Numerische Mathematik (2022) 150:975–992
https://doi.org/10.1007/s00211-022-01282-3

Numerische Mathematik

Stability of mixed FEMs for non-selfadjoint indefinite second-order linear elliptic PDEs

C. Carstensen[1,2] · Neela Nataraj[2] · Amiya K. Pani[2]

Received: 27 October 2020 / Revised: 1 March 2022 / Accepted: 2 March 2022 /
Published online: 1 April 2022

Abstract

For a well-posed non-selfadjoint indefinite second-order linear elliptic PDE with general coefficients $\mathbf{A}$, $\mathbf{b}$, γ in L^∞ and symmetric and uniformly positive definite coefficient matrix $\mathbf{A}$, this paper proves that mixed finite element problems are uniquely solvable and the discrete solutions are uniformly bounded, whenever the underlying shape-regular triangulation is sufficiently fine. This applies to the Raviart-Thomas and Brezzi-Douglas-Marini finite element families of any order and in any space dimension and leads to the best-approximation estimate in $H(\text{div}) \times L^2$ as well as in in $L^2 \times L^2$ up to oscillations. This generalises earlier contributions for piecewise Lipschitz continuous coefficients to L^∞ coefficients. The compactness argument of Schatz and Wang for the displacement-oriented problem does *not* apply immediately to the mixed formulation in $H(\text{div}) \times L^2$. But it allows the uniform approximation of some L^2 contributions and can be combined with a recent L^2 best-approximation result from the medius analysis. This technique circumvents any regularity assumption and the application of a Fortin interpolation operator.

Mathematics Subject Classification 65N12 · 65N15 · 65N30

✉ C. Carstensen
cc@math.hu-berlin.de

Neela Nataraj
neela@math.iitb.ac.in

Amiya K. Pani
akp@math.iitb.ac.in

1 Humboldt-Universität zu Berlin, Berlin, Germany

2 Department of Mathematics, Indian Institute of Technology Bombay, Powai,Mumbai 400076, India

1 Introduction

This section introduces the non-selfadjoint indefinite second-order linear elliptic PDE and its mixed formulations. A brief review of earlier results is followed by the assertion of the stability and the best-approximation results.

1.1 Non-selfadjoint indefinite second-order linear elliptic PDEs

The strong formulations for second-order elliptic problems with coefficients $\mathbf{A}$, $\mathbf{b}$, γ componentwise in $L^\infty(\Omega)$ and $f \in L^2(\Omega)$ read $\mathcal{L}_j u_j = f$ a.e. in a polyhedral bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$ with homogeneous Dirichlet boundary condition $u_j = 0$ on $\partial\Omega$ for $j = 1, 2$ and any dimension $n \geq 2$. For all $v \in H_0^1(\Omega)$, the two differential operators (referred to as conservative resp. divergence form throughout this paper) read

$$\mathcal{L}_1 v := -\nabla \cdot (\mathbf{A}\nabla v + v\,\mathbf{b}) + \gamma\, v \quad \text{and} \quad \mathcal{L}_2 v := -\nabla \cdot (\mathbf{A}\nabla v) + \mathbf{b} \cdot \nabla v + \gamma\, v. \tag{1.1}$$

The assumption on ellipticity means that the $n \times n$ coefficient matrix $\mathbf{A}(x)$ is symmetric and positive definite with eigenvalues in one universal compact interval of positive reals for a.e. $x \in \Omega$. This makes $\mathcal{L}_1, \mathcal{L}_2 : H_0^1(\Omega) \to H^{-1}(\Omega)$ Fredholm operators of index zero and their weak formulations $a(v, w) := \langle \mathcal{L}_1 v, w\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle \mathcal{L}_2 w, v\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}$, for all $v, w \in H_0^1(\Omega)$, are dual to each other in the duality bracket $\langle \bullet, \bullet, \rangle H^{-1}(\Omega) \times H_0^1(\Omega)$ of $H^{-1}(\Omega)$, the dual of $H_0^1(\Omega)$.

Throughout this paper, zero eigenvalues are excluded and the kernel (of one of these operators) $\mathcal{L}_j$ is supposed to be trivial, so that $\mathcal{L}_1$ and $\mathcal{L}_2$ are bijections. It is known from the theory of bilinear forms in reflexive Banach spaces [2, 3] that this implies well-posedness and the continuous inf-sup condition (e.g., when $H_0^1(\Omega)$ is endowed with the norm $\|\nabla \bullet\|$)

$$0 < \alpha := \inf_{v \in H_0^1(\Omega)\setminus\{0\}} \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{a(v, w)}{\|\nabla v\|\,\|\nabla w\|}. \tag{1.2}$$

The inf-sup constant is the same for the original and the dual problem; $a(v, w)$ could be replaced by $a(w, v)$ with the same α. The finite element error analysis is enormously simplified under additional conditions on the coefficients that lead to an ellipticity of $a(\bullet, \bullet)$ and allow an application of the Lax-Milgram lemma [2–4]. The present situation of a general non-selfadjoint indefinite second-order linear elliptic PDE avoids any of those assumptions and examines coefficients in L^∞, which satisfy the following.

Assumption (A) There exist two global constants $0 < \underline{\alpha} \leqslant \overline{\alpha} < \infty$ such that $\mathbf{A} \in L^\infty(\Omega; \mathbb{R}^{n\times n})$ satisfies $\underline{\alpha} \leqslant \lambda_1(\mathbf{A}(x)) \leqslant \cdots \leqslant \lambda_n(\mathbf{A}(x)) \leqslant \overline{\alpha}$ for the eigenvalues $\lambda_1(\mathbf{A}(x)) \leqslant \cdots \leqslant \lambda_n(\mathbf{A}(x))$ of the SPD $\mathbf{A}(x)$ for a.e. $x \in \Omega$. The functions $\mathbf{b}, \mathbf{b}_1, \mathbf{b}_2 \in L^\infty(\Omega; \mathbb{R}^n)$ and $\gamma \in L^\infty(\Omega)$ are componentwise bounded in the bounded polyhedral Lipschitz domain $\Omega \subset \mathbb{R}^n$.

Given the various applications to porous media and ground-water flow with rough and oscillating coefficients merely bounded in a well-posed PDE, this contribution gives an affirmative answer to the fundamental question whether the mixed finite element method be used (and then is stable and provides best-approximation property at least for fine triangulations).

1.2 Earlier contributions

For conforming finite element discretizations and sufficiently small mesh sizes, [17] establishes the existence and uniqueness of conforming finite element solutions under assumption (A). The mixed formulation for the conservation (resp. divergence) equation $\mathcal{L}_1 u = f$ (resp. $\mathcal{L}_2 u = f$) introduces the flux variable $\sigma = -\mathbf{A}\nabla u - u\,\mathbf{b}$ (resp. $\sigma = -\mathbf{A}\nabla u$) and seeks the solution $x = (\sigma, u) \in H$ to

$$b(x, y) = (f, v)_{L^2(\Omega)} \quad \text{for all } y = (\tau, v) \in H := H(\mathrm{div}, \Omega) \times L^2(\Omega) \tag{1.3}$$

with $\mathbf{b}_1 := \mathbf{A}^{-1}\mathbf{b}$, $\mathbf{b}_2 := 0$ (resp. $\mathbf{b}_1 := 0$, $\mathbf{b}_2 := \mathbf{b}\cdot\mathbf{A}^{-1}$) and $(\sigma, \tau)_{\mathbf{A}^{-1}} := (\mathbf{A}^{-1}\sigma, \tau)_{L^2(\Omega)}$ in

$$\begin{aligned} b(x, y) = (\sigma, \tau)_{\mathbf{A}^{-1}} - (u, \mathrm{div}\,\tau)_{L^2(\Omega)} + (v, \mathrm{div}\,\sigma)_{L^2(\Omega)} \\ + (u, \mathbf{b}_1\cdot\tau)_{L^2(\Omega)} - (v, \mathbf{b}_2\cdot\sigma)_{L^2(\Omega)} + (\gamma\, u, v)_{L^2(\Omega)}. \end{aligned} \tag{1.4}$$

The equivalence to the boundary value problems associated with the linear differential operators in (1.1) and their well-posedness on the continuous level can be found in [5,Sect. 2]. This implies the continuous inf-sup conditions [2, 3]

$$0 < \beta := \inf_{x\in H\setminus\{0\}} \sup_{y\in H\setminus\{0\}} \frac{b(x, y)}{\|x\|_H \|y\|_H} = \inf_{y\in H\setminus\{0\}} \sup_{x\in H\setminus\{0\}} \frac{b(x, y)}{\|x\|_H \|y\|_H}. \tag{1.5}$$

The existence and uniqueness of discrete solutions and optimal L^2 error estimates were introduced in [9] for sufficiently fine triangulations in two and three space dimensions under high regularity assumptions, where the pair (σ_h, u_h) is approximated in $RT_k(\mathcal{T}) \times P_k(\mathcal{T})$ with the Raviart-Thomas (RT) for $2D$ (resp. Raviart-Thomas-Nedelec for $3D$) finite elements. Global L^∞ and global L^2 and negative norm estimates for the conservation form were discussed in [10, 14] for smooth coefficients.

Provided the coefficients $\mathbf{A}$ and $\mathbf{b}$ are Lipschitz continuous, γ is piecewise Lipschitz continuous and H^2 regularity of the adjoint system, an interesting convergence phenomenon for the BDM finite element family is clarified in the fairly general framework of [8].

Let $M_k(\mathcal{T})$ be any RT or BDM finite element space of degree $k \in \mathbb{N}_0$ and define the discrete space $V(\mathcal{T}) := M_k(\mathcal{T}) \times P_k(\mathcal{T}) \subset H$ based on a shape-regular triangulation $\mathcal{T}$ with mesh-sizes $\leqslant \delta$, written $\mathcal{T} \in \mathbb{T}(\delta)$.

In case $\mathbf{A}$ and $\mathbf{b}$ are globally Lipschitz continuous and γ is piecewise Lipschitz continuous, the convergence results in [8] also establish stability in the sense

$$0 < \beta_0 \leqslant \inf_{\mathcal{T}\in\mathbb{T}(\delta)} \inf_{x_h\in V(\mathcal{T})\setminus\{0\}} \sup_{y_h\in V(\mathcal{T})\setminus\{0\}} \frac{b(x_h, y_h)}{\|x_h\|_H \|y_h\|_H} (=: \beta_h) \tag{1.6}$$

for some positive δ and β_0. With extra work and refined arguments along the lines of [8], but with reduced elliptic regularity and solution $u_j \in H_0^1(\Omega) \cap H^{1+s}(\Omega)$ to $\mathcal{L}_j u_j = f$ for some $s > 0$. Those arguments are *not* valid under Assumption (A).

Modern trends in the mathematics of mixed finite element schemes include local stable projections with commuting properties [11–13]; those techniques do not seem to allow the proof of discrete stability and best-approximation under assumption (A).

Piecewise Lipschitz continuous coefficients with regularity in H^{1+s} (for some positive s) lead in [5] to stability for the lowest-order RT FEM. The equivalence to nonconforming Crouzex-Raviart finite elements holds more generally [1] and the combination with the arguments from [17] and [5] might lead to stability results under the assumption (A) for more examples. In comparison, the methodology of this paper provides stability for any degree k and any dimension n (RT and BDM merely serve as popular model examples).

1.3 Contribution of this paper

Under the Assumption (A) and for any RT or BDM finite element space $V(\mathcal{T}) := M_k(\mathcal{T}) \times P_k(\mathcal{T}) \subset H := H(\mathrm{div}, \Omega) \times L^2(\Omega)$ of degree $k \in \mathbb{N}_0$ [2–4], the discrete stability (1.6) is established for small mesh-sizes, where either $\mathbf{b}_1 := \mathbf{A}^{-1}\mathbf{b}$ and $\mathbf{b}_2 := 0$ or $\mathbf{b}_1 := 0$ and $\mathbf{b}_2 := \mathbf{b} \cdot \mathbf{A}^{-1}$ in (1.4).

Theorem 1.1 (discrete stability) *For each (positive) $\beta_0 < \beta$ with β from* (1.5)*, there exists $\delta > 0$ such that* (1.6) *holds.*

This theorem implies [2, 3] that the mixed finite element problems for the RT and the BDM finite element families of any degree k and in any space dimension n are (i) uniquely solvable, (ii) uniformly bounded in H, and (iii) fullfil quasi-optimal error estimates in the norm of H, whenever the underlying shape-regular triangulation is sufficiently fine.

Theorem 1.1 and the tools of this paper lead to L^2 best-approximation up to oscillations.

Theorem 1.2 *(L^2 best approximation) Suppose $\delta > 0$ satisfies* (1.6) *with $b(\bullet, \bullet)$ defined for general $\mathbf{b}_1, \mathbf{b}_2 \in L^\infty(\Omega; \mathbb{R}^n)$ under Assumption (A). Assume $\mathcal{T} \in \mathbb{T}(\delta)$ and that $x := (\sigma, u) \in H$ (resp. $x_h \equiv (\sigma_h, u_h) \in V_h := V(\mathcal{T}) := M_k(\mathcal{T}) \times P_k(\mathcal{T})$) satisfy $b(x - x_h, y_h) = 0$ for all $y_h \in V_h$. Then the following results (a) and (b) hold.*
(a) There exists a positive constant C_1, which exclusively depends on $\beta_0 > 0$, the L^∞ norms of (all the components of) $\mathbf{A}^{1/2}\mathbf{b}_1$, $\mathbf{A}^{1/2}\mathbf{b}_2$, and γ, as well as on the shape-regularity of $\mathbb{T}$, such that the piecewise mesh size $h_\mathcal{T}$ in $\mathcal{T}$ and the L^2 projection Π_k onto $P_k(\mathcal{T})$ satisfy

$$C_1^{-1} \left(\|\sigma - \sigma_h\|_{\mathbf{A}^{-1}} + \|u - u_h\|\right) \leqslant \min_{\tau_h\in M_k(\mathcal{T})} \|\sigma - \tau_h\|_{\mathbf{A}^{-1}} + \|u - \Pi_k u\| + \|h_\mathcal{T}(1 - \Pi_k)\,\mathrm{div}\,\sigma\|.$$

(b) Suppose $\mathbf{b}_1 = 0$ *and that the scalar* $\gamma(x)$ *is Lipschitz continuous in* $x \in \operatorname{int}(T)$, *the interior of* $T \in \mathcal{T}$, *with a Lipschitz constant smaller than or equal to* $\mathcal{L}\mathrm{ip}(\gamma)$. *Then there exists a positive constant* C_2, *which depends exclusively depends on* $\beta_0 > 0$, $\|\mathbf{A}^{1/2}\mathbf{b}_2\|_{L^\infty(\Omega)}$, $\mathcal{L}\mathrm{ip}(\gamma)$, *and the shape-regularity of* $\mathbb{T}$, *such that*

$$C_2^{-1}\|\sigma - \sigma_h\|_{\mathbf{A}^{-1}} \leqslant \min_{\tau_h \in M_k(\mathcal{T})} \|\sigma - \tau_h\|_{\mathbf{A}^{-1}} + \|h_{\mathcal{T}}(u - \Pi_k u)\| + \|h_{\mathcal{T}}(1 - \Pi_k)\operatorname{div}\sigma\|.$$

The additional oscillations $\|h_{\mathcal{T}}(u - \Pi_k u)\|$ and $\|h_{\mathcal{T}}(1 - \Pi_k)(\operatorname{div}\sigma)\|$ can be higher-order contributions and then these terms explain the improved convergence of one variant for the BDM finite element family in [8] under Assumption (A).

This article, thus, generalises earlier contributions [5, 8–10, 12–14] for smooth or piecewise Lipschitz continuous coefficients to L^∞ coefficients without any further assumptions. The compactness argument of Schatz and Wang [17] for the displacement-oriented problem does *not* apply immediately to the mixed formulation in $H(\operatorname{div}) \times L^2$. Remark 12 below explains that no uniform L^2 approximation of the divergence component holds. This paper therefore compensates the lack of compactness by the computation and analysis of an optimal test function (the dual solution y in (1.7) of Sect. 1.4 below). Recent best-approximation for the flux in L^2 from the medius analysis [6, 15] combines with the compactness for the (dual) PDE. This and a careful shift of the discrete divergence circumvents the aforementioned lack of compactness in the divergence variable. In fact, this new methodology avoids any regularity argument and any Fortin interpolation at all.

1.4 Motivation

This subsection outlines the proof of the discrete inf-sup stability (1.6) in an abstract framework to guide the reader through the arguments. Suppose $X_h \times Y_h$ is a finite dimensional subspace of $H \times H$ with dual $X_h^* \times Y_h^*$ and let $x_h \in S(X_h)$, i.e., x_h belongs to X_h and has norm $\|x_h\|_H = 1$. Recall (1.5) and the well-posedness of the problem (1.3). Then, the dual problem is well-posed as well and $\langle x_h, \bullet\rangle_H = b(\bullet, y)$ has a unique dual solution y in the Hilbert space $(H, \langle\bullet,\bullet\rangle_H)$. The continuous inf-sup condition (1.5) shows

$$\beta\,\|y\|_H \leqslant \|b(\bullet, y)\|_{H^*} = \|x_h\|_H = 1, \quad \text{whence } \|y\|_H \leqslant 1/\beta \tag{1.7}$$

is bounded. Suppose that $y_h \in Y_h$ is a close approximation to y with $\|y - y_h\|_H \leqslant \varepsilon$ for some positive $\varepsilon < 1/\|b\|$, where $\|b\|$ is the operator norm of the bilinear form $b(\bullet,\bullet)$. Since

$$1 = b(x_h, y) = b(x_h, y_h) + b(x_h, y - y_h) \leqslant \|b(x_h, \bullet)\|_{Y_h^*}\|y_h\|_H + \varepsilon\|b\|,$$

it remains to bound $\|y_h\|_H$, e.g., with the triangle inequality

$$\|y_h\|_H \leqslant \|y\|_H + \|y - y_h\|_H \leqslant 1/\beta + \varepsilon.$$

The combination of the previous two displayed formulas gives a lower bound for $\|b(x_h, \bullet)\|_{Y_h^*}$. Under the assumption that ε is independent of y and so of x_h, this estimate reads

$$\beta \frac{1-\varepsilon\|b\|}{1+\varepsilon\beta} \leqslant \beta_h := \inf_{x_h \in X_h \setminus\{0\}} \sup_{y_h \in Y_h \setminus\{0\}} \frac{b(x_h, y_h)}{\|x_h\|_H \|y_h\|_H}. \tag{1.8}$$

This proves $\beta_0 \leqslant \beta(1-\varepsilon\|b\|)/(1+\varepsilon\beta)$ provided the approximation error $\|y-y_h\|_H$ is small independently of $X_h \times Y_h$ and $x_h \in S(X_h)$. A detailed investigation in Sect. 3.3 below reveals that the above strong form of a uniform approximation appears *neither* available in the norm of $H = H(\mathrm{div}, \Omega) \times L^2(\Omega)$ (cf. Remark 3.4) *nor* necessary for the stability under assumption (A). Recent results from a medius analysis [6, 15] and a careful shift of the discrete divergence variable successfully circumvent a uniform approximation in H.

1.5 Structure of the paper

Section 2 starts with the pre-compactness for uniform approximation and the precise assumptions on the set of admissible triangulations $\mathbb{T}$. The other two preliminary subsections concern the L^2 best-approximation of the fluxes and some discrete approximation result for the RT finite element family.

The stability analysis in Sect. 3 is based on the dual solution y in the conservative formulation characterised in Sect. 3.1. One contribution of y involves the PDE $\mathcal{L}_2\phi = g$ and allows for some pre-compacness and uniform approximation in Sect. 3.2. The proof of Theorem 1.1 concludes Sect. 3. A combination of the stability result (1.6) with the approximation arguments leads in Sect. 4 to Theorem 1.2, which generalises [6, 15] to non-selfadjoint indefinite second-order linear elliptic problems.

2 Preliminaries

This section introduces notations used in the paper, fixes the assumptions on the admissible triangulation $\mathbb{T}$, discusses an abstract version of compactness argument in [17], and then recalls some L^2 best-approximation property and concludes with an observation for the RT finite element family.

2.1 Notation

Standard notation on Lebesgue and Sobolev spaces $L^2(\Omega)$, $L^\infty(\Omega)$, $H_0^1(\Omega)$, $H^{-1}(\Omega) \equiv H_0^1(\Omega)^*$, and $H(\mathrm{div}, \Omega)$ apply throughout this paper. The L^2 scalar product $(\bullet, \bullet)_{L^2(\Omega)}$ induces the norm $\|\bullet\| := \|\bullet\|_{L^2(\Omega)}$ and the orthogonality relation $\perp$.

Whereas $\|\bullet\|$ denotes the norm in $L^2(\Omega)$ with the exception of the abbreviation $\|b\|$ for the bound of the bilinear form $b(\bullet, \bullet)$, the vector space $L^2(\Omega; \mathbb{R}^n)$ is endowed with the weighted scalar product $(\bullet, \bullet)_{\mathbf{A}^{-1}} := (\mathbf{A}^{-1}\bullet, \bullet)_{L^2(\Omega)}$ and induced norm $\|\bullet\|_{\mathbf{A}^{-1}} := \|\mathbf{A}^{-1/2}\bullet\|$ and so, for any $\tau \in L^2(\Omega; \mathbb{R}^n)$, is its distance $\mathrm{dist}(\tau, M_h) :=$

$\min_{\tau_h \in M_h} \|\tau - \tau_h\|_{\mathbf{A}^{-1}}$ to any subspace M_h of $L^2(\Omega; \mathbb{R}^n)$. The norm $\|(\tau, v)\|_H$ in the Hilbert space $H(\mathrm{div}, \Omega)$ is weighted with $\mathbf{A}^{-1}$ in $L^2(\Omega; \mathbb{R}^n)$ for the flux variable so the Hilbert space $H \equiv H(\mathrm{div}, \Omega) \times L^2(\Omega)$ has the weighted scalar product $\langle \bullet, \bullet, \rangle H$ with the induced norm $\|(\tau, v)\|_H$,

$$\|(\tau, v)\|_H^2 := \|\tau\|_{\mathbf{A}^{-1}}^2 + \|\mathrm{div}\,\tau\|^2 + \|v\|^2 \quad \text{for all } (\tau, v) \in H. \tag{2.1}$$

Duality brackets have the dual pairing as an index as in $\langle \bullet, \bullet, \rangle H^{-1}(\Omega) \times H_0^1(\Omega)$ above. To abbreviate the definition of inf-sup constants throughout this paper, let $S(V) := \{v \in V : \|v\|_V = 1\}$ for any normed linear space $(V, \|\bullet\|_V)$.

All emerging generic positive constants $C_1, \ldots, C_8$ in this paper exclusively depend on $\underline{\alpha}, \overline{\alpha}, \|\mathbf{b}\|_{L^\infty}, \|\mathbf{b}_1\|_{L^\infty}, \|\mathbf{b}_2\|_{L^\infty}$, and $\|\gamma\|_{L^\infty}$ as well as on α in (1.2) and β in (1.5) and on the class of admissible triangulations $\mathbb{T}$ specified in Subsection 2.2.

2.2 Assumptions on the discretization

The finite element spaces are based on admissible triangulations, the set $\mathbb{T}$ of all of those has certainly infinite cardinality; the point is that the constants in standard interpolation error estimates become universal through uniform shape regularity.

Definition 2.1 (*admissible triangulations*) The set of admissible triangulations $\mathbb{T}$ is a set of shape-regular triangulations of the polyhedral bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$ into simplices with uniform shape regularity and arbitrary small mesh sizes. Let $h_{\max}(\mathcal{T}) := \max h_{\mathcal{T}}$ for the piecewise constant mesh-size $h_{\mathcal{T}}$ for $\mathcal{T} \in \mathbb{T}$, defined by $h_{\mathcal{T}}|_T := \mathrm{diam}(T)$ in $T \in \mathcal{T}$, and abbreviate $\mathcal{T}(\delta) := \{\mathcal{T} \in \mathbb{T} : h_{\max}(\mathcal{T}) \leqslant \delta\}$.

Given $\mathcal{T} \in \mathbb{T}$, let $P_k(T)$ denote the polynomials of total degree at most $k \in \mathbb{N}_0$ seen as functions on $T \in \mathcal{T} \in \mathbb{T}$ and set $P_k(\mathcal{T}) := \{v_k \in L^\infty(\Omega) : \forall T \in \mathcal{T},\ v_k|_T \in P_k(T)\}$. Let $\Pi_k : L^2(\Omega) \to L^2(\Omega)$ be the L^2 projection onto $P_k(\mathcal{T})$ with respect to $\mathcal{T} \in \mathbb{T}$.

Definition 2.2 (*discrete spaces*) Any $\mathcal{T} \in \mathbb{T}$ is associated to the finite-dimensional subspace $V(\mathcal{T}) = M_k(\mathcal{T}) \times P_k(\mathcal{T})$ of $V := L^2(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ with $M_k(\mathcal{T}) := RT_k(\mathcal{T})$ or $M_k(\mathcal{T}) := BDM_k(\mathcal{T})$ of order $k \in \mathbb{N}_0$ from [2].

The best-approximation error reads $\mathrm{dist}(v, V(\mathcal{T})) := \inf\{\|v - v_h\| : v_h \in V(\mathcal{T})\}$ with the weighted L^2 norm, $\|v\|^2 = \|\tau\|_{\mathbf{A}^{-1}}^2 + \|w\|^2$ for $v = (\tau, w) \in V$. The density of smooth functions and standard approximation results for smooth functions proves the well-known pointwise convergence in the sense that each $v \in V$ satisfies [2]

$$\lim_{\delta \to 0^+} \sup_{\mathcal{T} \in \mathbb{T}(\delta)} \mathrm{dist}(v, V(\mathcal{T})) = 0. \tag{2.2}$$

2.3 Pre-compactness

This subsection adopts the key argument of [17].

Lemma 2.3 (uniform approximation on compact sets) *Suppose that K is a non-empty pre-compact subset of $V := L^2(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ with (2.2) for each $v \in K$. Then*

$$\lim_{\delta \to 0^+} \sup_{v \in K} \sup_{\mathcal{T} \in \mathbb{T}(\delta)} \operatorname{dist}(v, V(\mathcal{T})) = 0. \tag{2.3}$$

Proof Given any $\varepsilon > 0$ and $v \in K$, let $B(v, \varepsilon/2)$ be the open ball in V with center v and radius $\varepsilon/2$. The open cover $\{B(v, \varepsilon/2) : v \in K\}$ of the compact set $\overline{K}$ contains a finite sub-cover and so there exist $k_1, \dots, k_J \in K$ with $K \subset \bigcup_{j=1,\dots,J} B(k_j, \varepsilon/2)$. For each $k_j \in K$, (2.2) leads to $\delta_j > 0$ such that $\mathcal{T} \in \mathcal{T}(\delta_j)$ implies $\operatorname{dist}(k_j, V(\mathcal{T})) < \varepsilon/2$. Then $\delta := \min\{\delta_1, \dots, \delta_J\}$ implies $\mathbb{T}(\delta) \subset \bigcap_{j=1,\dots,J} \mathbb{T}(\delta_j)$. Given any $\mathcal{T} \in \mathbb{T}(\delta)$ and any $v \in K \subset \bigcup_{j=1,\dots,J} B(k_j, \varepsilon/2)$, there exists $j \in \{1, \dots, J\}$ with $\|v - k_j\| < \varepsilon/2$. Since $\mathcal{T} \in \mathbb{T}(\delta_j)$, $\operatorname{dist}(k_j, V(\mathcal{T})) < \varepsilon/2$. This and a triangle inequality show $\operatorname{dist}(v, V(\mathcal{T})) \leqslant \|v - k_j\| + \operatorname{dist}(k_j, V(\mathcal{T})) < \varepsilon/2 + \varepsilon/2 = \varepsilon$. □

The application of the previous lemma to the finite element approximation of the solution of the PDE reads as follows.

Lemma 2.4 (uniform approximation of solutions) *For any $\varepsilon > 0$ there exists some $\delta > 0$ such that, given any $g \in L^2(\Omega)$ and the weak solution $\phi \in H_0^1(\Omega)$ to $\mathcal{L}_2\phi = g$ (with $\mathcal{L}_2$ from (1.1)), the vector $v := (\mathbf{A}\nabla\phi, \mathbf{b} \cdot \nabla\phi + \gamma\phi) \in V$ satisfies* $\sup_{\mathcal{T} \in \mathbb{T}(\delta)} \operatorname{dist}(v, V(\mathcal{T})) \leqslant \epsilon \|g\|$.

Proof The linear and bounded bijective differential operator $\mathcal{L}_2 : H_0^1(\Omega) \to H^{-1}(\Omega)$ has a bounded inverse. The embedding $\iota : L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ is compact and so is the composition $\mathcal{L}_2^{-1} \circ \iota : L^2(\Omega) \to H_0^1(\Omega)$. Define the operator $T : L^2(\Omega) \to L^2(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ for any $g \in L^2(\Omega)$ by

$$T(g) := (\mathbf{A}\nabla\phi, \mathbf{b} \cdot \nabla\phi + \gamma\phi) \quad \text{with} \quad \phi := \mathcal{L}_2^{-1} g.$$

Since $\mathcal{L}_2^{-1} \circ \iota$ is compact, $K := T(S(L^2(\Omega))$ is pre-compact in $V = L^2(\Omega; \mathbb{R}^n) \times L^2(\Omega)$. Given any $\varepsilon > 0$ the approximation result (2.2) and Lemma 2.3 lead to a positive δ with (2.3). Consequently, the assertion $\sup_{\mathcal{T} \in \mathbb{T}(\delta)} \operatorname{dist}(v, V(\mathcal{T})) \leqslant \epsilon \|g\|$ holds for all $g \in S(L^2(\Omega))$ and corresponding $v := (\mathbf{A}\nabla\phi, \mathbf{b} \cdot \nabla\phi + \gamma\phi) \in V$. A rescaling proves the result for all $g \in L^2(\Omega)$.

2.4 L^2 best-approximation of the fluxes

The medius analysis of mixed finite element methods employs arguments from *a priori* and *a posteriori* error analysis [6, 15] to prove new L^2 best-approximation results. Recall that Π_k is the L^2 projection onto $P_k(\mathcal{T})$ and $h_\mathcal{T}$ is the mesh-size associated to $\mathcal{T}$.

Lemma 2.5 *(*flux L^2 best-approximation*) There exists a constant C_3, which depends on the shape-regularity in $\mathcal{T}$, on Ω and on $\underline{\alpha}, \overline{\alpha}$, such for any $\mathbf{p} \in H(\operatorname{div}, \Omega)$ and any $\mathcal{T} \in \mathbb{T}$, there exists $\mathbf{p}_h \in M_k(\mathcal{T})$ such that* $\operatorname{div} \mathbf{p}_h = \Pi_k \operatorname{div} \mathbf{p}$ *and*

$$C_3^{-1} \|\mathbf{p} - \mathbf{p}_h\|_{\mathbf{A}^{-1}} \leqslant \operatorname{dist}(\mathbf{p}, M_k(\mathcal{T})) + \|h_\mathcal{T}(1 - \Pi_k) \operatorname{div} \mathbf{p}\|.$$

This is the L^2 best-approximation result from [15,Lemma 5.1] for mixed finite element approximations for the unit matrix $\mathbf{A}$. Although with a different focus, the paper [6] introduces a general framework with a mesh-dependent norm $\|\bullet\|_h$ in $P_k(\mathcal{T})$; while [11,Eq (3.6)] presents a localized refinement of this lemma.

Proof Given $\mathbf{p} \in H(\mathrm{div}, \Omega)$, the right-hand sides $F(w) := (w, \mathrm{div}\,\mathbf{p})_{L^2(\Omega)}$ and $G(\mathbf{q}) := (\mathbf{p}, \mathbf{q})$ lead in the elliptic mixed formulation (for the Laplacian)

$$(\sigma, \mathbf{q}) - (u, \mathrm{div}\,\mathbf{q})_{L^2(\Omega)} + (w, \mathrm{div}\,\mathbf{p})_{L^2(\Omega)} = G(\mathbf{q}) + F(w) \quad \text{for all } (\mathbf{q}, w) \in H$$

to the unique solution $(\sigma, u) \equiv (\mathbf{p}, 0) \in H$. Its straight-forward mixed finite element discretisation substitutes H by $V_h := M_k(\mathcal{T}) \times P_k(\mathcal{T})$ and leads to a unique discrete solution $(\mathbf{p}_h, v_h) \in V_h$ with $\mathrm{div}\,\mathbf{p}_h = \Pi_k \,\mathrm{div}\,\mathbf{p}$. This and [6,Thm 2.2] lead to the asserted best-approximation result (in terms of (non-weighted) L^2 norms)

$$C_4^{-1}\|\mathbf{p} - \mathbf{p}_h\| \leqslant \inf_{\mathbf{q}_h \in M_k(\mathcal{T})} \|\mathbf{p} - \mathbf{q}_h\| + \|h_{\mathcal{T}}(1 - \Pi_k)\,\mathrm{div}\,\mathbf{p}\|.$$

The constant C_4 from [6, 15] does not depend on the coefficients $\mathbf{A}$, $\mathbf{b}$, γ but depends on the shape-regularity in $\mathcal{T}$ and on Ω. The equivalence of norms concludes the proof and leads to the asserted constant C_3, which depends on C_4 and $\underline{\alpha}$, $\overline{\alpha}$. □

2.5 A discrete approximation result for Raviart-Thomas functions

In any space-dimension n and degree k, the RT functions satisfy a rather particular approximation estimate with the componentwise L^2 projection Π_k onto $P_k(\mathcal{T})$.

Lemma 2.6 *Any* $\tau_{RT} \in RT_k(\mathcal{T})$ *satisfies* $\|\tau_{RT} - \Pi_k \tau_{RT}\| \leqslant \frac{n}{(n+1)(n+k)}\|h_{\mathcal{T}}\,\mathrm{div}\,\tau_{RT}\|$.

The proof will be postponed to the appendix because of its focus on the RT finite element shape functions. The statement of the above lemma fails for the BDM finite element family.

3 Stability analysis

This section deals with approximation of fluxes and stability result. The design of a test function in the proof of a discrete inf-sup condition is based on the characterisation and approximation of a dual solution.

3.1 Dual solution and conservative formulation

The inner structure of the dual solution y exploits the elliptic PDE and generates some compactness argument in the subsequent subsection. Recall that the operator $\mathcal{L}_2 : H_0^1(\Omega) \to H^{-1}(\Omega)$ from (1.1) is bijective.

Theorem 3.1 (dual solution in conservative formulation) *Suppose* $\mathbf{b}_1 := \mathbf{A}^{-1}\mathbf{b}$ *and* $\mathbf{b}_2 \equiv 0$ *a.e. in* Ω *in* (1.4). *Then* $x = (\sigma, u) \in H$ *and* $y = (\zeta, z) \in H$ *satisfy* $\langle x, \bullet\rangle_H = b(\bullet, y)$ *in* H *if and only if*

$$\zeta = \sigma - \mathbf{A}\nabla\phi \quad \textit{and} \quad z = \operatorname{div}\sigma - \phi \quad \textit{a.e. in } \Omega$$

for the weak solution $\phi \in H^1_0(\Omega)$ *to* $\mathcal{L}_2\phi = g := \mathbf{b}\cdot\mathbf{A}^{-1}\sigma + (\gamma - 1)\ \operatorname{div}\sigma - u \in L^2(\Omega)$.

The function ϕ originates from a known L^2 orthogonal decomposition

$$L^2(\Omega; \mathbb{R}^n) = \nabla H^1_0(\Omega) \oplus H(\operatorname{div}{=}0, \Omega) \tag{3.1}$$

with $H(\operatorname{div}{=}0, \Omega) := \{\tau \in H(\operatorname{div}, \Omega) : \operatorname{div}\tau = 0 \text{ a.e. in } \Omega\}$. The decomposition (3.1) is also useful in the proof of equivalence of the displacement formulation with the differential operators in (1.1) to the mixed formulations with (1.3).

Proof of Theorem 3.1 For the general version of the bilinear form $b(\bullet, \bullet)$, the equation $\langle x, \bullet\rangle_H = b(\bullet, y)$ is equivalent to $u = \mathbf{b}_1 \cdot \zeta + \gamma\, z - \operatorname{div}\zeta$ a.e. in Ω and

$$(\tau, \mathbf{A}^{-1}(\zeta - \sigma) - z\mathbf{b}_2)_{L^2(\Omega)} + (z - \operatorname{div}\sigma, \operatorname{div}\tau)_{L^2(\Omega)} = 0 \quad \text{for all } \tau \in H(\operatorname{div}, \Omega). \tag{3.2}$$

The test with $\tau \in H(\operatorname{div}{=}0, \Omega)$ proves that $\mathbf{A}^{-1}(\sigma - \zeta) + z\mathbf{b}_2 \perp H(\operatorname{div}{=}0, \Omega)$ and so (3.1) leads to $\phi \in H^1_0(\Omega)$ with

$$\mathbf{A}\nabla\phi = \sigma - \zeta + z\mathbf{A}\mathbf{b}_2 \quad \text{a.e. in } \Omega.$$

This identity allows the substitution of $\mathbf{A}^{-1}(\zeta - \sigma) - z\mathbf{b}_2$ in the above formula (3.2) with general $\tau \in H(\operatorname{div}, \Omega)$. Then, an integration by parts shows the resulting identity $(\phi + z - \operatorname{div}\sigma, \operatorname{div}\tau) = 0$. The surjectivity of $\operatorname{div} : H(\operatorname{div}, \Omega) \to L^2(\Omega)$ proves

$$\operatorname{div}\sigma = \phi + z \quad \text{a.e. in } \Omega.$$

The combination of the three preceding identities leads to the PDE

$$\begin{aligned} -\operatorname{div}(\mathbf{A}\nabla\phi) + \mathbf{b}_1 \cdot \mathbf{A}\nabla\phi + \gamma\phi = & -\operatorname{div}(z\mathbf{A}\mathbf{b}_2) + (\mathbf{b}_1 \cdot \mathbf{A}\mathbf{b}_2)\, z \\ & + \mathbf{b}_1 \cdot \sigma + (\gamma - 1)\ \operatorname{div}\sigma - u \end{aligned}$$

in the sense of distributions. Since $\mathbf{b}_2 = 0$, the right-hand side g belongs to L^2. This proves one direction of the assertion; the direct proof of the converse is omitted. □

Remark 3.2 (no divergence formulation) The proof shows the extra term $-\operatorname{div}(z\mathbf{A}\mathbf{b}_2) \in H^{-1}(\Omega)$ in case (1.4) is considered for non-zero $\mathbf{b}_2 \in L^\infty(\Omega; \mathbb{R}^n)$. This term does *not* belong to $L^2(\Omega)$ under Assumption (A) and is, therefore, excluded.

3.2 Approximation of the fluxes

The subsequent lemma describes the uniform approximation of the flux variable by a combination of the compactness argument and the L^2 best-approximation of Subsections 2.3 and 2.4.

Lemma 3.3 (flux approximation) *Given any $\varepsilon > 0$, there exists $\delta > 0$ such that the following holds for all $\mathcal{T} \in \mathbb{T}(\delta)$ and $g \in L^2(\Omega)$. There exists some $\mathbf{p}_h \in M_k(\mathcal{T})$ that approximates $\mathbf{p} := \mathbf{A}\nabla\phi \in H(\operatorname{div}, \Omega)$ for the weak solution $\phi \in H_0^1(\Omega)$ to $\mathcal{L}_2\phi = g$ with*

$$\operatorname{div}\mathbf{p}_h = \Pi_k \operatorname{div}\mathbf{p} \quad \textit{and} \quad \|\mathbf{p} - \mathbf{p}_h\|_{\mathbf{A}^{-1}} \leqslant \epsilon\|g\|.$$

Proof Given any $\varepsilon > 0$ and the constant C_3 from Lemma 2.5, Lemma 2.4 leads to a positive $\delta \leqslant \min\{1, 2^{-1}\epsilon/C_3\}$ with

$$\sup_{\mathcal{T}\in\mathbb{T}(\delta)} \operatorname{dist}((\mathbf{A}\nabla\phi, \mathbf{b}\cdot\nabla\phi + \gamma\phi), V(\mathcal{T})) \leqslant 2^{-3/2}\epsilon/C_3\,\|g\|$$

(the distance is with respect to the weighted norm $\|\bullet\|_{\mathbf{A}^{-1}}$ in $L^2(\Omega;\mathbb{R}^n)$ and $\|\bullet\|$ in $L^2(\Omega)$). Lemma 2.5 applies to $\mathbf{p} := \mathbf{A}\nabla\phi$ with $\operatorname{div}\mathbf{p} = \mathbf{b}\cdot\nabla\phi + \gamma\phi - g \in L^2(\Omega)$ and, for any $\mathcal{T} \in \mathbb{T}(\delta)$, leads to some approximation $\mathbf{p}_h \in M_k(\mathcal{T})$ with $\operatorname{div}\mathbf{p}_h = \Pi_k \operatorname{div}\mathbf{p}$ and

$$\begin{aligned}
C_3^{-1}\|\mathbf{p} - \mathbf{p}_h\|_{\mathbf{A}^{-1}} &\leqslant \operatorname{dist}(\mathbf{p}, M_k(\mathcal{T})) + \delta\|(1-\Pi_k)(\mathbf{b}\cdot\nabla\phi + \gamma\phi - g)\| \\
&\leqslant \operatorname{dist}(\mathbf{p}, M_k(\mathcal{T})) + \operatorname{dist}(\mathbf{b}\cdot\nabla\phi + \gamma\phi, P_k(\mathcal{T})) + \delta\,\|g\| \\
&\leqslant 2^{1/2}\operatorname{dist}((\mathbf{p}, \mathbf{b}\cdot\nabla\phi + \gamma\phi), V(\mathcal{T})) + \delta\,\|g\| \leqslant C_3^{-1}\epsilon\,\|g\|.
\end{aligned}$$

This concludes the proof. □

Remark 3.4 (no uniform approximation in $H(\operatorname{div})$) Lemma 3.3 does *not* state a uniform approximation estimate for the divergence and, in fact, an estimate of the form $\|\operatorname{div}(\mathbf{p}-\mathbf{p}_h)\| \leqslant \epsilon\|g\|$ *cannot* hold in general. To see this, adopt the notation of the proof of Lemma 3.3 and a reverse triangle inequality for

$$\|g - \Pi_k g\| - \|\operatorname{div}(\mathbf{p} - \mathbf{p}_h)\| \leqslant \|(1-\Pi_k)(\mathbf{b}\cdot\nabla\phi + \gamma\phi)\| \leqslant \epsilon/(2C_3)\,\|g\|.$$

The flux $\mathbf{p} = \mathbf{A}\nabla\mathcal{L}_2^{-1}g$ depends on $g \in S(L^2(\Omega))$ and so does the crucial term $\|\operatorname{div}(\mathbf{p} - \mathbf{p}_h)\| = \|(1-\Pi_k)\operatorname{div}\mathbf{p}\|$. Hence, $\sup\{\|g - \Pi_k g\| : g \in S(L^2(\Omega))\} = 1$ implies

$$1 - \epsilon/(2C_3) \leqslant \sup\{\|(1-\Pi_k)\operatorname{div}(\mathbf{A}\nabla\mathcal{L}_2^{-1}g)\| : g \in S(L^2(\Omega))\}.$$

Therefore, the approximation error $\|\operatorname{div}(\mathbf{p} - \mathbf{p}_h)\|$ will *not* tend to zero uniformly for all $g \in S(L^2(\Omega))$ as ϵ and δ tend to zero. □

Example 3.5 (RT approximation in $H(\mathrm{div})$ for particular g) The stability analysis in Subsection 1.4 concerns a discrete $x_h := (\sigma_h, u_h)$ with norm $\|(\sigma_h, u_h)\|_H = 1$ and leads in Theorem 3.1 to the particular right-hand side $g := \mathbf{b}\cdot\mathbf{A}^{-1}\sigma_h + (\gamma - 1)\ \mathrm{div}\,\sigma_h - u_h$ with $\|g\| \leqslant C_5 < \infty$ for the essential supremum C_5^2 of $|\mathbf{A}^{-1/2}\mathbf{b}|^2 + |\gamma - 1|^2 + 1$ in Ω. This g allows for a uniform approximation of $\mathbf{p}$ by $\mathbf{p}_h$ in $H(\mathrm{div}, \Omega)$ for the RT finite element family.

For instance, in the extreme case of piecewise constant coefficients, $g - \Pi_k g = \mathbf{b}\cdot\mathbf{A}^{-1}(1 - \Pi_k)\sigma_h$. With $C_6 := \|\mathbf{A}^{-1}\mathbf{b}\|_{L^\infty(\Omega)}$, Lemma 2.6 shows $\|g - \Pi_k g\| \leqslant \delta C_6$. The combination with Lemma 3.3 lead to $\mathbf{p}_h$ with $\|\mathbf{p} - \mathbf{p}_h\|_{H(\mathrm{div},\Omega)} \leqslant \epsilon\, C_5 + \delta\, C_6$. This and the arguments of Subsection 1.4 lead to the discrete stability (1.6).

3.3 Proof of theorem 1.1

Given any $0 < \varepsilon < \beta/\|b\|$, choose $\delta > 0$ as in Lemma 3.3. Suppose $\mathcal{T} \in \mathbb{T}(\delta)$ and let $x_h = (\sigma_h, u_h) \in V_h := M_k(\mathcal{T}) \times P_k(\mathcal{T})$ have norm $\|x_h\|_H = 1$ and define $g := \mathbf{b}\cdot\mathbf{A}^{-1}\sigma_h + (\gamma - 1)\ \mathrm{div}\,\sigma_h - u_h$. Replace x by x_h in Theorem 3.1 and let $\phi \in H_0^1(\Omega)$ solve $\mathcal{L}_2\phi = g$ to define $\zeta = \sigma_h - \mathbf{A}\nabla\phi$ and $z = \mathrm{div}\,\sigma_h - \phi$. Then $y = (\zeta, z) \in H$ is the dual solution and solves $\langle x_h, \bullet\rangle_H = b(\bullet, y)$ in H for the bilinear form (1.4) (with $\mathbf{b}_1 := \mathbf{A}^{-1}\mathbf{b}$ and $\mathbf{b}_2 \equiv 0$ a.e. in Ω). Lemma 3.3 applies to $\mathbf{p} := \mathbf{A}\nabla\phi$ and leads to $\mathbf{p}_h$ with $\mathrm{div}\,\mathbf{p}_h = \Pi_k\,\mathrm{div}\,\mathbf{p}$ and $\|\mathbf{p} - \mathbf{p}_h\|_{\mathbf{A}^{-1}} \leqslant \varepsilon\, C_5$. Hence, $\zeta_h := \sigma_h - \mathbf{p}_h$ and $z_h := \Pi_k z$ define $y_h := (\zeta_h, z_h) \in V_h$ with

$$\|y_h\|_H \leqslant \|y\|_H + \|\zeta - \zeta_h\|_{\mathbf{A}^{-1}} \leqslant \|y\|_H + \varepsilon\, C_5 \tag{3.3}$$

as $\|\,\mathrm{div}\,\zeta_h\| = \|\Pi_k\,\mathrm{div}\,\zeta\| \leqslant \|\,\mathrm{div}\,\zeta\|$, $\|z_h\| = \|\Pi_k z\| \leqslant \|z\|$ and $\mathrm{div}\,\sigma_h \in P_k(\mathcal{T})$. Moreover,

$$\|\zeta - \zeta_h\|_{\mathbf{A}^{-1}}^2 + \|z - z_h\|^2 \leqslant \varepsilon^2\, C_5^2 + \|\phi - \Pi_k\phi\|^2.$$

Piecewise Poincaré inequalities (with the Payne-Weinberger constant $1/\pi$ for convex domains [16]) show $\|\phi - \Pi_k\phi\| \leqslant \delta/\pi\,\|\nabla\phi\|$. Recall that $H_0^1(\Omega)$ is endowed with the seminorm $\|\nabla \bullet\|$ and let C_F denote the constant in the Friedrichs inequality $\|\bullet\| \leqslant C_F\|\nabla\bullet\|$ in $H_0^1(\Omega)$. Note that (1.2) leads to $\alpha\,\|\nabla\phi\| \leqslant \sup\{a(\psi, \phi) : \psi \in H_0^1(\Omega),\ \|\nabla\psi\| = 1\}$. Hence $a(\psi, \phi) = \langle\mathcal{L}_2\phi, \psi\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle g, \psi\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \leqslant C_F\|g\|$ implies $\alpha\,\|\nabla\phi\| \leqslant C_F\, C_5$. The combination with Poincaré inequalities shows $\|\phi - \Pi_k\phi\| \leqslant \delta C_F\, C_5/(\alpha\,\pi)$ and so

$$\|\zeta - \zeta_h\|_{\mathbf{A}^{-1}}^2 + \|z - z_h\|^2 \leqslant \varepsilon^2\, C_5^2 + \delta^2 C_F^2\, C_5^2/(\alpha^2\,\pi^2) =: (\varepsilon')^2.$$

Since $u_h \perp \mathrm{div}(\zeta - \zeta_h)$ and $z - z_h \perp \mathrm{div}\,\sigma_h$ ($\perp$ denotes orthogonality in $L^2(\Omega)$),

$$\begin{aligned} b(x_h, y - y_h) &= (\sigma_h + \mathbf{b}u_h, \zeta - \zeta_h)_{\mathbf{A}^{-1}} + (\gamma\, u_h, z - z_h) \\ &\leqslant \varepsilon'\left(\|\sigma_h + \mathbf{b}u_h\|_{\mathbf{A}^{-1}}^2 + \|(1 - \Pi_k)(\gamma\, u_h)\|^2\right)^{1/2} \end{aligned}$$

$$\leqslant \varepsilon' \left(2\|\sigma_h\|_{\mathbf{A}^{-1}}^2 + 2\|\mathbf{A}^{-1/2}\mathbf{b}\|_{L^\infty(\Omega)}^2 \|u_h\|^2 + \|\gamma\|_{L^\infty(\Omega)}^2 \|u_h\|^2\right)^{1/2}$$
$$\leqslant \varepsilon' C_7$$

with the constant $C_7^2 := \max\{2, 2\|\mathbf{A}^{-1/2}\mathbf{b}\|_{L^\infty(\Omega)}^2 + \|\gamma\|_{L^\infty(\Omega)}^2\}$. The arguments of Sect. 1.4 lead to

$$1 = \|x_h\|_H^2 = b(x_h, y) = b(x_h, y_h) + b(x_h, y - y_h) \leqslant \|b(x_h, \bullet)\|_{V_h^*} \|y_h\|_H + \varepsilon' C_7. \tag{3.4}$$

Since $\beta\|y\|_H \leqslant \|b(\bullet, y)\|_{H^*} = \|x_h\|_H = 1$ implies $\|y\|_H \leqslant 1/\beta$, (3.3) reads $\|y_h\|_H \leqslant 1/\beta + \varepsilon\, C_5$. This and (3.4) verify

$$1 - \varepsilon' C_7 \leqslant \|b(x_h, \bullet)\|_{V_h^*} \|y_h\|_H \leqslant \|b(x_h, \bullet)\|_{V_h^*} (1/\beta + \varepsilon\, C_5).$$

Since x_h was arbitrary in $S(V_h)$ (with V_h endowed with the norm in H),

$$\beta \frac{1 - \varepsilon' C_7}{1 + \varepsilon\, \beta\, C_5} \leqslant \beta_h := \inf_{x_h \in S(V_h)} \sup_{y_h \in S(V_h)} b(x_h, y_h).$$

Relabelling ϵ and δ proves the assertion: For any $0 < \beta_0 < \beta$, there exists $\delta > 0$ with (1.6). This establishes the theorem for the conservative version of the mixed finite element discretisation with $\mathbf{b}_1 := \mathbf{A}^{-1}\mathbf{b}$ and $\mathbf{b}_2 \equiv 0$ a.e. in Ω.

To deduce the same inf-sup constant for the other variant, define $b_1(\bullet, \bullet)$ (resp. $b_2(\bullet, \bullet)$) by (1.6) for $\mathbf{b}_1 := \mathbf{A}^{-1}\mathbf{b}$ and $\mathbf{b}_2 \equiv 0$ (resp. $\mathbf{b}_1 := 0$ and $\mathbf{b}_2 := \mathbf{b} \cdot \mathbf{A}^{-1}$) a.e. in Ω. The above proof shows $0 < \beta_0 \leqslant \beta_h$ and it is elementary to see that

$$\beta_h = \inf_{(\tau_h, v_h) \in S(V_h)} \sup_{(\sigma_h, u_h) \in S(V_h)} b_1((\tau_h, -v_h), (\sigma_h, -u_h)).$$

A direct calculation shows $b_1((\tau, -v), (\sigma, -u)) = b_2((\sigma, u), (\tau, v))$ for all (σ, u), $(\tau, v) \in H$. This and a duality argument (singular values of a square matrix coincide with those of its transposed) in the last equality show

$$\beta_h = \inf_{(\tau_h, v_h) \in S(V_h)} \sup_{(\sigma_h, u_h) \in S(V_h)} b_2((\sigma_h, u_h), (\tau_h, v_h)) = \inf_{x_h \in S(V_h)} \sup_{y_h \in S(V_h)} b_2(x_h, y_h).$$

Hence, the divergence formulation has the same discrete inf-sup constant β_h. □

Remark 3.6 (δ dependence) The size of δ in (1.6) is hidden behind a compactness argument of Lemma 3.3. Besides the norms and parameters mentioned in Assumption (A), the mapping properties of $\mathcal{L}_2^{-1}$ are of relevance as well. A review of the proofs of this paper shows that there is a finite sub-cover of $S(L^2(\Omega))$ with small balls in $H^{-1}(\Omega)$ that leads to a finite number of (without loss of generality) smooth functions $k_1, \ldots, k_J$ as in the proof of Lemma 2.3. The size of δ is related to the approximation properties of the weak solutions Φ_j to $\mathcal{L}_2 \Phi_j = k_j$ a.e. The regularity properties of $\Phi_j \in H_0^1(\Omega)$

are *not* characterised for Assumption (A): In fact, it is unknown whether Φ_j belongs to any $H^{1+s}(\Omega)$ for any $s > 0$. Under Assumption (B) and reduced elliptic regularity, however, the afore mentioned approximation properties could be quantified more and reveal further information on δ.

4 L^2 best-approximation

The notation of Theorem 1.2 applies throughout this section with continuous and discrete solutions $x = (\sigma, u)$ and $x_h = (\sigma_h, u_h)$.

4.1 Proof of theorem 1.2.a

Given $\mathbf{p} := \sigma \in H(\mathrm{div}, \Omega)$ and $\mathcal{T} \in \mathbb{T}(\delta)$, choose $\sigma_h^* := \mathbf{p}_h \in M_k(\mathcal{T})$ as in Lemma 2.5, and define $e_h := (\sigma_h - \sigma_h^*, u_h - \Pi_k u) \in V_h$. Given $\beta_0 > 0$ in (1.6) there exists some $y_h = (\tau_h, v_h) \in V_h$ with $\|y_h\|_H = 1$ and

$$\beta_0 \|e_h\|_H \leqslant b(e_h, y_h) = b((\sigma_h - \sigma_h^*, u_h - \Pi_k u), y_h) = b((\sigma - \sigma_h^*, u - \Pi_k u), y_h).$$

Since $u - \Pi_k u \perp \mathrm{div}\, \tau_h$ and $v_h \perp \mathrm{div}(\sigma - \sigma_h^*)$, the last term is equal to

$$\begin{aligned}&(\sigma - \sigma_h^*, \tau_h - v_h\, \mathbf{A}\mathbf{b}_2)_{\mathbf{A}^{-1}} + (u - \Pi_k u, \mathbf{b}_1 \cdot \tau_h + \gamma\, v_h)_{L^2(\Omega)}\\ &\quad \leqslant C_8\, \|(\sigma - \sigma_h^*, u - \Pi_k u)\|_L\end{aligned}$$

in terms of the weighted L^2 norm $\|\bullet\|_L \leqslant \|\bullet\|_H$ in H with $\|(\tau_h, v_h)\|_L^2 := \|\tau_h\|_{\mathbf{A}^{-1}}^2 + \|v_h\|^2 \leqslant 1$ and with $C_8^2 = 1 + \|\mathbf{A}^{1/2}\mathbf{b}_1\|_{L^\infty(\Omega)}^2 + \|\mathbf{A}^{1/2}\mathbf{b}_2\|_{L^\infty(\Omega)}^2 + \|\gamma\|_{L^\infty(\Omega)}^2$. Consequently, $\|e_h\|_H \leqslant \Big(C_8/\beta_0\Big)\, \|(\sigma - \sigma_h^*, u - \Pi_k u)\|_L$. Lemma 2.5 shows

$$C_3^{-1}\|\sigma - \sigma_h^*\|_{\mathbf{A}^{-1}} \leqslant \mathrm{dist}(\sigma, M_k(\mathcal{T})) + \|h_{\mathcal{T}}(1 - \Pi_k)\,\mathrm{div}\,\sigma\|. \tag{4.1}$$

This and the distance dist_L (measured in the norm $\|\bullet\|_L$) lead to

$$\|e_h\|_H \leqslant C_8 \max\{1, C_3\}/\beta_0\ (\mathrm{dist}_L((\sigma, u), V_h) + \|h_{\mathcal{T}}(1 - \Pi_k)\,\mathrm{div}\,\sigma\|)\,.$$

This, (4.1), and a triangle inequality conclude the proof. □

4.2 Proof of theorem 1.2.b

Throughout this subsection, let $\mathbf{b}_1 \equiv 0$ and $\mathbf{b}_2 := \mathbf{A}^{-1}\mathbf{b}$ a.e. in Ω in (1.4) and let $f \in L^2(\Omega)$ be a fixed right-hand side for the continuous and discrete problem $\mathcal{L}_2 u = f$.

Return to the proof of the previous subsection with e_h and follow the first lines until

$$\beta_0 \|e_h\|_H \leqslant b(e_h, y_h) = (\sigma - \sigma_h^*, \tau_h - v_h\, \mathbf{Ab}_2)_{\mathbf{A}^{-1}} + (u - \Pi_k u, \gamma\, v_h)_{L^2(\Omega)}.$$

Abbreviate $\overline{\gamma} := \Pi_0\gamma$ and utilize the Lipschitz continuity of the coefficients γ on each simplex

$$\|\Pi_k(\gamma(u - \Pi_k u))\| = \|\Pi_k((\gamma - \overline{\gamma})(u - \Pi_k u))\| \leqslant \mathcal{L}\mathrm{ip}(\gamma)\|h_{\mathcal{T}}(u - \Pi_k u)\|.$$

This controls the above term $(u - \Pi_k u, \gamma\, v_h)_{L^2(\Omega)} \leqslant \|\Pi_k(\gamma(u - \Pi_k u))\|\,\|v_h\|$ and leads to

$$\begin{aligned}\|\sigma_h^* - \sigma_h\|_{\mathbf{A}^{-1/2}} \leqslant \|e_h\|_L \leqslant \|e_h\|_H \leqslant C_8/\beta_0\, \|\sigma - \sigma_h^*\|_{\mathbf{A}^{-1}} \\ + \mathcal{L}\mathrm{ip}(\gamma)/\beta_0\, \|h_{\mathcal{T}}(u - \Pi_k u)\|.\end{aligned}$$

A triangle inequality in $L^2(\Omega; \mathbb{R}^n)$ is followed by (4.1) in the proof of

$$\begin{aligned}\|\sigma - \sigma_h\|_{\mathbf{A}^{-1/2}} \leqslant C_3(1 + C_8/\beta_0)\Big(\mathrm{dist}(\sigma, M_h) + \|h_{\mathcal{T}}(1 - \Pi_k)\,\mathrm{div}\,\sigma\|\Big) \\ + \mathcal{L}\mathrm{ip}(\gamma)/\beta_0\, \|h_{\mathcal{T}}(u - \Pi_k u)\|.\end{aligned}$$

This completes the rest of the proof. □

4.3 Conservative formulation

Theorem 2.a includes an error estimate for the conservative formulation with $\mathbf{b}_1 := \mathbf{A}^{-1}\mathbf{b}$ and $\mathbf{b}_2 := 0$ in (1.4) and $\sigma = -\mathbf{A}\nabla u - u\,\mathbf{b}$ with $\mathrm{div}\,\sigma = f - \gamma u$. The refined analog of Theorem 2.b is not expected because of an extra term exemplified in the extreme case of piecewise constant coefficients $\mathbf{b}_1$ and γ. The arguments of Subsection 4.1 lead to

$$\begin{aligned}\beta_0\, \|e_h\|_H &\leqslant b((\sigma - \sigma_h^*, u - \Pi_k u), y_h) \\ &= (\sigma - \sigma_h^*, \tau_h)_{\mathbf{A}^{-1}} + ((u - \Pi_k u)\mathbf{b}_1, \tau_h - \Pi_k \tau_h)_{L^2(\Omega)}.\end{aligned}$$

The last term is not of higher order for the BDM finite element family as pointed out in [8] through numerical evidence. For the RT finite element family, however, Lemma 2.6 shows $\|\tau_h - \Pi_k \tau_h\| \lesssim \|h_{\mathcal{T}}\ \mathrm{div}\, \tau_h\|$ and then leads to a higher-order contribution in the asserted inequality of Theorem 1.2.b as the final result. □

The arguments could be generalised, but those result are of limited relevance as the convergence order is *not* generally improved in comparison with Theorem 2.a. An exception is the example of [7,Sect 3.5] (with $\mathbf{b} = 0 = \gamma$ on the unit ball) when Theorem 1.2.b guarantees $O(\delta^2)$ for the L^2 flux error for $k = 0$.

Acknowledgements The research of the first author has been supported by the Deutsche Forschungsgemeinschaft in the Priority Program 1748 under the project "foundation and application of generalied mixed FEM towards nonlinear problems in solid mechanics" (CA 151/22-2). The finalization of this paper has been supported by DST SERB MATRICS grant No. MTR/2017/000199 (NN), MATRICS grant No. MTR/201S/000309 (AKP) and SPARC project (id 235) entitled *the mathematics and computation of plates*.

Funding Open Access funding enabled and organized by Projekt DEAL.

Appendix: proof of Lemma 2.6

For any simplex $T \subset \mathbb{R}^n$, let $P_k(T;\mathbb{R}^n)$ be the linear space of vector-valued polynomials q_k of degree at most k in any component and let $\|\bullet\|$ abbreviate the L^2 norm $\|\bullet\|_{L^2(T)}$ on T. The particular structure of the RT function τ_{RT} leads to some polynomial $g \in P_k(T)$ and

$$\tau_{RT} = g(x)\,x + p_k \quad \text{for all } x \in T \quad \text{and some } p_k \in P_k(T;\mathbb{R}^n).$$

The argument x (will always belong to T) is often neglected as in $\tau_{RT} := \tau_{RT}(x)$ or $p_k = p_k(x)$, while (with a small inconsistency, but the right emphasis) written out in the leading term $g(x)\,x$. The latter polynomial is either identically zero or of exact degree $k+1$ in the sense that g is a sum of monomials of exact degree k. Adopt a multi-index notation with $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$ and the monomial $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ of degree $k = |\alpha| := \alpha_1 + \cdots + \alpha_n$ for any $x = (x_1, \dots, x_n) \in T$. With real coefficients c_α for any $\alpha \in \mathbb{N}_0^n$ of degree $|\alpha| = k$,

$$g(x) = \sum_{|\alpha|=k} c_\alpha x^\alpha \quad \text{for all } x \in T.$$

(The symbol $|\alpha| = k$ under the sum sign abbreviates the set of all multi-indices $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$ of degree k). The divergence

$$\operatorname{div}(g(x)\,x) = n\,g(x) + x \cdot \nabla g(x)$$

of the vector-valued polynomial $g(x)\,x$ of degree $k+1$ with respect to x is computed with the observation that, for any $\alpha \in \mathbb{N}_0^n$ with $|\alpha| = k$,

$$x \cdot \nabla(x^\alpha) = \sum_{j=1}^n x_j \partial(x^\alpha)/\partial x_j = \sum_{j=1}^n \alpha_j x^\alpha = k\,x^\alpha.$$

Consequently, $x \cdot \nabla g(x) = k\, g(x)$ and

$$\operatorname{div}(g(x)\, x) = (n+k)\, g(x) \quad \text{for all } x \in T.$$

This proves $\operatorname{div} \tau_{RT} = \operatorname{div}(g(x)\, x) + q_{k-1} = (n+k)\, g(x) + q_{k-1}$ for some $q_{k-1} \in P_{k-1}(T)$. The comparison with $\tau_{RT} = g(x)\, x + p_k$ leads to some polynomial remainder $r_k \in P_k(T; \mathbb{R}^n)$ in

$$\tau_{RT} = (n+k)^{-1} (\operatorname{div} \tau_{RT})\, x + r_k \quad \text{for all } x \in T.$$

In other words, since $\operatorname{div} \tau_{RT} \in P_k(T)$,

$$(n+k)(1-\Pi_k)\tau_{RT} = (1-\Pi_k)\, ((\operatorname{div} \tau_{RT})\, x) = (1-\Pi_k)\, ((\operatorname{div} \tau_{RT})\, (x-c))$$

for any constant vector c. For instance, the center of inertia $c = \operatorname{mid}(T)$ of T with diameter h_T satisfies $|x - \operatorname{mid}(T)| \leqslant (n/(n+1))\, h_T$ for all $x \in T$. This leads to

$$\begin{aligned}(n+k)\|\tau_{RT} - \Pi_k \tau_{RT}\| &= \|(1-\Pi_k)\, ((\operatorname{div} \tau_{RT})(x - \operatorname{mid}(T)))\,\| \\ &\leqslant \|(\operatorname{div} \tau_{RT})(x - \operatorname{mid}(T))\| \leqslant (n/(n+1))\, h_T\, \|\operatorname{div} \tau_{RT}\|.\end{aligned}$$

This proves $\|\tau_{RT} - \Pi_k \tau_{RT}\|_{L^2(T)} \leqslant \frac{n\ h_T}{(n+1)(n+k)} \|\operatorname{div} \tau_{RT}\|_{L^2(T)}$. □

References

1. Arbogast, T., Chen, Z.: On the implementation of mixed methods as nonconforming methods for second-order elliptic problems. Math. Comput. **64**(211), 943–972 (1995)
2. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications, Springer Series in Computational Mathematics, vol. 44. Springer, Heidelberg (2013)
3. Braess, D.: Finite Elements, 3rd edn. Cambridge University Press, Cambridge (2007)
4. Brenner, S.C., Scott, L.R.: The mathematical theory offinite element methods. In: Marsden, J.E., Sirovich, L., Antman, S.S. (eds.) Texts in Applied Mathematics, 3rd edn., vol. **15**. Springer, New York (2008)
5. Carstensen, C., Dond, A.K., Nataraj, N., Pani, A.K.: Error analysis of nonconforming and mixed FEMs for second-order linear non-selfadjoint and indefinite elliptic problems. Numer. Math. **133**(3), 557–597 (2016)
6. Carstensen, C., Gallistl, D., Schedensack, M.: L^2 best approximation of the elastic stress in the Arnold-Winther FEM. IMA J. Numer. Anal. **36**(3), 1096–1119 (2016)
7. Carstensen, C., Peterseim, D., Schedensack, M.: Comparison results of finite element methods for the Poisson model problem. SIAM J. Numer. Anal. **50**(6), 2803–2823 (2012)
8. Demlow, A.: Suboptimal and optimal convergence in mixed finite element methods. SIAM J. Numer. Anal. **39**(6), 1938–1953 (2002)
9. Douglas, J., Jr., Roberts, J.E.: Mixed finite element methods for second order elliptic problems. Mat. Apl. Comput. **1**(1), 91–103 (1982)
10. Douglas, J., Jr., Roberts, J.E.: Global estimates for mixed methods for second order elliptic equations. Math. Comp. **44**(169), 39–52 (1985)
11. Ern, A., Gudi, T., Smears, I., Vohralik, M.: Equivalence of local- and global-best approximations, a simple stable local commuting projector, and optimal hp approximation estimates in H(div). IMA J. Numer. Anal. 1–27 (2021). https://doi.org/10.1093/imanum/draa103
12. Ern, A., Guermond, J.-L.: Theory and Practices of Finite Elements, 1st edn. Springer-Verlag, New York (2004)

13. Ern, A., Guermond, J.-L.: Mollification in strongly Lipschitz domains with application to continuous and discrete de Rham complexes. Comput. Methods Appl. Math. **16**(1), 51–75 (2016)
14. Gastaldi, L., Nochetto, R.H.: On L^{∞}-accuracy of mixed finite element methods for second order elliptic problems. Mat. Apl. Comput. **7**(1), 13–39 (1988)
15. Huang, J., Xu, Y.: Convergence and complexity of arbitrary order adaptive mixed element methods for the Poisson equation. Sci. China Math. **55**(5), 1083–1098 (2012)
16. Payne, L.E., Weinberger, H.F.: An optimal Poincaré inequality for convex domains. Arch. Ration. Mech. Anal. **5**(1), 286–292 (1960)
17. Schatz, A.H., Wang, J.P.: Some new error estimates for Ritz-Galerkin methods with minimal regularity assumptions. Math. Comput. **65**(213), 19–27 (1996)