4 can be simplified as follows: for all $g \in Y$ we have $w_g \in H^2(\Omega)$ with $\|w_g\|_{H^2(\Omega)} \le C\|g\|_{L^2(\Omega)}$; thus, in (56) one can simply consider $\chi := w_g$ and use (37).

4 A priori analysis

Lemma 4 *If the hypotheses of Theorem 5 are satisfied, then for all $h > 0$,*

$$\|u - u^h\|_{H^1(\Omega)} \le C(h^\ell + \|u - u^h\|_{L^2(\Omega)}), \tag{58}$$

where C is independent of h.

Proof Let $\xi^h = u^h - v^h$ with $v^h = \mathscr{I}_h u$. Using (12), we have

$$\begin{aligned}\lambda\|\xi^h\|^2_{H^1(\Omega)} \le A_h(u^h; u^h - v^h, \xi^h) &= A_h(u^h; u^h, \xi^h) - A(u; u, \xi^h)\\ &+ A(u; u - v^h, \xi^h)\\ &+ A(u; v^h, \xi^h) - A(v^h; v^h, \xi^h)\\ &+ A(v^h; v^h, \xi^h) - A_h(v^h; v^h, \xi^h)\\ &+ A_h(v^h; v^h, \xi^h) - A_h(u^h; v^h, \xi^h).\end{aligned}$$

We now bound each of the five above terms. For the first term using (8), (14) and (15) we have

$$|A_h(u^h; u^h, \xi^h) - A(u; u, \xi^h)| = |F_h(\xi^h) - F(\xi^h)| \le Ch^\ell.$$

For the second term using (7) and (37) yields

$$A(u; u - v^h, \xi^h) \le Ch^\ell\|\xi^h\|_{H^1(\Omega)}.$$

For the third term using (2),(30), (37), (38) and the inequality $\|u - v^h\|_{L^3(\Omega)} \le C\|u - v^h\|_{H^1(\Omega)}$ we obtain

$$\begin{aligned}|A(u; v^h, \xi^h) - A(v^h; v^h, \xi^h)| &\le \|(a(\cdot, u) - a(\cdot, v^h))\nabla v^h\|_{L^2(\Omega)}\|\nabla\xi^h\|_{L^2(\Omega)}\\ &\le C\|u - v^h\|_{L^3(\Omega)}\|v^h\|_{W^{1,6}(\Omega)}\|\nabla\xi^h\|_{L^2(\Omega)}\\ &\le Ch^\ell\|\xi^h\|_{H^1(\Omega)}.\end{aligned}$$

Similarly for the fifth term using (32) gives

$$\begin{aligned}|A_h(v^h;v^h,\xi^h)-A_h(u^h;v^h,\xi^h)| &\le \|(a(\cdot,v^h)-a(\cdot,u^h))\nabla v^h\|_{\mathscr{T}_h,2}\|\nabla\xi^h\|_{\mathscr{T}_h,2}\\ &\le C\|\xi^h\|_{\mathscr{T}_h,3}\|\nabla v^h\|_{\mathscr{T}_h,6}\|\nabla\xi^h\|_{\mathscr{T}_h,2}\\ &\le C\|v^h\|_{W^{1,6}(\Omega)}\|\xi^h\|_{L^3(\Omega)}\|\nabla\xi^h\|_{L^2(\Omega)}. \qquad (59)\end{aligned}$$

For the fourth term we use Proposition 1. We obtain

$$\|\xi^h\|_{H^1(\Omega)} \le C(h^\ell + \|\xi^h\|_{L^3(\Omega)}), \qquad (60)$$

where we used (38) in the inequality (59). Using the Gagliardo-Nirenberg inequality (31) and the Young inequality, we have

$$\|\xi^h\|_{L^3(\Omega)} \le C\eta^{-1}\|\xi^h\|_{L^2(\Omega)} + C\eta\|\xi^h\|_{H^1(\Omega)},$$

for all $\eta > 0$. Choosing η small enough, this together with (60) and the triangular inequalities $\|u-u^h\| \le \|u-\mathscr{I}_h u\| + \|\xi^h\|$, $\|\xi^h\| \le \|u-u^h\| + \|u-\mathscr{I}_h u\|$ (respectively for the H^1 and L^2 norms), and (37) yields the desired estimate (58). □

Proof of Theorem 4. Inspecting the proof of Lemma 4 reveals, using $\|\xi^h\|_{L^3(\Omega)} \le C\|\xi^h\|_{H^1(\Omega)}$ in (60),

$$\|u-u^h\|_{H^1(\Omega)} \le Ch^\ell + C_1\|u-u^h\|_{H^1(\Omega)},$$

with C independent of h and $C_1 = C_2\Lambda_1\lambda^{-1}\|u\|_{H^2(\Omega)}$, where Λ_1, λ are the constants in (2),(3), and the constant C_2 depends only on Ω and the FE space $(S_0^\ell(\Omega,\mathscr{T}_h))_{h>0}$. Then, if we assume that $C_1 < 1$, we immediately obtain the estimate (24).

Assuming such smallness hypothesis on u, we can also prove the uniqueness of u^h for all h small enough as follows. Let (u^h) and $(\widetilde{u}^h)$ be two sequences of solutions of (14). We show that $\xi^h = \widetilde{u}^h - u^h$ is zero for all h small enough. Using (12) and (32), we have, similarly to (59),

$$\begin{aligned}\lambda\|\xi^h\|_{H^1(\Omega)} \le A_h(\widetilde{u}^h;\xi^h,\xi^h) &= ((a(\cdot,u^h)-a(\cdot,\widetilde{u}^h))\nabla u^h,\nabla\xi^h)_h\\ &\le C\Lambda_1\|\xi^h\|_{L^6(\Omega)}\|u^h\|_{W^{1,3}(\Omega)}\|\xi^h\|_{H^1(\Omega)}.\end{aligned}$$

Using Lemma 1 and $\|\xi^h\|_{L^6(\Omega)} \le C\|\xi^h\|_{H^1(\Omega)}$ ($\dim\Omega \le 3$), we obtain for all $h \le h_0$,

$$\|\xi^h\|_{H^1(\Omega)} \le C_0\Lambda_1\lambda^{-1}\|u\|_{H^2(\Omega)}\|\xi^h\|_{H^1(\Omega)}.$$

If one assumes $C_0\Lambda_1\lambda^{-1}\|u\|_{H^2(\Omega)} < 1$ in the above inequality, then $\xi^h = 0$, which implies the uniqueness of u^h. □

For deriving the L^2 error estimate (26), we consider the operator obtained by linearizing (4) and its adjoint

$$L\varphi = -\nabla\cdot(a(\cdot,u)\nabla\varphi + \varphi a_u(\cdot,u)\nabla u), \qquad (61)$$
$$L^*\varphi = -\nabla\cdot(a(\cdot,u)^T\nabla\varphi) + a_u(\cdot,u)\nabla u\cdot\nabla\varphi. \qquad (62)$$

It has been shown in [13] that these linear operators play an important role. We assume here that L^* satisfies

$$\|\varphi\|_{H^2(\Omega)} \le C(\|L^*\varphi\|_{L^2(\Omega)} + \|\varphi\|_{H^1(\Omega)}), \quad \text{for all } \varphi \in H^2(\Omega)\cap H_0^1(\Omega). \tag{63}$$

We recall here that (63) is also required for L^2 estimates in the case of linear problems [12], and that it is automatically satisfied if the domain is a convex polyhedron.

We consider the bilinear form corresponding to L^* and its discrete counterpart (linearized at $\mathscr{I}_h u$) obtained by numerical quadrature

$$B(v,w) := (a(\cdot,u)\nabla w, \nabla v) + (a_u(\cdot,u)\nabla u\cdot\nabla v, w), \quad \forall v,w \in H_0^1(\Omega), \tag{64}$$

$$\begin{aligned} B_h(v^h,w^h) :=\ & (a(\cdot,\mathscr{I}_h u)\nabla w^h, \nabla v^h)_h \\ &+ (a_u(\cdot,\mathscr{I}_h u)\nabla \mathscr{I}_h u\cdot\nabla v^h, w^h)_h, \quad \forall\, v^h,w^h \in S_0^\ell(\Omega,\mathscr{T}_h). \end{aligned} \tag{65}$$

For $\xi \in L^2(\Omega)$, we then seek $\varphi \in H_0^1(\Omega)$, $\varphi^h \in S_0^\ell(\Omega,\mathscr{T}_h)$ such that

$$B(\varphi,w) = (\xi,w), \quad \forall w \in H_0^1(\Omega), \tag{66}$$

$$B_h(\varphi^h,w^h) = (\xi,w^h), \quad \forall w^h \in S_0^\ell(\Omega,\mathscr{T}_h). \tag{67}$$

Lemma 5 *Assume the hypotheses of Theorem 5 are satisfied. Then, for $\xi \in L^2(\Omega)$ and for all h small enough, the problems* (66) *and* (67) *have unique solutions $\varphi \in H^2(\Omega)$, $\varphi^h \in S_0^\ell(\Omega,\mathscr{T}_h)$. They satisfy*

$$\|\varphi - \varphi^h\|_{H^1(\Omega)} \le Ch\|\xi\|_{L^2(\Omega)}, \tag{68}$$

$$\|\varphi^h\|_{\tilde{H}^2(\Omega)} + \|\varphi^h\|_{W^{1,6}(\Omega)} \le C\|\xi\|_{L^2(\Omega)}, \tag{69}$$

where C is independent of h.

Proof We show that Proposition 4 applies with $\ell' = 1$ to the operator $\mathscr{L} = L^\star$, with the bilinear forms (64) and (65). Using (70) below, this proves (68). Lemma 1 next yields the estimate (69) for φ^h.

Using the assumption $u \in W^{1,\infty}(\Omega)$ and the Cauchy-Schwarz inequality, we obtain that the bilinear form $B(\cdot,\cdot)$ satisfies the bound (48), and the Gårding inequalities (47), (49) are obtained using (35), (3) and the Young inequality. Notice that $B(\cdot,\cdot)$ is the bilinear form associated to the operator L^* defined in (62). Since the operator $L : H_0^1(\Omega) \to H^{-1}(\Omega)$ in (62) is in divergence form, it can be shown (see [14] and also [20, Corollary 8.2]) that L is injective. Since the Gårding inequality (47) is satisfied by $B(\cdot,\cdot)$, using the Fredholm alternative, this implies (see [21]) that the operator $L^* : H_0^1(\Omega) \to H^{-1}(\Omega)$ is an isomorphism. Next, from (63), we have the estimate

$$\|\varphi\|_{H^2(\Omega)} \le C\|\xi\|_{L^2(\Omega)}. \tag{70}$$

It remains to prove (50) (with $\ell' = 1$). Consider the following bilinear form,

$$\begin{aligned} \overline{B}_h(v^h,w^h) :=\ & (a(\cdot,\mathscr{I}_h u)\nabla w^h, \nabla v^h) \\ &+ (a_u(\cdot,\mathscr{I}_h u)\nabla \mathscr{I}_h u\cdot\nabla v^h, w^h), \quad \forall v^h,w^h \in S_0^\ell(\Omega,\mathscr{T}_h). \end{aligned}$$

Using

$$a_u(\cdot,\mathscr{I}_h u)\nabla\mathscr{I}_h u - a_u(\cdot,u)\nabla u = (a_u(\cdot,\mathscr{I}_h u) - a_u(\cdot,u))\nabla\mathscr{I}_h u + a_u(\cdot,u)\nabla(\mathscr{I}_h u - u), \tag{71}$$

(35) and the Hölder inequality (33), we obtain

$$\begin{aligned}|B(\mathscr{I}_h v, w^h) - \overline{B}_h(\mathscr{I}_h v, w^h)| &\le C\|\mathscr{I}_h u - u\|_{L^6(\Omega)}\|\nabla\mathscr{I}_h v\|_{L^3(\Omega)}\|w^h\|_{H^1(\Omega)}\\ &\quad + C\|\mathscr{I}_h u - u\|_{H^1(\Omega)}\|\nabla\mathscr{I}_h v\|_{L^3(\Omega)}\|w^h\|_{L^6(\Omega)}\\ &\le Ch^\ell\|v\|_{H^2(\Omega)}\|w^h\|_{H^1(\Omega)}\\ &\le Ch\|v\|_{H^2(\Omega)}\|w^h\|_{H^1(\Omega)}\end{aligned}$$

where we used the continuous injection $H^1(\Omega)\subset L^6(\Omega)$ and (37). Similarly, we have

$$\begin{aligned}|B(v^h,\mathscr{I}_h w) - \overline{B}_h(v^h,\mathscr{I}_h w)| &\le C\|\mathscr{I}_h u - u\|_{L^6(\Omega)}\|\nabla v^h\|_{L^2(\Omega)}\|\mathscr{I}_h w\|_{W^{1,3}(\Omega)}\\ &\quad + C\|\mathscr{I}_h u - u\|_{H^1(\Omega)}\|\nabla v^h\|_{H^1(\Omega)}\|\mathscr{I}_h w\|_{L^\infty(\Omega)}\\ &\le Ch\|v^h\|_{H^1(\Omega)}\|w\|_{H^2(\Omega)}.\end{aligned}$$

We finally show that (50) with $\ell' = 1$ holds with B replaced by $\overline{B}_h$. Indeed, for the first term in $B_h(\cdot,\cdot), \overline{B}_h(\cdot,\cdot)$, we apply Proposition 2 with $\alpha=\infty, \beta = 2$, and the same proposition with tensor a replaced by a^T, and we use (35) for $z = u$. For the second term we apply Proposition 3. This proves (50) and concludes the proof of Lemma 5. □

Lemma 6 *Assume the hypotheses of Theorem 5 are satisfied. Then,*

– *for $\mu = 0$, we have for all h small enough,*

$$\|u - u^h\|_{L^2(\Omega)} \le C(h^\ell + \|u - u^h\|^2_{H^1(\Omega)}), \tag{72}$$

– *for $\mu = 1$, we have for all h small enough,*

$$\|u - u^h\|_{L^2(\Omega)} \le C(h^{\ell+1} + \|u - u^h\|^2_{H^1(\Omega)}), \tag{73}$$

where C is independent of h.

Proof Let $v^h \in S^\ell_0(\Omega,\mathscr{T}_h)$ and $\xi^h = v^h - u^h$. Let φ, φ^h be the solutions of (66), (67) respectively, with right-hand side ξ^h. We have:

$$\begin{aligned}\|\xi^h\|^2_{L^2(\Omega)} &= B_h(\varphi^h,\xi^h)\\ &= A_h(v^h;v^h,\varphi^h) - A_h(v^h;u^h,\varphi^h) + (\xi^h a_u(\cdot,v^h)\nabla v^h,\nabla\varphi^h)_h.\end{aligned}$$

A short computation using integration by parts shows that

$$\begin{aligned}&- A_h(v^h;u^h,\varphi^h) + (\xi^h a_u(\cdot,v^h)\nabla v^h,\nabla\varphi^h)_h\\ &= A_h(u^h,u^h,\varphi^h) + (\xi^h\bar{a}_u\nabla\xi^h - \bar{a}_{uu}(\xi^h)^2\nabla v^h,\nabla\varphi^h)_h\end{aligned}$$

where

$$\overline{a}_u(x) := \int_0^1 a_u(x, v^h(x) - t\xi^h(x))dt,$$
$$\overline{a}_{uu}(x) := \int_0^1 (1-t)a_{uu}(x, v^h(x) - t\xi^h(x))dt.$$

Thus we obtain

$$\|\xi^h\|^2_{L^2(\Omega)} = A_h(v^h; v^h, \varphi^h) - A_h(u^h, u^h, \varphi^h) + (\xi^h \overline{a}_u \nabla \xi^h - \overline{a}_{uu}(\xi^h)^2 \nabla v^h, \nabla \varphi^h)_h. \tag{74}$$

Using (32), the boundedness of a_u, a_{uu} on $\overline{\Omega} \times \mathbb{R}$ and Sobolev embeddings, we have

$$\begin{aligned}(\xi^h \overline{a}_u \nabla \xi^h - \overline{a}_{uu}(\xi^h)^2 \nabla v^h, \nabla \varphi^h)_h &= (\overline{a}_u \nabla \xi^h - \overline{a}_{uu}\xi^h \nabla v^h, \xi^h \nabla \varphi^h)_h \\ &\leq C\Big(\big(\|\nabla \xi^h\|_{\mathscr{T}_h,2} + \|\xi^h \nabla v^h\|_{\mathscr{T}_h,2}\big)\|\xi^h \nabla \varphi^h\|_{\mathscr{T}_h,2}\Big) \\ &\leq C(1 + \|v^h\|_{W^{1,6}(\Omega)})\|\xi^h\|_{H^1(\Omega)}\|\xi^h\|_{L^3(\Omega)}\|\varphi^h\|_{W^{1,6}(\Omega)}.\end{aligned}$$

The first term in (74) can be written as

$$\begin{aligned}I := A_h(v^h, v^h, \varphi^h) - A_h(u^h, u^h, \varphi^h) &= A_h(v^h, v^h, \varphi^h) - A(v^h, v^h, \varphi^h) \\ &+ A(v^h, v^h, \varphi^h) - A(u, v^h, \varphi^h) \\ &+ A(u, v^h - u, \varphi^h - \varphi) \\ &+ A(u, v^h - u, \varphi) \\ &+ A(u, u, \varphi^h) - A_h(u^h, u^h, \varphi^h).\end{aligned}$$

We now distinguish two cases to bound the above quantity I.

- For the case $\mu = 0$, we take $v^h = \mathscr{I}_h u$. Using (8), (14), (37), (15), (40), (69), we obtain similarly to the proof of Lemma 4, $I \leq C\|\xi^h\|_{L^2(\Omega)} h^\ell$.
- For the case $\mu = 1$, we take $v^h = \Pi_h u$ equal to the L^2-orthogonal projection of u on the finite element space $S_0^\ell(\Omega, \mathscr{T}_h)$. We have $\|\Pi_h u - u\|_{L^2(\Omega)} \leq Ch^{\ell+1}$ and $\|\Pi_h u - u\|_{H^1(\Omega)} \leq Ch^\ell$. Using (68) we obtain

  $$A(u, \Pi_h u - u, \varphi^h - \varphi) \leq C\|\Pi_h u - u\|_{H^1(\Omega)}\|\varphi^h - \varphi\|_{H^1(\Omega)} \leq Ch^{\ell+1}\|\varphi\|_{H^2(\Omega)}.$$

  Using Green's formula yields

  $$A(u, \Pi_h u - u, \varphi) \leq C\|\Pi_h u - u\|_{L^2(\Omega)}\|\varphi\|_{H^2(\Omega)} \leq h^{\ell+1}\|\varphi\|_{H^2(\Omega)}$$

  Using (8), (14), (16), (41) and (69) we deduce $I \leq C\|\xi^h\|_{L^2(\Omega)} h^{\ell+1}$.

Using (69) and $\|v^h\|_{W^{1,6}(\Omega)} \leq C\|u\|_{H^2(\Omega)}$ for $v^h = \mathscr{I}_h u$ or $v^h = \Pi_h u$, we obtain

$$\|\xi^h\|_{L^2(\Omega)} \leq C(h^{\ell+\mu} + \|\xi^h\|_{H^1(\Omega)}\|\xi^h\|_{L^3(\Omega)}) \leq C(h^{\ell+\mu} + \|\xi^h\|^2_{H^1(\Omega)}).$$

Finally the triangle inequality $\|u - u^h\| \leq \|\xi^h\| + \|v^h - u\|$ with the L^2 and H^1 norms, respectively, gives the estimates (72), (73). □

Proof of Theorem 5. We first prove the H^1 estimate (25) and then the L^2 estimate (26). We postpone to the end of Section 4.1 the proof of the uniqueness of the numerical solution u^h.
i) *Proof of the a-priori estimate* (25).
We know from Theorem 2 that a numerical solution u^h exists for all h. Substituting (72) of Lemma 6 into (58) of Lemma 4, we obtain that for all $h \le h_1$ any solution u^h satisfies an inequality of the form

$$\|u-u^h\|_{H^1(\Omega)} \le C(h^\ell + \|u-u^h\|^2_{H^1(\Omega)}),$$

with some constant C, or equivalently,

$$(1-C\|u-u^h\|_{H^1(\Omega)})\|u-u^h\|_{H^1(\Omega)} \le Ch^\ell. \tag{75}$$

From Theorem 3 together with Proposition 2 ($\alpha = 2$, $\beta = \infty$) and (15), we have that $\|u^h-u\|_{L^2(\Omega)} \to 0$ for $h \to 0$. Using Lemma 4, we deduce

$$\|u^h-u\|_{H^1(\Omega)} \to 0 \text{ for } h \to 0.$$

Then there exists h_2 such that for all $h \le h_2$, $1-C\|u^h-u\|_{H^1(\Omega)} \ge 1/2$. Finally we set $h_0 = \min(h_1,h_2)$ and the proof of (25) is complete.
ii) *Proof of the a-priori estimate* (26).
The L^2 estimate (26) is an immediate consequence of the H^1 estimate (25) and (73) in Lemma 6. □

4.1 Newton's method

Consider for all $z^h \in S^\ell_0(\Omega,\mathscr{T}_h)$ the bilinear form $N_h(z^h;\cdot,\cdot)$ defined on $S^\ell_0(\Omega,\mathscr{T}_h) \times S^\ell_0(\Omega,\mathscr{T}_h)$ by

$$N_h(z^h;v^h,w^h) := (a(\cdot,z^h)\nabla v^h,\nabla w^h)_h + (v^h a_u(\cdot,z^h)\nabla z^h,\nabla w^h)_h.$$

The Newton method for approximating u^h by a sequence (z^h_k) in $S^\ell_0(\Omega,\mathscr{T}_h)$ can be written as

$$N_h(z^h_k;z^h_{k+1}-z^h_k,v^h) = F_h(v^h) - A_h(z^h_k;z^h_k,v^h), \quad \forall v^h \in S^\ell_0(\Omega,\mathscr{T}_h), \tag{76}$$

where $z^h_0 \in S^\ell_0(\Omega,\mathscr{T}_h)$ is an initial guess.

In this section, we show that under the hypotheses of Theorem 5, the Newton method (76) can be used to compute the numerical solution u^h of the nonlinear system (14). We also prove the uniqueness of the finite element solution u^h of (14) for all h small enough. This generalizes the results in [13] to the case of numerical quadrature.

Consider for all h the quantity

$$\sigma_h = \sup_{v^h \in S^\ell_0(\Omega,\mathscr{T}_h)} \frac{\|v^h\|_{L^\infty(\Omega)}}{\|v^h\|_{H^1(\Omega)}}.$$

Using (9), one can show the estimates

$$\sigma_h \le C(1+|\ln h|)^{1/2} \ \text{ for } d=2, \qquad \sigma_h \le Ch^{-1/2} \ \text{ for } d=3,$$

where C is independent of h. The above estimates are a consequence of the inverse inequality (34) with $m=n=0$, $q=\infty$ and the continuous injection $H^1(\Omega)\subset L^p(\Omega)$ with $p=6$ for $d=3$ and with all $1\le p<\infty$ for $d=2$. For $d=1$, we simply have $\sigma_h \le C$. Notice that for all dimensions $d\le 3$, we have $h\sigma_h \to 0$ for $h\to 0$.

To prove that the Newton method (76) is well defined and converges, the following lemma is a crucial ingredient.

Lemma 7 *Let $\tau>0$. Under assumptions of Theorem 5, there exist $h_0,\delta>0$ such that if $0<h\le h_0$, and $z^h\in S_0^\ell(\Omega,\mathscr{T}_h)$ with*

$$\|z^h\|_{W^{1,6}(\Omega)} \le \tau \quad \textit{and} \quad \sigma_h\|z^h-\mathscr{I}_h u\|_{H^1(\Omega)} \le \delta,$$

then for all linear form G on $S_0^\ell(\Omega,\mathscr{T}_h)$, there exists one and only one solution $v^h\in S_0^\ell(\Omega,\mathscr{T}_h)$ of

$$N_h(z^h;v^h,w^h)=G(w^h), \quad \forall w^h\in S_0^\ell(\Omega,\mathscr{T}_h). \tag{77}$$

Moreover, v^h satisfies

$$\|v^h\|_{H^1(\Omega)} \le C\|G\|_{H^{-1}(\Omega)} \tag{78}$$

where we write $\|G\|_{H^{-1}(\Omega)}=\sup_{w^h\in S_0^\ell(\Omega,\mathscr{T}_h)}|G(w^h)|/\|w^h\|_{H^1(\Omega)}$, and C is a constant independent of h and z^h.

Proof It is sufficient to prove (78), since it implies that the solution is unique and hence exists in the finite-dimensional space $S_0^\ell(\Omega,\mathscr{T}_h)$. Assume that v^h is a solution of (77). Using (12), (33) and (31), we have

$$\begin{aligned}
\lambda\|v^h\|^2_{H^1(\Omega)} &\le A_h(z^h;v^h,v^h)=G(v^h)-(v^h a_u(\cdot,z^h)\nabla z^h,\nabla v^h)_h\\
&\le (\|G\|_{H^{-1}(\Omega)}+C\|a_u(\cdot,z^h)\nabla z^h\|_{L^6(\Omega)}\|v^h\|_{L^3(\Omega)})\|v^h\|_{H^1(\Omega)}\\
&\le (\|G\|_{H^{-1}(\Omega)}+C\tau\|v^h\|^{1/2}_{L^2(\Omega)}\|v^h\|^{1/2}_{H^1(\Omega)})\|v^h\|_{H^1(\Omega)}.
\end{aligned}$$

From the Young inequality, we deduce

$$\|v^h\|_{H^1(\Omega)} \le C(\|G\|_{H^{-1}(\Omega)}+\|v^h\|_{L^2(\Omega)}). \tag{79}$$

Next, applying Lemma 5, with $\xi=v$ in (67), let φ^h be the solution for h small enough of

$$N_h(\mathscr{I}_h u;w^h,\varphi^h)=(v^h,w^h) \quad \forall w^h\in S_0^\ell(\Omega,\mathscr{T}_h);$$

it satisfies the bound

$$\|\varphi^h\|_{H^1(\Omega)} \le C\|v^h\|_{L^2(\Omega)}. \tag{80}$$

We obtain using an identity similar to (71) and the Cauchy-Schwarz inequality,

$$\begin{aligned}\|v^h\|^2_{L^2(\Omega)} &= N_h(\mathscr{I}_h u; v^h, \varphi^h)\\ &= G(\varphi^h) + N_h(\mathscr{I}_h u; v^h, \varphi^h) - N_h(z^h; v^h, \varphi^h)\\ &\le (\|G\|_{H^{-1}(\Omega)} + C\|\mathscr{I}_h u - z^h\|_{L^\infty(\Omega)}\|v^h\|_{H^1(\Omega)}\\ &+ C\|v^h\|_{L^\infty(\Omega)}\|\mathscr{I}_h u - z^h\|_{H^1(\Omega)})\|\varphi^h\|_{H^1(\Omega)}.\end{aligned}$$

Using (80), we deduce

$$\begin{aligned}\|v^h\|_{L^2(\Omega)} &\le C(\|G\|_{H^{-1}(\Omega)} + 2\sigma_h\|\mathscr{I}_h u - z^h\|_{H^1(\Omega)}\|v^h\|_{H^1(\Omega)})\\ &\le C(\|G\|_{H^{-1}(\Omega)} + \delta\|v^h\|_{H^1(\Omega)})\end{aligned}$$

Substituting into (79), we obtain

$$(1 - C\delta)\|v^h\|_{H^1(\Omega)} \le C\|G\|_{H^{-1}(\Omega)}.$$

We choose $\delta > 0$ so that $1 - C\delta > 0$ which concludes the proof. □

We may now state in the following theorem that the Newton method (76) is well defined and converges. This results generalizes to the case of numerical quadrature the result of [13, Thm. 2].

Theorem 6 *Consider u^h a solution of* (14)*. Under assumptions of Theorem 5, there exist $h_0, \delta > 0$ such that if $h \le h_0$ and $\sigma_h\|z_0^h - u^h\|_{H^1(\Omega)} \le \delta$, then the sequence (z_k^h) for the Newton method* (76) *is well defined, and $e_k = \|z_k^h - u^h\|_{H^1(\Omega)}$ is a decreasing sequence that converges quadratically to* 0 *for $k \to \infty$, i.e.*

$$e_{k+1} \le C\sigma_h e_k^2, \tag{81}$$

where C is a constant independent of h, k.

Proof The proof is a consequence of Lemma 7 and is obtained following the lines of the proof of [13, Thm. 2]. For the convenience of the reader, a detailed proof is given in the Appendix. □

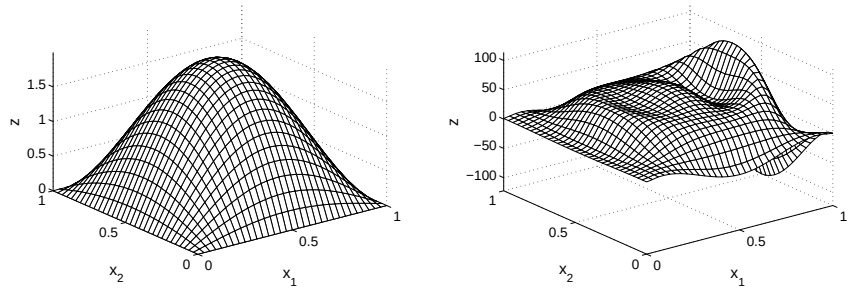
Using Theorem 6, we may now show the uniqueness of the numerical solution u^h of (14) for all H small enough.

Proof of Theorem 5.
iii) uniqueness of the numerical solution.
We know from Theorem 2 that a solution of (14) exists for all h. Consider two solutions $u^h, \widetilde{u}^h \in S_0^\ell(\Omega, \mathscr{T}_h)$ of (14). Using (25), there exists $h_1 > 0$ (independent of the choice of $u^h, \widetilde{u}^h$) such that

$$\text{for all } h \le h_1, \quad \|u^h - u\|_{H^1(\Omega)} \le Ch^\ell \text{ and } \|\widetilde{u}^h - u\|_{H^1(\Omega)} \le Ch^\ell.$$

This yields

$$\sigma_h\|\widetilde{u}^h - u^h\|_{H^1(\Omega)} \le \sigma_h\|\widetilde{u}^h - u\|_{H^1(\Omega)} + \sigma_h\|u^h - u\|_{H^1(\Omega)} \le 2C\sigma_h h^\ell \to 0 \text{ for } h \to 0.$$



(a) Solution u^h with mesh size 32×32.

(b) L^2–projection of the source f on the finite element space with mesh size 32×32.

Fig. 1 Problem (84)-(85).

Thus, we have $\sigma_h \|\widetilde{u}^h - u^h\|_{H^1(\Omega)} \leq \delta$ for all $h \leq h_2$ for some $h_2 > 0$. Then, applying Theorem 6 with initial guess $z_0^h = \widetilde{u}^h$, we have that the sequence $(z_k^h)_{k\geq 0}$ of the Newton method is well defined by (76), and $\|z_k^h - u^h\|_{H^1(\Omega)} \to 0$ for $k \to \infty$. Since z_k^h is in fact independent of k (because $\widetilde{u}^h$ solves (14)), we obtain $\widetilde{u}^h = u^h$ for all $h \leq h_0 := \min(h_1, h_2)$. □

5 Numerical experiments

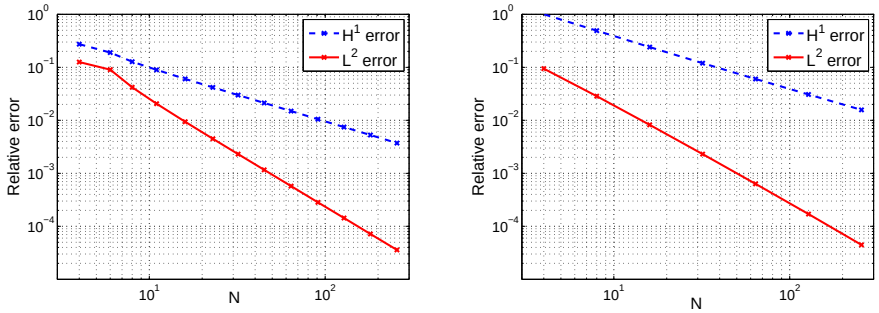
In this section, we present two test problems in dimension two to illustrate numerically that the H^1 and L^2 estimates between the finite element solution and the exact solution in Theorem 5 are sharp.

We consider the numerical resolution of non-linear problems of the form (1), with Dirichlet and also more general boundary conditions, on the square domain $\Omega = [0,1]^2$ discretized by a uniform mesh with $N \times N$ $\mathscr{Q}^1$-quadrilateral elements or a uniform mesh with $N \times N$ couples of $\mathscr{P}^1$-triangular elements which corresponds in both cases to $\mathscr{O}(N^2)$ degrees of freedom. Notice that we obtain similar results when considering either quadrilateral or triangular elements. For each quadrilateral element, we consider the Gauss quadrature formula with $J = 4$ nodes, while for triangular elements we consider the quadrature formula with $J = 1$ node at the baricenter.

Evaluating L^2 and H^1 errors. The L^2 and H^1 relative errors between the finite element solutions u^h and the exact solution u are approximated by quadrature formulas. We compute

$$e_{L^2}^2 := \|u\|_{L^2(\Omega)}^{-2} \sum_{K\in\mathscr{T}_h} \sum_{j=1}^{J} \omega_{K_j} |u^h(x_{K_j}) - u(x_{K_j})|^2, \tag{82}$$

$$e_{H^1}^2 := \|\nabla u\|_{L^2(\Omega)}^{-2} \sum_{K\in\mathscr{T}_h} \sum_{j=1}^{J} \omega_{K_j} \|\nabla u^h(x_{K_j}) - \nabla u(x_{K_j})\|^2, \tag{83}$$



(a) Problem (84)-(85). $\mathscr{Q}^1$-quadrilateral FEs. (b) Problem (88)-(87). $\mathscr{P}^1$-triangular FEs.

Fig. 2 e_{L^2} error (solid lines) and e_{H^1} error (dashed lines) as a function of the size N of a uniform $N \times N$ mesh.

so that

$$e_{L^2} \approx \frac{\|u-u^h\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}}, \qquad e_{H^1} \approx \frac{\|\nabla(u-u^h)\|_{L^2(\Omega)}}{\|\nabla u\|_{L^2(\Omega)}}.$$

Here the values $u(x_{K_j})$ and $\nabla u(x_{K_j})$ for the exact solution are computed either analytically, or approximated using a very fine mesh. In (82)-(83), for each quadrilateral element, we consider the Gauss quadrature formula with $J=4$ nodes, which is exact on $\mathscr{Q}^3(K)$, while for triangular elements we use a quadrature formula with $J=6$ nodes on each triangle (the nodes and the middle of the edges) which is exact on $\mathscr{P}^2(K)$. This way, the additional numerical quadrature error introduced in (82)-(83) is negligible compared to the accuracy of the studied finite element method.

Test problem. We first consider the non-linear problem

$$\begin{aligned} &-\nabla\cdot(a(x,u(x))\nabla u(x)) = f(x) \ \text{ in } \Omega \\ &u(x) = 0 \ \text{ on } \partial\Omega \end{aligned} \tag{84}$$

with Dirichlet boundary conditions and the anisotropic tensor

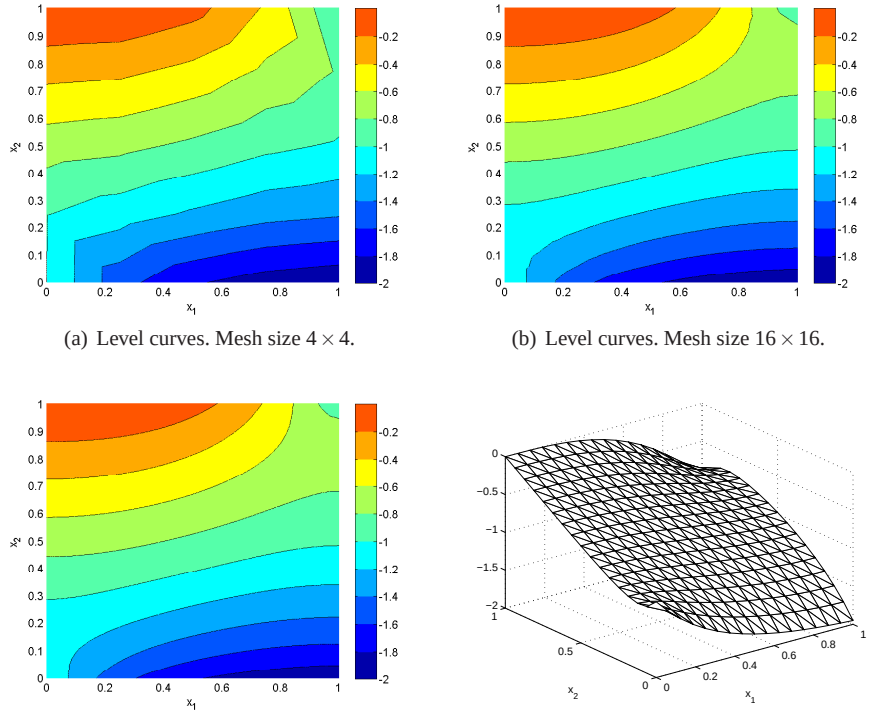
$$a(x,s) = \begin{pmatrix} 1+x_1\sin(\pi s) & 0 \\ 0 & 2+\arctan(s) \end{pmatrix}. \tag{85}$$

The source f in (84) is adjusted analytically so that the exact solution is

$$u(x) = 8\sin(\pi x_1)x_2(1-x_2), \tag{86}$$

see the numerical solution on a 32×32 mesh in Figure 1(a). We also give a graphical representation of the source f projected on the finite element space in Figure 1(b).

In Figure 2(a), we plot the L^2 and H^1 relative errors (82)-(83) for the numerical solution compared to the analytical solution (86), as a function of the size N of the meshes made of $N\times N$ elements of quadrilateral type with size $h=1/N$. As predicted by Theorem 5, we observe that the error for the H^1 norm has size $\mathscr{O}(h)$ (line of slope one), and for the L^2 norm, we observe an error of size $\mathscr{O}(h^2)$ (line of slope two).



(a) Level curves. Mesh size 4×4.

(b) Level curves. Mesh size 16×16.

(c) Level curves. Mesh size 32×32.

(d) Surface plot. Mesh size 16×16.

Fig. 3 Porous media flow problem (88)-(87). Numerical solutions on various uniform meshes with $N \times N$ couples of $\mathscr{P}^1$-triangular elements.

Concerning the Newton iterations (76), using the (artificial) initial guess $z_0^h = \Pi_h\big(10x_1(1-x_1)x_2(1-x_2)\big)$, we observe that it requires about 7 iterations to converge to u^h up to machine precision for all meshes considered in Figure 2(a).

Richards' equation for porous media flows. Consider Richards' parabolic equation for describing the fluid pressure $u(x,t)$ in an unsaturated porous medium, with permeability tensor $a(s)$ and volumetric water content Θ,

$$\frac{\partial \Theta(u)}{\partial t} - \nabla \cdot (a(u)\nabla u)) + \frac{\partial a(u)}{\partial x_2} = f$$

where x_2 is the vertical coordinate, and f corresponds to possible sources or sinks. We consider an exponential model for the permeability tensor a similar to the one in [31], which we slightly modify to simulate an anisotropic porous media,

$$a(s) = \begin{pmatrix} e^s & 0 \\ 0 & 1.1e^{1.2s} \end{pmatrix}. \tag{87}$$

For our numerical simulation, we are interested only in the stationary state (where $\partial u/\partial t = 0$). We therefore arrive at the following non-linear elliptic problem. For simplicity, we let the source term be identically zero ($f(x) \equiv 0$),

$$-\nabla \cdot (a(u(x))\nabla(u(x) - x_2)) = 0 \ \text{ in } \Omega. \tag{88}$$

We add mixed boundary conditions of Dirichlet and Neumann types,

$$\begin{aligned}
&u(x) = g_1(x) \ \text{ on } \partial\Omega_{D_1} = [0,1] \times \{1\},\\
&u(x) = g_2(x) \ \text{ on } \partial\Omega_{D_2} = [0,1] \times \{0\},\\
&n \cdot a(u(x))\nabla(u(x) - x_2) = 0 \ \text{ on } \partial\Omega_N = \{0\} \times [0,1] \cup \{1\} \times [0,1].
\end{aligned}$$

We put Neumann conditions on the left and right boundaries of the domain (n denotes the vector normal to the boundary). On the top boundary $\partial\Omega_{D_1}$ and the bottom boundary $\partial\Omega_{D_2}$, we take respectively

$$\begin{aligned}
g_1(x) &= -x_1^3,\\
g_2(x) &= -2 + e^{-3x_1}.
\end{aligned}$$

Notice that (88) is not exactly of the form (1), but can be cast into this form using the change of variable $v(x) := u(x) - x_2$. The corresponding tensor is then $a(x,s) = a(s + x_2)$. Since no analytical formula for the solution $u(x)$ is available, we compute a reference a finescale solution on a uniform mesh with 1024×1024 couples of $\mathscr{P}^1$-triangular elements (one million degrees of freedom). Here, the Newton iterations (76) converge in about 6 iterations with the initial guess $z_0^h \equiv 0$.

In Figure 3 we represent the levels curves of the the numerical solutions on uniform meshes of various sizes. Notice that the level curves for the finescale solution look nearly identical to those of the solution with $N = 32$ in Figure 3(c).

In Figure 2(b), we plot the H^1 and L^2 relative errors on various uniform meshes with $N \times N$ couples of $\mathscr{P}^1$-triangular elements with size $h = 1/N$. Similarly to the previous experiment, we observe an error of size $\mathscr{O}(h)$ in the H^1 norm as predicted by Theorem 5 (line of slope 0.97 for the meshes with $N = 64, 128, 256$), and $\mathscr{O}(h^2)$ in the L^2 norm (line of slope 1.91 for the meshes with $N = 64, 128, 256$).

6 Appendix

We give here the proofs of Theorem 3, Lemmas 2, 3, Prop. 3 and Theorem 6.

Proof of Theorem 3. As mentioned in Remark 4 we have to make sure that Theorem 3 remains true for general simplicial and quadrilateral FEs. We use a compactness argument similar to [17, Thm. 2.6] or [13, 893]. From Theorem 2, the numerical solution exists for all h, and for any choice of the numerical solution, the sequence $(u^h)_{h>0}$ is bounded in $H_0^1(\Omega)$. Since the injection $H^1(\Omega) \subset L^2(\Omega)$ is compact, from any sequence of $\{h\}$ tending to zero, there exists a subsequence $\{h_k\}$ such that for some $w \in H^1(\Omega)$, $u^{h_k} \to w$ strongly in $L^2(\Omega)$ and weakly in $H^1(\Omega)$. To conclude the proof that $\|u^h - u\|_{L^2(\Omega)} \to 0$ for $h \to 0$, it is sufficient to show that the limit is unique

with $w = u$. Let $v \in C_0^\infty(\Omega)$ and $v^{h_k} := \mathscr{I}_{h_k} v$. Using (36) yields $\|v - v^{h_k}\|_{W^{1,\infty}(\Omega)} \to 0$ for $k \to \infty$ and $\|v^{h_k}\|_{W^{2,\infty}(\Omega)} \le C\|v\|_{W^{2,\infty}(\Omega)}$. Using (8), we have

$$\begin{aligned} A(w,w,v) - F(v) &= A(w, w-u^{h_k}, v) + (A(w,u^{h_k},v) - A(u^{h_k},u^{h_k},v)) \\ &+ A(u^{h_k},u^{h_k},v-v^{h_k}) \\ &+ (A(u^{h_k},u^{h_k},v^{h_k}) - A_h(u^{h_k},u^{h_k},v^{h_k})) \\ &+ (A_h(u^{h_k},u^{h_k},v^{h_k}) - F_h(v^{h_k})) \\ &+ (F_h(v^{h_k}) - F(v^{h_k})) + F(v^{h_k} - v). \end{aligned}$$

Using (18), (19) it is straightforward that the right-hand side of the above equality tends to zero for $k \to \infty$. Thus we obtain that w satisfies

$$A(w;w,v) = F(v),\ \forall v \in C_0^\infty(\Omega),$$

and hence w is solution of (8) because $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$. Since the solution of (8) is unique (Theorem 1), we obtain $w = u$. □

Proof of Lemma 2. As the functional E_K in (39) is linear, we shall get the error estimates for the expression $E_K(a(\cdot,z)v_{(m)}w_{(n)})$, where $a(\cdot,\cdot)$ is a scalar function denoting a component of the tensor $(a_{mn}(x,s))_{1\le m,n\le d}$ and $v_{(m)}, w_{(n)}$ denote the components of $\nabla v^h|_K$, $\nabla w^h|_K$. Consider a reference element $\hat{K}$. We use the notations $\hat{a}(x,\cdot) := a(F_K(x),\cdot)$, $\hat{z}(x) := z(F_K(x))$, $\hat{v}_{(m)}(x) := v_{(m)}(F_K(x))$ and similarly for $w_{(n)}$, where $F_K : \hat{K} \to K$ is defined in Section 2.2. We have

$$E_K(a(\cdot,z)v_{(m)}w_{(n)}) = |\det \partial F_K| E_{\hat{K}}(\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}\hat{w}_{(n)}). \tag{89}$$

i) Proof of estimate (42).
We adapt the proof of [11, Thm. 28.2]. We start by applying the Bramble-Hilbert Lemma [11, Thm. 28.1] to the linear form $\hat{\varphi} \mapsto E_{\hat{K}}(\hat{\psi}\hat{\varphi})$ with $\hat{\psi}$ a polynomial on $\hat{K}$. This is a linear bounded functional on $W^{\ell,\infty}(\Omega)$ which vanishes on $\mathscr{P}^{\ell-1}(\hat{K})$ if $\hat{\psi} \in \mathscr{P}^{\ell-1}(\hat{K})$ (due to the assumption **(Q2)** for simplicial FEs) and if $\hat{\psi} \in (\mathscr{Q}^\ell(\hat{K}))'$ [4] (due to the assumption **(Q2)** for quadrilateral FEs). Thus, in either cases,

$$E_{\hat{K}}(\hat{\psi}\hat{\varphi}) \le C\|\hat{\psi}\|_{L^2(\hat{K})}|\hat{\varphi}|_{W^{\ell,\infty}(\hat{K})}, \quad \forall \hat{\varphi} \in W^{\ell,\infty}(\hat{K}). \tag{90}$$

We now take $\hat{\varphi} = \hat{a}(\cdot,\hat{z})\hat{v}_{(m)}$ and $\hat{\psi} = \hat{w}_{(n)}$, where $\hat{z}, \hat{v}, \hat{w} \in \mathscr{P}^\ell(\hat{K})$ or $\mathscr{Q}^\ell(\hat{K})$ (and thus $\hat{\psi}$ is in $\mathscr{P}^{\ell-1}(\hat{K})$ or in $(\mathscr{Q}^{\ell-1}(\hat{K})'$, respectively). We obtain

$$|E_{\hat{K}}(\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}\hat{w}_{(n)})| \le C|\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}|_{W^{\ell,\infty}(\hat{K})}\|\hat{w}_{(n)}\|_{L^2(\hat{K})}.$$

Using the equivalence of norms on a finite dimensional space of polynomials, we have

$$|\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}|_{W^{\ell,\infty}(\hat{K})} \le C\sum_{j=0}^{\ell} |\hat{a}(\cdot,\hat{z})|_{W^{j,\infty}(\hat{K})}|\hat{v}_{(m)}|_{H^{\ell-j,}(\hat{K})},$$

[4] We denote by $(\mathscr{Q}^\ell(\hat{K}))'$ the space of all derivative of polynomials belonging to $(\mathscr{Q}^\ell(\hat{K}))$

where we note that the sum stops at $\ell-1$ if $\nu \in \mathscr{P}^\ell(K)$. Using the Faà-di-Bruno formula[5], $|\hat{a}(\cdot,\hat{z})|_{W^{j,\infty}(\hat{K})}$ can be bounded by a sum of terms of the form

$$\|\partial_{\hat{x}}^{\nu}\partial_u^k \hat{a}(\cdot,\hat{z})\|_{L^\infty(\hat{K})}|\hat{z}|_{W^{r_1,\infty}(\hat{K})}\cdots|\hat{z}|_{W^{r_k,\infty}(\hat{K})} \tag{91}$$

where $\nu \in \mathbb{N}^d$ is a multi-index and $|\nu|+r_1+\ldots+r_k = j$, with $k \geq 0$ and $r_i \geq 1$ for all i. We recall the following inequalities [11, Theorems 15.1 and 15.2], for all $0 \leq j \leq \ell-1$,

$$\|\partial_{\hat{x}}^{\nu}\partial_u^k \hat{a}\|_{L^\infty(\hat{K}\times\mathbb{R})} \leq Ch_k^{|\nu|}\|\partial_{\hat{x}}^{\nu}\partial_u^k a\|_{L^\infty(K\times\mathbb{R})},\ 0 \leq k+|\nu| \leq \ell, \tag{92}$$

$$|\hat{v}|_{W^{j,q}(\hat{K})} \leq Ch_k^j|\det\partial F_K|^{-1/q}|v|_{W^{j,q}(K)},\ \forall v \in W^{j,q}(K),\ 1 \leq q < \infty, \tag{93}$$

$$|\hat{v}|_{W^{j,\infty}(\hat{K})} \leq Ch_k^j|v|_{W^{j,\infty}(K)},\ \forall v \in W^{j,\infty}(K). \tag{94}$$

Using the equivalence of norms, the term for $k=1, |\nu|=0, j=\ell$ can be bounded as

$$\begin{aligned}\|\partial_u \hat{a}(\cdot,\hat{z})\|_{L^\infty(\hat{K})}|\hat{z}|_{W^{\ell,\infty}(\hat{K})}|\hat{v}_{(m)}|_{L^\infty(\hat{K})} &\leq C|\hat{a}|_{W^{1,\infty}(\hat{K}\times\mathbb{R})}|\hat{z}|_{W^{\ell,\alpha}(\hat{K})}\|\hat{v}_{(m)}\|_{L^\beta(\hat{K})}\\ &\leq Ch^\ell|\det\partial F_K|^{-1/2}|a|_{W^{1,\infty}(K\times\mathbb{R})}|z|_{W^{\ell,\alpha}(K)}\|v_{(m)}\|_{L^\beta(K)}\end{aligned}$$

where we use (93) with $q=2,\alpha,\beta$ $(1/\alpha+1/\beta=1/2)$. For all other terms in (91) we use the estimates (92) and (94). We obtain

$$\begin{aligned}|E_{\hat{K}}(\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}w_{(n)})| \leq{}& Ch^\ell|\det\partial F_K|^{-1}\|a\|_{W^{\ell,\infty}(K\times\mathbb{R})}\|w_{(n)}\|_{L^2(K)}\\ &\Big(\|v_{(m)}\|_{H^\gamma(K)}(1+\|z\|^\ell_{W^{\ell-1,\infty}(K)})+|z|_{W^{\ell,\alpha}(K)}\|v_{(m)}\|_{L^\beta(K)}\Big),\end{aligned}$$

where $\gamma=\ell-1$ if $v\in\mathscr{P}^\ell(K)$ and $\gamma=\ell$ if $v\in\mathscr{Q}^\ell(K)$ (in the above estimate $\|z\|^\ell_{W^{\ell-1,\infty}(K)}$) can be omitted for $\ell=1$). Finally, using (89) concludes the proof of (42).

ii) Proof of estimate (43).

We adapt the proof of [12, Thm. 2]. Consider the linear operator $\hat{\Pi}_0 : L^1(\hat{K}) \to \mathscr{P}^0(\hat{K})$ defined as

$$\hat{\Pi}_0(\hat{\psi}) = \frac{1}{|\hat{K}|}\int_{\hat{K}}\hat{\psi}(\hat{x})d\hat{x}.$$

Let $\hat{\varphi} \in W^{\ell+1,\infty}(\hat{K})$ and $\hat{\psi} \in (\mathscr{R}^\ell(\hat{K}))'$. Then, we have

$$E_{\hat{K}}(\hat{\psi}\hat{\varphi}) = E_{\hat{K}}((\hat{\Pi}_0\hat{\psi})(\hat{\Pi}_0\hat{\varphi}_1)\hat{\varphi}_2)+E_{\hat{K}}((\hat{\Pi}_0\hat{\psi})(\hat{\varphi}_1-\Pi_0\hat{\varphi}_1)\hat{\varphi}_2)+E_{\hat{K}}((\hat{\psi}-\hat{\Pi}_0\hat{\psi})\hat{\varphi}). \tag{95}$$

where we set $\hat{\varphi} := \hat{\varphi}_1\hat{\varphi}_2$. We apply the Bramble-Hilbert Lemma three times, to estimate each of the above terms. Using **(Q2)**, the first term as a function of $\hat{\varphi}_2$ is a linear form which vanishes on $\mathscr{P}^\ell(\hat{K})$ (since $\hat{\Pi}_0\hat{\psi}\in\mathscr{P}^0(\hat{K})$), while the second and third terms as functions of $\hat{\varphi}_2$, $\hat{\varphi}$ respectively are linear forms which vanish on

[5] Here we use the fact that all functions in $W^{1,\infty}(\mathbb{R})$ are Lipschitz continuous. This implies that the usual chain rule applies for differentiating with respect to x the composition $a(x,z(x))$ of $s\mapsto a(x,s)$ (where s evolves in $\mathbb{R}$) with a smooth scalar function $z(x)$ defined on K.

$\mathscr{P}^{\ell-1}(\hat{K})$. We use $\|\hat{\Pi}_0\hat{\psi}\|_{L^2(\hat{K})} \le C\|\hat{\psi}\|_{L^2(\hat{K})}$ and $\|\hat{\psi}-\hat{\Pi}_0\hat{\psi}\|_{L^2(\hat{K})} \le C|\hat{\psi}|_{H^1(\hat{K})}$ (applying the Bramble-Hilbert Lemma to the linear form $\hat{\psi} \mapsto \hat{\psi}-\hat{\Pi}_0\hat{\psi}$ which vanishes on $\mathscr{P}^0(\hat{K})$). This yields

$$\begin{aligned}|E_{\hat{K}}(\hat{\psi}\hat{\varphi})| \le C(&\|\hat{\psi}\|_{L^2(\hat{K})}\|\hat{\varphi}_1\|_{L^2(\hat{K})}|\hat{\varphi}_2|_{W^{\ell+1,\infty}(\hat{K})}\\ &+ \|\hat{\psi}\|_{L^2(\hat{K})}|\hat{\varphi}_1|_{H^1(\hat{K})}|\hat{\varphi}_2|_{W^{\ell,\infty}(\hat{K})} + |\hat{\psi}|_{H^1(\hat{K})}|\hat{\varphi}|_{W^{\ell,\infty}(\hat{K})}).\end{aligned}$$

Similarly to i), we take $\hat{\varphi}_2 = \hat{a}(\cdot,\hat{z})$, $\hat{\varphi}_1 = \hat{v}_{(m)}$ and $\hat{\psi} = \hat{w}_{(n)}$. We obtain

$$\begin{aligned}|E_{\hat{K}}(\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}\hat{w}_{(n)}| \le C(&|\hat{a}(\cdot,\hat{z})|_{W^{\ell+1,\infty}(\hat{K})}\|\hat{v}_{(m)}\|_{L^2(\hat{K})}\|\hat{w}_{(n)}\|_{L^2(\hat{K})}\\ &+ |\hat{a}(\cdot,\hat{z})|_{W^{\ell,\infty}(\hat{K})}|\hat{v}_{(m)}|_{H^1(\hat{K})}\|\hat{w}_{(n)}\|_{L^2(\hat{K})}\\ &+ |\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}|_{W^{\ell,\infty}(\hat{K})}|\hat{w}_{(n)}|_{H^1(\hat{K})}).\end{aligned}$$

In the above estimate, the quantity $|\hat{a}(\cdot,\hat{z})\hat{v}_{(m)}|_{W^{\ell,\infty}(\hat{K})}$ can be bounded exactly as in the proof in i). It remains to bound the first two terms in the above estimate. We use again the Faà-di-Bruno formula for computing the derivatives up to order $\ell+1$ of $\hat{a}(\cdot,\hat{z})$. For the case where $\hat{z}$ is differentiated ℓ or $\ell+1$ times, we obtain terms of the form

$$\begin{aligned}&\|\partial_u\partial_{\hat{x}_i}\hat{a}\|_{L^\infty(\hat{K}\times\mathbb{R})}|\hat{z}|_{H^\ell(\hat{K})}\|\hat{v}_{(m)}\|_{L^\alpha(\hat{K})}\|\hat{w}_{(n)}\|_{L^\beta(\hat{K})},\\ &\|\partial_u\hat{a}\|_{L^\infty(\hat{K}\times\mathbb{R})}|\hat{z}|_{W^{\ell,\alpha}(\hat{K})}|\hat{v}_{(m)}|_{H^1(\hat{K})}\|\hat{w}_{(n)}\|_{L^\beta(\hat{K})},\\ &\|\partial_u\hat{a}\|_{L^\infty(\hat{K}\times\mathbb{R})}|\hat{z}|_{H^{\ell+1}(\hat{K})}\|\hat{v}_{(m)}\|_{L^\alpha(\hat{K})}\|\hat{w}_{(n)}\|_{L^\beta(\hat{K})},\end{aligned}$$

where we use the equivalences of norms for spaces of polynomials on $\hat{K}$. For derivatives of z of order $j<\ell$, we consider the norms $|\hat{z}|_{W^{j,\infty}(\hat{K})}$, $|\hat{v}_{(m)}|_{H^{j'}(\hat{K})}$ and $\|\hat{w}_{(n)}\|_{L^2(\hat{K})}$. We conclude the proof using (92), (93), (94) and (89), similarly to the proof in i). □

Remark 7 Notice that in the above proof ii) of (43) in Lemma 2, in the case of simplicial elements, instead of (95), one can simply consider

$$E_{\hat{K}}(\hat{\psi}\hat{\varphi}) = E_{\hat{K}}((\hat{\Pi}_0\hat{\psi})\hat{\varphi}) + E_{\hat{K}}((\hat{\psi}-\hat{\Pi}_0\hat{\psi})\hat{\varphi}),$$

then take $\hat{\psi} = \hat{w}_{(n)}$ and $\hat{\varphi} = \hat{a}(\cdot,\hat{z})\hat{v}_{(m)}$, and use $|\hat{v}_{(m)}|_{H^\ell(\hat{K})} = |\hat{v}_{(m)}|_{H^{\ell+1}(\hat{K})} = 0$. For quadrilateral elements, we had to use twice the projection $\hat{\Pi}_0$ in (95) because we have $|\hat{v}_{(m)}|_{H^{\ell+1}(\hat{K})} \neq 0$ in general.

Proof of Lemma 3. For simplicial FEs with $\ell = 1$, the result was first shown in [17, Lemma 2.5]. For general simplicial or quadrilateral FEs, we apply the Bramble-Hilbert Lemma [11, Thm. 28.1] to the functional $E_K(\hat{\psi}\cdot)$ with $\hat{\psi}$ a polynomial in $(\mathscr{R}^\ell(\hat{K}))'$. This is a linear bounded functional on $W^{1,\infty}(\Omega)$ which vanishes on $\mathscr{P}^0(\hat{K})$ (as **Q2**) holds). Thus,

$$E_K(\hat{\psi}\hat{\varphi}) \le C\|\hat{\psi}\|_{L^2(\hat{K})}|\hat{\varphi}|_{W^{1,\infty}(\hat{K})}, \quad \forall\hat{\varphi}\in W^{1,\infty}(\hat{K}). \tag{96}$$

Then we take $\hat{\psi} = \hat{v}_{(m)}$ and $\hat{\varphi} = \hat{a}(\cdot,\hat{z})\hat{w}_{(n)}$. The rest of the proof is similar to i) in the proof of Lemma 2. □

Proof of Proposition 3. We follow the lines of the proofs of Proposition 2 and Lemma 3, and take in the estimate (96) the functions $\hat{\varphi} = \hat{a}_u(\cdot,\hat{z})z_{(n)}\hat{w}$, $\hat{\psi} = \hat{v}_{(m)}$ and $\hat{\varphi} = \hat{a}_u(\cdot,\hat{z})v_{(m)}\hat{w}$, $\hat{\psi} = \hat{z}_{(n)}$, respectively. This yields for all $z^h, v^h, w^h \in S_0^\ell(\Omega,\mathscr{T}_h)$ the two estimates

$$\begin{aligned}
&(a_u(\cdot,z^h)\nabla z^h\cdot\nabla v^h, w^h)_h - (a_u(\cdot,z^h)\nabla z^h\cdot\nabla v^h, w^h)\\
&\le Ch\|v^h\|_{H^1(\Omega)}\big((1+\|z^h\|^2_{W^{1,\infty}(\Omega)})\|w^h\|_{L^2(\Omega)} + \|z^h\|_{\bar{H}^2(\Omega)}\|w^h\|_{L^\infty(\Omega)}\\
&\qquad + \|z^h\|_{W^{1,\infty}(\Omega)}\|w^h\|_{H^1(\Omega)}\big)\\
&\le Ch\|z^h\|_{W^{1,\infty}(\Omega)}\big((1+\|z^h\|_{W^{1,\infty}(\Omega)})\|v^h\|_{H^1(\Omega)}\|w^h\|_{L^2(\Omega)}\\
&\qquad + \|v^h\|_{\bar{H}^2(\Omega)}\|w^h\|_{L^2(\Omega)} + \|v^h\|_{\bar{H}^2(\Omega)}\|w^h\|_{H^1(\Omega)}\big).
\end{aligned}$$

We conclude the proof of Proposition 3 by taking $z^h =: \mathscr{I}_h u$, and $w^h := \mathscr{I}_h w$, $v^h := \mathscr{I}_h v$ respectively, and using (36), (38). □

Proof of Theorem 6. The proof follows closely the lines of the proof of [13, Thm. 2]. We first show that Lemma 7 applies with $z^h = u^h$ for all $h \le h_0$ small enough. Indeed, we have from Theorem 5 that $\sigma_h\|u^h - \mathscr{I}_h u\|_{H^1(\Omega)} \le C\sigma_h h^\ell \to 0$ for $h \to 0$, and we obtain from Lemma 1 that $\|u^h\|_{W^{1,6}(\Omega)} \le C$ where C is independent of h.

We show that given z_k^h satisfying $\sigma_h\|u^h - z_k^h\|_{H^1(\Omega)} \le \delta$, the next approximation z_{k+1}^h exists and is uniquely defined. Since $S_0^\ell(\Omega,\mathscr{T}_h)$ is finite-dimensional, it is sufficient to show for all $v^h \in S_0^\ell(\Omega,\mathscr{T}_h)$ that

$$N_h(z_k^h; v^h, w^h) = 0, \quad \forall w^h \in S_0^\ell(\Omega,\mathscr{T}_h) \tag{97}$$

implies $v^h = 0$. Indeed, using (97) we have

$$N_h(u^h; v^h, w^h) = G(w^h), \quad \forall w^h \in S_0^\ell(\Omega,\mathscr{T}_h),$$

where

$$\begin{aligned}
G(w^h) &= ((a(\cdot,u^h) - a(\cdot,z_k^h))\nabla v^h, \nabla w^h)_h\\
&\quad + (v^h\big((a_u(\cdot,u^h) - a_u(\cdot,z_k^h))\nabla(u^h) - a_u(\cdot,z_k^h)\nabla(z_k^h - u^h)\big), \nabla w^h)_h.
\end{aligned}$$

Then, $\|G\|_{H^{-1}(\Omega)} \le C\sigma_h\|u^h - z_k^h\|_{H^1(\Omega)}\|v^h\|_{H^1(\Omega)}$, and Lemma 7 yields

$$\|v^h\|_{H^1(\Omega)} \le C\sigma_h\|u^h - z_k^h\|_{H^1(\Omega)}\|v^h\|_{H^1(\Omega)} \le C\delta\|v^h\|_{H^1(\Omega)}.$$

If δ is chosen small enough, we have $C\delta < 1$ and thus $v^h = 0$.

We now show (81). We have

$$\begin{aligned}
N_h(u^h; z_{k+1}^h - u^h, w^h) &= N_h(u^h; z_k^h - u^h, w^h) + A_h(u^h; u^h, w^h) - A_h(z_k^h; z_k^h, w^h)\\
&\quad + N_h(u^h; z_{k+1}^h - z_k^h, w^h) - N_h(z_k^h; z_{k+1}^h - z_k^h, w^h)\\
&= G_1(w^h) + G_2(w^h) = G(w^h), \quad \forall w^h \in S_0^\ell(\Omega,\mathscr{T}_h),
\end{aligned}$$

where the first and second lines are equal to G_1, G_2 respectively. Then, similarly as in the proof of Lemma 6, we have

$$\begin{aligned} G_1(w^h) &= (\frac{1}{2}\widetilde{a}_{uu}(z_k^h-u^h)^2\nabla u^h + \widetilde{a}_u(z_k^h-u^h)\nabla(u^h-z_k^h),\nabla w^h)_h \\ &\le C\sigma_h e_k^2 \|w^h\|_{H^1(\Omega)}, \end{aligned}$$

where $\widetilde{a}_{uu}$ and $\widetilde{a}_u$ are certain averages of a_{uu} and a_u. Similarly,

$$\begin{aligned} G_2(w^h) &= ((a(\cdot,u^h)-a(\cdot,z_k^h))\nabla(z_{k+1}^h-z_k^h) + (z_{k+1}^h-z_k^h)(a(\cdot,z_k^h)\nabla(u^h-z_k^h)),\nabla w^h)_h \\ &+ ((z_{k+1}^h-z_k^h)\big((a(\cdot,u^h)-a(\cdot,z_k^h))\nabla u^h,\nabla w^h)_h \\ &\le C\sigma_h\|z_k^h-u^h\|_{H^1(\Omega)}(2\|z_k^h-u^h\|_{H^1(\Omega)} + \|z_{k+1}^h-u^h\|_{H^1(\Omega)})\|w^h\|_{H^1(\Omega)} \\ &+ C\sigma_h\|z_k^h-u^h\|_{H^1(\Omega)}\|u^h\|_{W^{1,6}(\Omega)}(\|z_k^h-u^h\|_{H^1(\Omega)} + \|z_{k+1}^h-u^h\|_{H^1(\Omega)})\|w^h\|_{H^1(\Omega)} \\ &\le C\sigma_h e_k(e_k+e_{k+1})\|w^h\|_{H^1(\Omega)}. \end{aligned}$$

Using Lemma 7 with $z^h = u^h$ we obtain

$$e_{k+1} \le C\sigma_h(e_k^2 + e_k e_{k+1})$$

which yields

$$(1-C\sigma_h e_k)e_{k+1} \le C\sigma_h e_k^2$$

and taking δ small enough, we have $1-C\sigma_h e_k \ge 1-C\delta > 0$ and this concludes the proof. □

References

1. A. Abdulle, *The finite element heterogeneous multiscale method: a computational strategy for multiscale PDEs*, GAKUTO Int. Ser. Math. Sci. Appl., 31 (2009), 135–184.
2. A. Abdulle, *A priori and a posteriori analysis for numerical homogenization: a unified framework*, Ser. Contemp. Appl. Math. CAM, 16, World Scientific Publishing Co. Pte. Ltd., Singapore, (2011), 280–305.
3. A. Abdulle and G. Vilmart, *The effect of numerical integration in the finite element method for nonmonotone nonlinear elliptic problems with application to numerical homogenization methods*, C. R. Acad. Sci. Paris, Ser. I, 349 (19-20) (2011), 1041–1046.
4. A. Abdulle and G. Vilmart, *Analysis of the finite element heterogeneous multiscale method for nonmonotone elliptic homogenization problems*, preprint, submitted for publication, http://infoscience.epfl.ch/record/163326.
5. H. Amann, *Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems*, Function Spaces, Differential Operators and Nonlinear Analysis (Friedrichroda, 1992), (Teubner-Texte Math., vol. 133) Teubner, Stuttgart 1993, pp. 9126.
6. N. André and M. Chipot, *Uniqueness and nonuniqueness for the approximation of quasilinear elliptic equations*, SIAM J. Numer. Anal. 33 (5) (1996), 1981–1994.
7. G.A. Baker and V.A. Dougalis, *The effect of quadrature errors on finite element approximations for second order hyperbolic equations*, SIAM J. Numer. Anal., 13 (1976), 577–598.
8. J. Bear and Y. Bachmat, *Introduction to modelling of transport phenomena in porous media*, Kluwer Academic, Dordrecht, 1991.
9. S. Brenner and R. Scott, *The mathematical theory of finite element methods.* Third edition. Texts in Applied Mathematics, 15. Springer, New York, 2008.
10. M. Chipot, *Elliptic equations: an introductory course*, Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Verlag, Basel, 2009.

11. P.G. Ciarlet, *Basic error estimates for elliptic problems*, Handb. Numer. Anal., Vol. 2, North-Holland, Amsterdam (1991), 17–351.
12. P.G. Ciarlet and P.A. Raviart, *The combined effect of curved boundaries and numerical integration in isoparametric finite element method*, in A. K Aziz (Ed), Math. Foundation of the FEM with Applications to PDE, Academic Press, New York, NY, (1972), 409–474.
13. J. Douglas, Jr. and T. Dupont, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comp., 29 (131) (1975), 689–696.
14. J. Douglas, Jr., T. Dupont, and J. Serrin, *Uniqueness and comparison theorems for nonlinear elliptic equations in divergence form*, Arch. Rational Mech. Anal., 42 (1971), 157–168.
15. W. E, B. Engquist, X. Li, W. Ren and E. Vanden-Eijnden, *Heterogeneous multiscale methods: a review*, Commun. Comput. Phys. 2 (3) (2007), 367–450.
16. B. Engquist and P.E. Souganidis, *Asymptotic and numerical homogenization*, Acta Numer. 17 (2008), 147–190.
17. M. Feistauer, M. Křížek, and V. Sobotíková, *An analysis of finite element variational crimes for a nonlinear elliptic problem of a nonmonotone type*, East-West J. Numer. Math. 1 (4) (1993), 267–285.
18. M. Feistauer and A. Ženíšek, *Finite element solution of nonlinear elliptic problems*, Numer. Math., 50 (4) (1987), 451–475.
19. M.G.D. Geers, A.G. Kouznetsova and W.A.M. Brekelmans, *Multi-scale computational homogenization: Trends and challenges*, J. Comput. Appl. Math., 234 (7) (2010), 2175–2182.
20. D. Gilbarg and N. Trudinger, *Elliptic partial differential equations of second order*. Reprint of the 1998 edition. Classics in Mathematics. Springer-Verlag, Berlin, 2001.
21. S. Hildebrandt and E. Wienholtz, *Constructive proofs of representation theorems in separable Hilbert space*, Comm. Pure Appl. Math. 17 (1964), 369–373.
22. I. Hlaváček, M. Křížek, and J. Malý, *On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type*, J. Math. Anal. Appl. 184 (1) (1994), 168–189.
23. S. Korotov, M. Křížek, *Finite element analysis of variational crimes for a quasilinear elliptic problem in 3D*, Numer. Math. 84 (4) (2000), 549–576.
24. L. Nirenberg, *On elliptic partial differential equations*, Ann. Scuola Norm. Sup. Pisa, 13 (3) (1959), 115–162.
25. J. A. Nitsche, *On L_∞-convergence of finite element approximations to the solution of a nonlinear boundary value problem*, Topics in numerical analysis, III (Proc. Roy. Irish Acad. Conf., Trinity Coll., Dublin, 1976), Academic Press, London-New York (1977) 317–325.
26. J. Poussin and J. Rappaz, *Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems*, Numer. Math., 69 (2) (1994), 213–231.
27. P.A. Raviart, *The use of numerical integration in finite element methods for solving parabolic equations*, in Miller, J. J. H. (ed.), Topics in Numerical Analysis,
28. A. H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp. 28 (1974) 959–962.
29. A. H. Schatz and J. P. Wang, *Some new error estimates for Ritz-Galerkin methods with minimal regularity assumptions*. Math. Comp. 65 (213) (1996) 19–27.
30. G. Strang, *Variational crimes in the finite element method*, in A. K Aziz (Ed), Math. Foundation of the FEM with Applications to PDE, Academic Press, New York, NY, (1972), 689–710
31. A.W. Warrick, *Time-dependent linearized infiltration: III. strip and disc sources*, Soil. Sci. Soc. Am. J., 40 (1976), 639-643.