HABILITATIONSSCHRIFT

On the fast, efficient and robust numerical solution of partial differential equations in poro- and solid mechanics

Eingereicht an der Universität
Duisburg-Essen
Fakultät für Mathematik

von

Dr. Maria Lymbery

May 10, 2023

DOI: 10.17185/duepublico/81590
URN: urn:nbn:de:hbz:465-20240222-105832-9

Contents

INTRODUCTION

This introductory part contains a summary in which we briefly discuss the content of the 11 original research papers included in this habilitation thesis. The articles have been ordered into three groups depending on the main focus of their studies, namely **Quasi-static poroelasticity** (6 papers), **Dynamic plate vibration** (1 paper) and **Scalar elliptic problems** (4 papers).

Ten of these articles have been published in the following renowned international scientific journals

- Numer. Linear Algebra Appl.: Papers I, VIII
- Multiscale Model. Simul.: Paper II
- Math. Models Methods Appl. Sci.: Paper III
- Comput. Methods Appl. Mech. Eng.: Paper IV
- Math. Comp.: Paper V
- J. Sci. Comput.: Paper VI
- Comput. Math. with Appl.: Paper VII
- SIAM J. Sci. Comput.: Paper X
- Math. Comput. Simul.: Paper XI

There is also one conference proceeding paper IX included.

Papers I, II, III, IV, V, VI gathered in the first thematic group **Quasi-static poroelasticity** study the robust and efficient numerical solution of Biot's consolidation model and its generalizations, that are, the multiple-network poroelastic theory (MPET) model and the Biot-Brinkman model. Biot's theory dating back to 1941 is certainly the most popular theory used to describe the displacement and flow within a fluid-filled linearly elastic porous media and has found many major applications, e.g., in geomechanics and petroleum engineering. In 1960 double-porosity models, extending upon Biot's single fluid network to the case of two interacting networks, were utilized to describe the motion of liquids in fissured rocks and later, in the context of reservoir modelling, multiple-network poroelasticity equations were used to study the behaviour of elastic media permeated by multiple networks characterised by different porosities, permeabilities and/or interactions. The MPET equations have more recently been used to model the heart as well as the brain and central nervous system. The incorporation of viscous fluid effects in the generalized Biot-Brinkman model furthermore extends its applicability to more complicated biomedical processes such as the perfusion of the heart and glymphatic system of the brain. Solving numerically these poroelasticity models is a rather onerous task as the many physical parameters in practical applications exhibit extremely large variations.

The second group, **Dynamic plate vibration**, contains article VII. The considered biharmonic wave equation is commonly used to describe the linear reaction of thin structures (also called "plates") to external forces, the study of which is important, e.g., in civil and aeronautical engineering. This dynamic model investigating the propagation of waves in the plates as well as standing waves and vibration modes can be furthermore investigated as a prototype model for more sophisticated Kirchhoff-type equations, such as the Euler–Bernoulli equation describing the deflection of viscoelastic plates. The numerical solution of wave equations is a challenging task due to the fact that solutions may exhibit sharp fronts, as for example in case of shock waves, and possibly create complicated patterns of interference due to reflections at boundaries, in particular for complicated domains and inhomogeneous materials.

In the third group, **Scalar elliptic problems**, articles VIII, IX, X, XI are included. Although the computer systems have evolved significantly since the first electronic computer was developed in the 1940s, the topic of constructing fast solvers for linear systems of large dimensions continues to be very relevant nowadays as the understanding of "large" has also changed. Finding the numerical solution of a physical problem described in terms of differential equations with some prescribed accuracy may require solving a sparse algebraic system of millions or even billions of unknowns. This task can become even more intricate, e.g., when the time needed to solve such a system is also of high priority as is in many applications. While a lot has been done in this direction in the recent years, the construction of fast and efficient iterative solvers for certain classes of problems still remains an open question.

QUASI-STATIC POROELASTICITY

I **Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models.**

Qingguo Hong, Johannes Kraus, Maria Lymbery and Fadi Philo

Numer. Linear Algebra Appl. 26(4), (2019), e2242

Article I is the first on the topic of poroelasticity and is a joint work with Qingguo Hong, Johannes Kraus and Fadi Philo. The studied model here is the MPET model which is a generalization of the quasi-static Biot's consolidation model when more than one fluid networks are considered. Its flux based formulation for n networks reads as follows:

$$-\mathrm{div}\,\boldsymbol{\sigma} + \sum_{i=1}^{n} \alpha_i \nabla p_i = \boldsymbol{f} \quad \text{in} \quad \Omega \times (0,T), \tag{0.1a}$$

$$\boldsymbol{v}_i = -K_i \nabla p_i \quad \text{in} \quad \Omega \times (0,T), \tag{0.1b}$$

$$-\alpha_i \mathrm{div}\,\dot{\boldsymbol{u}} - \mathrm{div}\,\boldsymbol{v}_i - c_{p_i}\dot{p}_i - \sum_{\substack{j=i \\ j\neq i}}^{n} \beta_{ij}(p_i - p_j) = g_i \quad \text{in} \quad \Omega \times (0,T), \tag{0.1c}$$

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\epsilon}(\boldsymbol{u}) + \lambda \mathrm{div}(\boldsymbol{u})\boldsymbol{I}, \tag{0.1d}$$

$$\boldsymbol{\epsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T), \tag{0.1e}$$

where $i = 1, \ldots, n$ and the studied physical fields are the displacement $\boldsymbol{u}$, fluxes $\boldsymbol{v}_i$ and corresponding pressures p_i.

Here, $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ is an open domain, $\boldsymbol{f}$ describes the body force density in (0.1a), whereas g_i represent forced fluid extractions or injections into the medium in (0.1c).

The Biot-Willis coefficients α_i and the effective stress tensor $\boldsymbol{\sigma}$ appearing in the mass balance equation (0.1a) couple the pore pressures with the displacement variable $\boldsymbol{u}$. Each fluid flux $\boldsymbol{v}_i$ relates to a specific negative pressure gradient $-\nabla p_i$ via Darcy's law in (0.1b), in which the parameter tensors K_i describe the hydraulic conductivities which provide an indication of the general permeability of a porous medium. In the momentum balance equation, (0.1c), the time derivatives of the displacement $\boldsymbol{u}$ and the pressure variables p_i are denoted by $\dot{\boldsymbol{u}}$ and $\dot{p}_i$, respectively, the coupling coefficients β_{ij} designate the network transfer coefficients, hence, $\beta_{ij} = \beta_{ji}$ and the constants c_{p_i}, connected to the compressibility of each fluid, are called constrained specific storage coefficients.

It is assumed that Hooke's law (0.1d) holds where the effective strain tensor $\boldsymbol{\epsilon}(\boldsymbol{u})$ defined in (0.1e) is the symmetric part of the gradient of the displacement, here $\boldsymbol{I}$ denotes the identity tensor.

The Lamé parameters λ and μ in (0.1d) are determined in terms of the Young's modulus E and the Poisson ratio $\nu \in [0, 1/2)$ by

$$\lambda := (\nu E)/[(1+\nu)(1-2\nu)], \qquad \mu := E/[2(1+\nu)].$$

The following boundary, (0.2), and initial, (0.3), conditions complement system (0.1) and guarantee its well posedness:

$$\begin{aligned}
p_i(\boldsymbol{x},t) &= p_{i,D}(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x} \in \Gamma_{p_i,D}, \quad t>0, && \text{(0.2a)}\\
\boldsymbol{v}_i(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x}) &= q_{i,N}(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x} \in \Gamma_{p_i,N}, \quad t>0, && \text{(0.2b)}\\
\boldsymbol{u}(\boldsymbol{x},t) &= \boldsymbol{u}_D(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x} \in \Gamma_{\boldsymbol{u},D}, \quad t>0, && \text{(0.2c)}\\
(\boldsymbol{\sigma}(\boldsymbol{x},t) - \sum_{i=1}^{n} \alpha_i p_i \boldsymbol{I})\,\boldsymbol{n}(\boldsymbol{x}) &= \boldsymbol{g}_N(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x} \in \Gamma_{\boldsymbol{u},N}, \quad t>0, && \text{(0.2d)}
\end{aligned}$$

$$\begin{aligned}
p_i(\boldsymbol{x},0) &= p_{i,0}(\boldsymbol{x}) \qquad \boldsymbol{x} \in \Omega, && \text{(0.3a)}\\
\boldsymbol{u}(\boldsymbol{x},0) &= \boldsymbol{u}_0(\boldsymbol{x}) \qquad \boldsymbol{x} \in \Omega, && \text{(0.3b)}
\end{aligned}$$

where $i = 1, \ldots, n$ and it is satisfied $\Gamma_{p_i,D} \cap \Gamma_{p_i,N} = \emptyset$, $\overline{\Gamma}_{p_i,D} \cup \overline{\Gamma}_{p_i,N} = \Gamma = \partial\Omega$, $\Gamma_{\boldsymbol{u},D} \cap \Gamma_{\boldsymbol{u},N} = \emptyset$ and $\overline{\Gamma}_{\boldsymbol{u},D} \cup \overline{\Gamma}_{\boldsymbol{u},N} = \Gamma$.

The stable discretization and efficient solution of system (0.1) is a very challenging task due to the big number highly of varying parameters. In this article we construct novel parameter-matrix-dependent norms. Based on them for the first time the uniform Ladyzhenskaya–Babuška–Brezzi (LBB) stability for this class of problems is proven which ultimately facilitates the design of uniformly stable discretizations and parameter-robust preconditioners for flux-based formulations of multiporosity/multipermeability systems.

It is important to note that the stability estimates presented here are uniform not only with respect to the Lamé parameters but also to all the other model parameters, such as the storage coefficients c_{p_i}, the permeability coefficients K_i, the network transfer coefficients $\beta_{ij}, i, j = 1, \ldots, n$ the scale of the networks n and the time step size τ involved in the time discretization process.

The considered discretizations in article I are shown to be strongly mass-conservative and meet the required conditions for parameter-robust LBB stability. Furthermore, optimal error estimates have been proven and the foundation for optimal and fully robust iterative solution methods is laid via the construction of canonical (norm-equivalent) operator preconditioners on both continuous and discrete level.

Finally, the results of the performed numerical tests given in the included tables fully support the theoretical findings and further give a clear view of the practical potential of the developed technique.

II **Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems.**

Qingguo Hong, Johannes Kraus, Maria Lymbery, and Mary F. Wheeler

Multiscale Model. Simul. 18(2), (2020), 10.1137/19M1253988

Article II is a continuation of the work presented in article I where the focus is on the construction of fast iterative solvers for the arising block systems on both continuous and discrete level. For stability reasons, as before, we discretize system (0.1) in time by an implicit method which results in a coupled static problem in each time step. Among the established iterative schemes to solve coupled problems arising in poromechanics is the fixed-stress split iteration which often is the method of choice due to being unconditionally stable and computationally cheaper to other approaches.

In this work, performed in collaboration with Qingguo Hong, Johannes Kraus and Mary Wheeler, the fixed-stress split method is derived for the first time for the discretized in time flux based quasi-static MPET model on both continuous and discrete level and convergence analysis proving that the contraction rate of the fixed point iteration in both cases is independent of any model and discretization parameters is provided.

The series of numerical tests presented in this article not only fully justify the elaborated theory but also demonstrate the advantage of the fixed-stress split scheme over a preconditioned Minimal Residual (MinRes) solver accelerated by norm-equivalent preconditioning.

III **Parameter-robust Uzawa-type iterative methods for double saddle point problems arising in Biot's consolidation and multiple-network poroelasticity models.**

Qingguo Hong, Johannes Kraus, Maria Lymbery and Fadi Philo

Math. Models Methods Appl. Sci. 30(13), (2020), 2523-2555

The experience gathered while working on the previous two projects naturally led to deeper understanding but also rose more questions regarding the fast and efficient solution of quasi-static multiple-network poroelasticity equations describing flow in elastic porous media that is permeated by single or multiple fluid networks - can we further accelerate the iterative process by decoupling all three fields of interest, $\boldsymbol{u}$, $\boldsymbol{v}_i$ and p_i, which is not the case in the fixed-stress split iteration? What would be the thereby preconditioner defining the new scheme and how would it affect the performance of classical iterative schemes, e.g., the generalized minimal residual (GMRES) algorithm?

Starting point to answer these questions is the observation that the MPET equations (0.1) when written in abstract canonical form exhibit a double saddle point structure

$$\begin{bmatrix} A_1 & 0 & B_1^T \\ 0 & A_2 & B_2^T \\ B_1 & B_2 & -C \end{bmatrix}$$

with A_1 and A_2 being symmetric positive definite (SPD) operators and C a symmetric positive semidefinite (SPSD) operator.

In article III we propose an approach in which we augment and split this three-by-three block system in such a way that the resulting block Gauss–Seidel preconditioner defines a fully decoupled iterative scheme for the flux-, pressure-, and displacement fields, thereby obtaining an augmented Lagrangian Uzawa-type method. The theoretical study of this algorithm is the main contribution of this paper. We show that the rate of contraction of this fixed-point iteration is strictly less than one independently of all physical and discretization parameters which proves its parameter-robust uniform linear convergence.

All the results of the numerical tests we have performed with the newly developed Lagrangian Uzawa-type algorithm coincide with the theoretical expectations. Moreover, we have compared the performance of the fully decoupled scheme to the very popular partially decoupled fixed-stress split iterative method, which decouples only flow from the mechanics problem – the flux and pressure fields remain coupled in this case – and also to the preconditioned GMRES accelerated by the block-triangular preconditioner defining the new scheme. In terms of computational work, the obtained results clearly demonstrate the superiority of the new algorithm over these methods.

A scaling test indicating the robustness of the preconditioned GMRES and augmented Uzawa-type algorithms with respect to the number of networks is included which further supports the theoretical findings in this work performed in collaboration with Qingguo Hong, Johannes Kraus and Fadi Philo.

IV **Uniformly well-posed hybridized discontinuous Galerkin/hybrid mixed discretizations for Biot's consolidation model.**

Johannes Kraus, Philip L. Lederer, Maria Lymbery, Joachim Schöberl

Comput. Methods Appl. Mech. Eng. 384 (2021), 113991

Article IV is a joint work Johannes Kraus, Philip L. Lederer and Joachim Schöberl and is again motivated by the desire to construct fast and memory efficient solvers for poroelasticity systems. Subject of the study is the flux based formulation of the quasi-static Biot's consolidation model. To achieve Stokes and Darcy stability and pointwise mass conservation of the discrete model we use an $\mathbf{H}(\mathrm{div})$-conforming ansatz for $\mathbf{u}$ and $\mathbf{v}$ together with an appropriate pressure space.

We propose a new family of higher-order mass conserving hybridized/hybrid mixed FE discretizations based on the combination of a hybridized discontinuous Galerkin (DG) method for the elasticity subproblem with a mixed method for the flow subproblem, handled also by hybridization, which ultimately allows for a static condensation step. The latter eliminates the seepage velocity from the system while at the same time preserving mass conservation. A key point here is that the system to be finally solved contains only degrees of freedom (DOF) related to $\mathbf{u}$ and p which is a consequence of the hybridization process.

The performed theoretical analysis of the proposed discretization technique involves proper scaled norms and conclusively shows the well-posedness of the resulting continuous and discrete problems. Furthermore, it guides the construction of norm-equivalent preconditioners and the derivation of optimal near best approximation estimates.

To validate our theoretical findings we have performed a number of numerical tests which show the expected orders of convergence when increasing the degree of the FE approximation and the parameter-robustness of the proposed preconditioners. The last tests we have included clearly show that the new technique provides a very cost-efficient family of physics-oriented space discretizations for poroelasticity problems, especially for higher-order approximations.

V **A new practical framework for the stability analysis of perturbed saddle-point problems and applications.**

Qingguo Hong, Johannes Kraus, Maria Lymbery and Fadi Philo

Math. Comp. 92 (2023), 607-634

Proving stability results has been a very substantial part of the scientific work performed in the previous articles by the author of this habilitation. All the achievements have been possible largely due to the involvement of specially chosen norms which to many readers can seem hard to grasp and apply themselves in practice.

Therefore, there was the inevitable need to perform a deeper and more comprehensive analysis which ultimately resulted in the development of a new practical framework for the stability analysis of perturbed saddle-point problems which is the topic of article V written together with Quingguo Hong, Johannes Kraus and Fadi Philo.

In this paper we consider perturbed saddle-point problems characterized by an operator/matrix of the form

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix},$$

where A and C are SPSD operators and for them we prove a new abstract stability result which is based on a norm fitting technique.

The stability condition according to Babuška's theory is derived in a combined abstract norm from a small inf-sup condition, similar to the famous Ladyzhenskaya-Babuška-Brezzi (LBB) condition, and the other standard assumptions in Brezzi's theory. The combined norm itself is formed from individual fitted norms that are composed from proper seminorms.

This result is fundamental because it not only shows how simpler (shorter) proofs of many stability results can be derived but also guides the design of parameter-robust norm-equivalent preconditioners. All these benefits have been demonstrated in practice on mixed variational formulations of generalized Poisson, Stokes, vector Laplace and Biot's equations. Furthermore, the framework allows to analyze variational problems not only on a continuous but also on a discrete level as has been exploited in the convergence analysis of the discretizations subject to the following two articles.

VI **Robust Approximation of Generalized Biot-Brinkman Problems.**

Qingguo Hong, Johannes Kraus, Miroslav Kuchta, Maria Lymbery, Kent-André Mardal, Marie E. Rognes

J. Sci. Comput. volume 93, Article number: 77 (2022)

In article VI we focus on flux-based three-field formulations of the generalized Biot-Brinkman model which encompasses effects of fluid viscosity by replacing Darcy's law (0.1b) for $i = 1, \dots, n$ by the Brinkman equation

$$-\nu_i \operatorname{div} \boldsymbol{\epsilon}(\boldsymbol{v}_i) + \boldsymbol{v}_i + K_i \nabla p_i = \boldsymbol{r}_i \ \text{ in } \ \Omega \times (0, T),$$

where for a fixed network i, ν_i denotes the fluid viscosity, r_i represents an external flux and $\boldsymbol{\epsilon}(\boldsymbol{v}_i)$ is the symmetric part of the gradient of the velocity $\boldsymbol{v}_i$.

This model has the advantage that it extends on the Biot, in case $n = 1$, and multiple-network poroelasticity equations on the one hand and Brinkman flow problems on the other hand which allows to encompass a range of singular perturbation problems in realistic parameter regimes. Therefore, it can be applied to investigate more complex poromechanical interactions, e.g., in biophysics and engineering sciences.

In this article, which is a joint work with Qingguo Hong, Johannes Kraus, Miroslav Kuchta, Kent-André Mardal and Marie E. Rognes, a class of finite element discretizations for the generalized Biot-Brinkman equations is introduced and theoretically analyzed using the framework presented in article V. We demonstrate that the proposed three-field formulation of the Biot-Brinkman problem on a continuous and discrete level is uniformly well-posed under the chosen norms and the associated preconditioning strategy is robust with respect to the relevant parameter regimes.

The theoretical analysis is complemented by numerical experiments in the last section of the article whose results confirm the stability properties of the finite element discretization of the generalized Biot-Brinkman model and the robust behaviour of the derived preconditioner. The latter has been tested within the MinRes algorithm where for the solution of the displacement and flux subsystems required for the application of the preconditioner, a geometric multigrid solver implemented in FireDrake has been employed to accelerate the computations. Note that the auxliary multigrid method which has been proposed in article VIII provides an optimal solver for these tasks as well.

DYNAMIC PLATE VIBRATION

VII **C^1-conforming variational discretization of the biharmonic wave equation.**

Markus Bause, Maria Lymbery and Kevin Osthues

Comput. Math. with Appl. 119 (2022), 208-219

Article VII is a joint work with Markus Bause and Kevin Osthues. Subject of this study is the biharmonic wave equation which is of big importance to various applications including thin plate analyses. In contrary to articles I, II, III, IV, VI the model considered here is fully dynamic and its analysis and fast and accurate numerical solution require the involvement of different techniques.

The novelty here comes from the proposed numerical approximation using a C^1-conforming in space and time finite element ansatz which offers preservation of the smoothness properties of the solutions to the continuous evolution problem. We discretize in time using a combined Galerkin and collocation technique, while for space discretization we apply the Bogner–Fox–Schmit element and we prove optimal order error estimates.

We demonstrate the convergence and performance properties of the presented discretization technique by a series of numerical experiments with complex wave profiles in homogeneous and heterogeneous media which not only fully support the theoretical findings in this work but also show that the approach offers high potential for sophisticated multi-physics and/or multi-scale systems.

Based on the collected experience, the author of this habilitation thesis plans to apply similar tools also for the error analysis of the more complicated fully dynamic poroelasticicy systems and is currently preparing a paper on this topic.

SCALAR ELLIPTIC PROBLEMS

VIII **Auxiliary space multigrid method based on additive Schur complement approximation**

Johannes Kraus, Maria Lymbery and Svetozar Margenov

Numer. Linear Algebra Appl. 22(6), (2015) 965-986

The evolution of a certain quantity with time and/or according to a structure variable such as space is typically expressed by partial differential equations (PDE) and therefore they play a key role in diverse fields such as physics, chemistry, biology, economics, engineering, and life sciences. Finding the numerical solution of PDE after the discretization process typically reduces a continuous problem to a discrete problem that finally is given in the form of one or more systems of linear algebraic equations.

These systems of equations are typically very large and sparse which brings the need to construct fast and efficient iterative solvers. Among all iterative solvers, multigrid (MG) methods have earned a reputation as an efficient and versatile approach for solving such systems as they can solve them in a small uniformly bounded number of iterations, independent of the dimension of the discrete problem, up to a prescribed accuracy, allow for low-memory implementations and often can be parallelized.

In article VIII the idea of auxiliary space MG methods, which differ from classical MG methods in replacing coarse-grid correction by auxiliary space correction, is introduced. We perform a two-level block factorization of local matrices that are associated with a partitioning of the domain into overlapping or non-overlapping subdomains. The two-level auxiliary space method is based on a coarse-grid operator obtained from additive Schur complement approximation and is analyzed in the framework of auxiliary space preconditioning. We prove condition number estimates for both the additive Schur complement approximation and the two-level preconditioner, the latter implying robust convergence of the related two-grid method. We extend the two-level algorithm recursively in order to define the auxiliary space multigrid (ASMG) algorithm, where so-called Krylov cycles are considered.

To demonstrate the efficiency of the new algorithm for multiscale problems we have performed a representative collection of numerical tests. The obtained results not only show the efficiency of the proposed auxiliary space multigrid algorithm but also address further directions for development, e.g., incorporating different smoothers and transfer mappings or shifting the focus to different problem classes.

Article VIII is a joint work with Johannes Kraus and Svetozar Margenov.

IX **Auxiliary Space Multigrid Method for Elliptic Problems with Highly Varying Coefficients**

Johannes Kraus and Maria Lymbery

(2016) In: Dickopf, T., Gander, M., Halpern, L., Krause, R., Pavarino, L. (eds) Domain Decomposition Methods in Science and Engineering XXII. Lecture Notes in Computational Science and Engineering, vol 104. Springer.

In article IX written together with Johannes Kraus we summarize the main steps of the construction of the ASMG method presented in article VIII on a less technical level. We study in detail the spectral properties of the proposed additive Schur complement approximation for a class of elliptic problems with highly varying coefficients and provide a condition number estimate implying the robustness of the two-level method.

A crucial step in the application of the two-level preconditioner is the choice and realization of the operator $\Pi_{\widetilde{D}}$ defining the fictitious space preconditioner. We suggest two different variants where the efficient implementation of the second requires additionally the incorporation of an inner iterative method such as a

preconditioned conjugate gradient (PCG) method. For reasons of efficiency the latter requires a uniform preconditioner. We propose the scaled one-level additive Schwarz preconditioner for this purpose and prove a uniform condition number estimate.

The included numerical results fully support the theoretical analysis and moreover demonstrate that the choice of the surjective mapping $\Pi_{\widetilde{D}}$ affects crucially the performance of the nonlinear AMLI-cycle ASMG method.

X **Preconditioning Heterogeneous $H(\mathrm{div})$ Problems by Additive Schur Complement Approximation and Applications**

Johannes Kraus, Raytcho Lazarov, Maria Lymbery, Svetozar Margenov, and Ludmil Zikatanov

SIAM J. Sci. Comput. 38(2), (2016), 10.1137/140974092

In article X the AMLI-cycle ASMG is again in focus, this time with respect to systems arising from a mixed finite element approximation of second-order elliptic problems describing processes in highly heterogeneous media. First, we prove the stability of the continuous and discrete variational formulations with respect to the contrast of the media, defined as the ratio between the maximum and minimum values of the coefficient of the elliptic operator.

For the numerical solution of the discrete problem, we propose a new preconditioner for the MinRes algorithm where for the efficient solution of the weighted $H(\mathrm{div})$ subsystem a nonlinear ASMG algorithm is utilized. The performed numerical tests demonstrate the high quality of the preconditioner and its desired robustness with respect to the material contrast. Several representative test cases have been considered, one of which is related to the SPE10 (Society of Petroleum Engineers) benchmark problem.

This article is a joint work Johannes Kraus, Raytcho Lazarov, Svetozar Margenov and Ludmil Zikatanov.

XI **Incomplete factorization by local exact factorization (ILUE)**

Johannes Kraus and Maria Lymbery

Math. Comput. Simul. 145 (2018), 50-61.

In article XI which is co-authored with Johannes Kraus we develop a special preconditioning strategy for symmetric positive (semi-)definite SP(S)D matrices which we refer to as incomplete factorization by local exact factorization (ILUE).

This technique is influenced by and therefore shares similarities with the auxiliary space multigrid algorithm presented in articles VIII, XI, X in the sense that it utilizes a splitting of the domain into overlapping or non-overlapping subdomains as in ASMG, in the present context for computing exact LU decompositions of small-sized local matrices. We define the ILUE preconditioner and provide an estimate for its relative condition number.

We test the performance of the ILUE preconditioner within the PCG algorithm for linear systems arising from the finite element (FE) discretization of a second order elliptic boundary value problem in mixed formulation. The obtained numerical results clearly show the robust behaviour of the new algorithm and its advantage over the classical ILU(p) and ILUT(τ) incomplete factorization preconditioners even for problems with highly oscillatory permeability coefficients.

It should be noted that many coupled problems in computational mechanics are most efficiently solved by applying preconditioning techniques which finally require only fast solvers for elliptic problems either in form of a second-order self-adjoint problem or of a first-order system in mixed form. An example of such a situation is article IV in which the application of the parameter-robust preconditioner for the poroelasticity system requires the inversion of Laplacians only.

A project proposal on the topic of fast solvers for the fully dynamic and non-linear MPET model is under preparation.

PAPERS

CONSERVATIVE DISCRETIZATIONS AND PARAMETER-ROBUST PRECONDITIONERS FOR BIOT AND MULTIPLE-NETWORK FLUX-BASED POROELASTICITY MODELS

Received: 7 October 2018 | Revised: 25 February 2019 | Accepted: 8 March 2019

DOI: 10.1002/nla.2242

RESEARCH ARTICLE

WILEY

Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models

Qingguo Hong[1] | Johannes Kraus[2] | Maria Lymbery[2] | Fadi Philo[2]

[1]Department of Mathematics, Pennsylvania State University, State College, Pennsylvania

[2]Faculty of Mathematics, University of Duisburg-Essen, Essen, Germany

Correspondence
Johannes Kraus, Faculty of Mathematics, University of Duisburg-Essen, 45127 Essen, Germany.
Email: johannes.kraus@uni-due.de

Summary

The parameters in the governing system of partial differential equations of multiple-network poroelasticity models typically vary over several orders of magnitude, making its stable discretization and efficient solution a challenging task. In this paper, we prove the uniform Ladyzhenskaya–Babuška–Brezzi (LBB) condition and design uniformly stable discretizations and parameter-robust preconditioners for flux-based formulations of multiporosity/multipermeability systems. Novel parameter-matrix-dependent norms that provide the key for establishing uniform LBB stability of the continuous problem are introduced. As a result, the stability estimates presented here are uniform not only with respect to the Lamé parameter λ but also to all the other model parameters, such as the permeability coefficients K_i; storage coefficients c_{p_i}; network transfer coefficients $\beta_{ij}, i,j = 1, \dots, n$; the scale of the networks n; and the time step size τ. Moreover, strongly mass-conservative discretizations that meet the required conditions for parameter-robust LBB stability are suggested and corresponding optimal error estimates proved. The transfer of the canonical (norm-equivalent) operator preconditioners from the continuous to the discrete level lays the foundation for optimal and fully robust iterative solution methods. The theoretical results are confirmed in numerical experiments that are motivated by practical applications.

KEYWORDS

Biot's consolidation model, multiple-network poroelastic theory (MPET), parameter-robust LBB stability, robust norm-equivalent preconditioners, strongly mass-conservative discretization

1 | INTRODUCTION

Multiple-network poroelastic theory (MPET) has been introduced into geomechanics[1,2] to describe mechanical deformation and fluid flow in porous media as a generalization of Biot's theory.[3,4] The deformable elastic matrix is assumed to be permeated by multiple fluid networks of pores and fissures with differing porosity and permeability.

Abbreviations: MPET, multiple-network poroelastic theory; LBB, Ladyzhenskaya–Babuška–Brezzi

During the last decade, MPET has acquired many important applications in medicine and biomechanics and therefore become an active area of scientific research. The biological MPET model captures flow across scales and networks in soft tissue and can be used as an embedding platform for more specific models, for example, to describe water transport in the cerebral environment and to explore hypotheses defining the initiation and progression of both acute and chronic hydrocephalus.[5]

In the works of Vardakis et al.[6,7] multicompartmental poroelasticity models have been proposed to study the effects of obstructing cerebrospinal fluid (CSF) transport within an anatomically accurate cerebral environment and to demonstrate the impact of aqueductal stenosis and fourth ventricle outlet obstruction (FVOO). As a consequence, the efficacy of treating such clinical conditions by surgical procedures that focus on relieving the buildup of CSF pressure in the brain's third or fourth ventricles could be explored by means of computer simulations, which could also assist in finding medical indications of oedema formation.[8]

Recently, the MPET model has also been used to better understand the influence of biomechanical risk factors associated with the early stages of Alzheimer's disease (AD), the most common form of dementia.[9] Modeling transport of fluid within the brain is essential in order to discover the underlying mechanisms currently being investigated with regard to AD, such as the amyloid hypothesis, according to which the accumulation of neurotoxic amyloid-β (Aβ) into parenchymal senile plaques or within the walls of arteries is a root cause of this disease.

Biot and multiple-network poroelasticity models are computationally challenging because the physical parameters in practical applications exhibit extremely large variations. To give a few examples, comparing typical geophysical and biolophysical systems, permeabilities range from 10^{-9} to $10^{-21}\mathrm{m}^2$ and 10^{-7} to $10^{-16}\mathrm{m}^2$, a Poisson ratio from 0.1 to 0.3 and 0.3 to almost 0.5, respectively; see the works of Wang,[10] Lee et al.,[11] and Coussy.[12] Young's modulus in geomechanics is of the order of GPa, whereas in soft tissues, it is KPa; see the works of Smith et al.[13] and Støverud et al.[14]

In the multiple-network poroelasticity model, recently proposed in the work of Vardakis et al.,[7] describing fluid flow in the human brain, permeability also depends on network type. Transfer coefficients between different networks are very small and vary from $10^{-19}\mathrm{kg}/(\mathrm{m}\cdot\mathrm{s})$ to $10^{-13}\mathrm{kg}/(\mathrm{m}\cdot\mathrm{s})$. For that reason, it is important that the problem is well-posed and that numerical methods for its solution are stable over the whole range of values of the physical and discretization parameters.

The stability of discretizations by finite difference or finite volume methods for the Biot problem has been studied in other works.[15–18] We focus here on the design and analysis of uniform LBB stable discretizations for static multiple-network poroelasticity problems. It is well known that the well-posedness of saddle-point problems in their weak formulation, apart from the boundedness of the underlying bilinear form, relies on a stability estimate that is often referred to as the LBB condition.[19,20] The LBB condition[21,22] is also crucial in the analysis of stable discretizations and the derivation of a priori error estimates. Inf-sup stability for the Darcy problem, as well as the Stokes and linear elasticity problems, has been established under rather general conditions, and various stable mixed discretizations for either of these problems have been proposed over the years; see, for example, the work of Boffi et al.[19] and the references therein.

The fully parameter-robust stability of Biot's classical three-field formulation holding Darcy's law has been established only recently in the work of Hong et al.[23] Alternative formulations that can be proven to be stable include a two-field formulation[24,25] and a new three-field formulation introducing a total pressure as the third unknown aside from the displacement and fluid pressure.[11,26] A four-field formulation for the Biot model keeping the stress tensor as a variable is considered in the work of Lee,[27] where the analysis is robust with respect to the Lamé parameter λ, but not uniform with respect to parameter K. Another formulation of Biot's model based on $\boldsymbol{\sigma}$, $\boldsymbol{u}$, p, and a Lagrange multiplier to weakly impose the symmetry of the stress tensor $\boldsymbol{\sigma}$ has recently been proposed and analyzed in the work of Bærland et al.[28]

The first attempt to design and analyze parameter-robust stable discretizations for the MPET model is presented in the work of Lee et al.[29] Motivated by the works of Lee et al.[11] and Oyarzúa et al.,[26] Lee et al.[29] propose a mixed finite element formulation based on introducing an additional total pressure variable. They show that the formulation is robust in the limits of incompressibility, vanishing storage coefficients, and vanishing transfer between networks.

There are various discretizations for the classic three-field formulation of Biot's model that fit in the framework of full parameter-robust stability analysis presented in the work of Hong et al.[23] For example, the triplets $CR_l/RT_{l-1}/P^{\mathrm{dc}}_{l-1}(l=1,2)$ together with the stabilization techniques suggested in the works of Hansbo et al.[30] and Hu et al.[31] (see also the work of Fortin et al.[32]); the triplets $P_2/RT_0/P^{\mathrm{dc}}_0$ (in 2D) and $P^{\mathrm{stab}}_2/RT_0/P^{\mathrm{dc}}_0$ (in 3D); $P_2/RT_1/P^{\mathrm{dc}}_1$; the stabilized discretization, recently advocated in the work of Rodrigo et al[33]; or the finite element methods proposed in the work of Lee[34] would qualify for such parameter robustness. Coupling continuous or discontinuous Galerkin (DG) approximations of the solid displacement with a mixed method for the pressure, error estimates were obtained in the works of Phillips et al.[35,36] Following the theoretical framework presented in this paper, these discretizations can be applied to the MPET model.

A priori error estimates for the continuous-in-time scheme and discontinuous Galerkin spatial discretization, similar to the work of Hong et al.,[23] have been presented in the work of Kanschat et al.[37] for the Biot model. Inspired by the approach proposed in the work of Hong et al.[23] in the context of the static Biot problem; we analyze the MPET system using novel parameter-matrix-dependent norms. Furthermore, we exploit the same DG technology for discretizing the displacement field. The aim of this work is to establish the results regarding the parameter-robust stability of the weak formulation of the continuous problem, as well as the stability of strongly mass-conservative discretizations, corresponding error estimates, and parameter-robust preconditioners for the $(2n + 1)$-field formulation of the n-network problem. The presented stability results, error estimates, and preconditioners are independent of all model and discretization parameters including the Lamé parameter λ; permeability coefficients K_i; arbitrarily small or even vanishing storage coefficients c_{p_i}; network transfer coefficients $\beta_{ij}, i,j = 1, \dots, n$; the scale of the networks n; the time step size τ; and mesh size h. To our knowledge, these are the first fully parameter-robust stability results for the MPET model in a flux-based formulation.

The paper is organized as follows. In Section 2, the multiple-network poroelasticity model is stated in a flux-based formulation. The governing partial differential equations are then rescaled and the static boundary-value problem resulting from semidiscretization in time by the implicit Euler method is presented in its weak formulation in the beginning of Section 3. The proofs of the uniform boundedness and the parameter-robust inf-sup stability of the underlying bilinear form are the main results that follow in this section. Section 4 then discusses a class of uniformly stable and strongly mass-conservative mixed finite element discretizations that are based on $H(\div)$-conforming DG approximations of the displacement field. Boundedness and LBB stability are shown to be independent of all model and discretization parameters. In consequence, parameter-robust preconditioners and uniform optimal error estimates are provided. Section 5 is devoted to numerical tests underlining and validating the theoretical results of this work. Finally, Section 6 provides a brief conclusion.

2 | MODEL PROBLEM

In an open domain $\Omega \subset \mathbb{R}^d, d = 2, 3$, the unknown physical variables in the MPET flux-based model are the displacement $\boldsymbol{u}$, fluxes $\boldsymbol{v}_i$, and corresponding pressures $p_i\ i = 1, \dots, n$. The equations describing the model are as follows:

$$-\operatorname{div} \boldsymbol{\sigma} + \sum_{i=1}^{n} \alpha_i \nabla p_i = \boldsymbol{f} \text{ in } \Omega \times (0, T), \tag{1a}$$

$$\boldsymbol{v}_i = -K_i \nabla p_i \text{ in } \Omega \times (0, T), \quad i = 1, \dots, n, \tag{1b}$$

$$-\alpha_i \operatorname{div} \dot{\boldsymbol{u}} - \operatorname{div} \boldsymbol{v}_i - c_{p_i} \dot{p}_i - \sum_{\substack{j=i \\ j \neq i}}^{n} \beta_{ij}(p_i - p_j) = g_i \text{ in } \Omega \times (0, T), \quad i = 1, \dots, n, \tag{1c}$$

$$\boldsymbol{\sigma} = 2\mu \boldsymbol{\epsilon}(\boldsymbol{u}) + \lambda \operatorname{div}(\boldsymbol{u}) \boldsymbol{I}, \tag{1d}$$

$$\boldsymbol{\epsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T). \tag{1e}$$

In Equation (1d), λ and μ denote the Lamé parameters defined in terms of the modulus of elasticity (Young's modulus) E and the Poisson ratio $\nu \in [0, 1/2)$ by $\lambda := (\nu E)/[(1 + \nu)(1 - 2\nu)], \mu := E/[2(1 + \nu)]$. The constants α_i appearing in (1a) couple n pore pressures p_i with the displacement variable $\boldsymbol{u}$ and are known in the literature as Biot–Willis parameters. The corresponding right-hand side $\boldsymbol{f}$ describes the body force density. Each fluid flux $\boldsymbol{v}_i$ is related to a specific negative pressure gradient $-\nabla p_i$ via Darcy's law in (1b). The tensors K_i denote the hydraulic conductivities, which give an indication of the general permeability of a porous medium. In (1c), $\dot{\boldsymbol{u}}$ and $\dot{p}_i$ express the time derivatives of the displacement $\boldsymbol{u}$ and the pressure variables p_i. The constants c_{p_i} are referred to as the constrained specific storage coefficients and are connected to compressibility of each fluid; for more details, see for example the work of Showalter[38] and the references therein. The parameters β_{ij} are the network transfer coefficients coupling the network pressures[5]; hence, $\beta_{ij} = \beta_{ji}$. The source terms g_i in (1c) represent forced fluid extractions or injections into the medium.

It is assumed that the effective stress tensor $\boldsymbol{\sigma}$ satisfies Hooke's law (1d) where the effective strain tensor $\boldsymbol{\epsilon}(\boldsymbol{u})$ is given by the symmetric part of the gradient of the displacement field; see (1e). Here, $\boldsymbol{I}$ is used to denote the identity tensor.

The following boundary and initial conditions guarantee the well-posedness of system (1):

$$p_i(\boldsymbol{x},t) = p_{i,D}(\boldsymbol{x},t) \qquad \text{for } \boldsymbol{x} \in \Gamma_{p_i,D}, \quad t>0, \quad i=1,\ldots,n, \tag{2a}$$

$$\boldsymbol{v}_i(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x}) = q_{i,N}(\boldsymbol{x},t) \qquad \text{for } \boldsymbol{x} \in \Gamma_{p_i,N}, \quad t>0, \quad i=1,\ldots,n, \tag{2b}$$

$$\boldsymbol{u}(\boldsymbol{x},t) = \boldsymbol{u}_D(\boldsymbol{x},t) \qquad \text{for } \boldsymbol{x} \in \Gamma_{\boldsymbol{u},D}, \quad t>0, \tag{2c}$$

$$\left(\boldsymbol{\sigma}(\boldsymbol{x},t) - \sum_{i=1}^{n}\alpha_i p_i \boldsymbol{I}\right)\boldsymbol{n}(\boldsymbol{x}) = \boldsymbol{g}_N(\boldsymbol{x},t) \qquad \text{for } \boldsymbol{x} \in \Gamma_{\boldsymbol{u},N}, \quad t>0, \tag{2d}$$

where, for $i=1,\ldots,n$, it is fulfilled $\Gamma_{p_i,D}\cap\Gamma_{p_i,N}=\emptyset$, $\overline{\Gamma}_{p_i,D}\cup\overline{\Gamma}_{p_i,N}=\Gamma=\partial\Omega$, and $\Gamma_{\boldsymbol{u},D}\cap\Gamma_{\boldsymbol{u},N}=\emptyset$, $\overline{\Gamma}_{\boldsymbol{u},D}\cup\overline{\Gamma}_{\boldsymbol{u},N}=\Gamma$.

The initial conditions

$$p_i(\boldsymbol{x},0) = p_{i,0}(\boldsymbol{x}) \qquad \boldsymbol{x}\in\Omega, \quad i=1,\ldots,n, \tag{3a}$$

$$\boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}) \qquad \boldsymbol{x}\in\Omega \tag{3b}$$

at the time $t=0$ have to satisfy (1a).

The stress variable $\boldsymbol{\sigma}$ is eliminated from the MPET system by substituting the constitutive equation (1d) in (1a), thus obtaining a flux-based formulation of the MPET model.

To solve numerically the time-dependent problem, the backward Euler method is employed for time discretization resulting in the following system of time-step equations:

$$-2\mu\,\mathrm{div}\,\boldsymbol{\epsilon}(\boldsymbol{u}^k) - \lambda\nabla\mathrm{div}\;\boldsymbol{u}^k + \sum_{i=1}^{n}\alpha_i\nabla p_i^k = \boldsymbol{f}^k, \tag{4a}$$

$$K_i^{-1}\boldsymbol{v}_i^k + \nabla p_i^k = \boldsymbol{0}, \quad i=1,\ldots,n, \tag{4b}$$

$$-\alpha_i\mathrm{div}\,\boldsymbol{u}^k - \tau\mathrm{div}\,\boldsymbol{v}_i^k - c_{p_i}p_i^k - \tau\sum_{\substack{j=1\\ j\neq i}}^{n}\beta_{ij}\left(p_i^k - p_j^k\right) = g_i^k, \quad i=1,\ldots,n. \tag{4c}$$

The unknown time-step functions $\boldsymbol{u}^k, \boldsymbol{v}_i^k, p_i^k$ for $i=1,\ldots,n$ yield approximations of $\boldsymbol{u}, \boldsymbol{v}_i, p_i$ at a given time $t_k = t_{k-1}+\tau$:

$$\begin{aligned}
&\boldsymbol{u}(x,t_k) \approx \boldsymbol{u}^k \in \boldsymbol{u} := \left\{\boldsymbol{u}\in H^1(\Omega)^d \,:\, \boldsymbol{u}=\boldsymbol{u}_D \text{ on } \Gamma_{\boldsymbol{u},D}\right\},\\
&\boldsymbol{v}_i(x,t_k) \approx \boldsymbol{v}_i^k \in \boldsymbol{V}_i := \left\{\boldsymbol{v}_i\in H(\mathrm{div},\Omega) \,:\, \boldsymbol{v}_i\cdot\boldsymbol{n}=q_{i,N} \text{ on } \Gamma_{p_i,N}\right\},\\
&p_i(x,t_k) \approx p_i^k \in P_i := L^2(\Omega).
\end{aligned}$$

The right-hand side time-step functions are given by

$$\begin{aligned}
&\boldsymbol{f}^k = \boldsymbol{f}(x,t_k),\\
&g_i^k = -\tau g_i(x,t_k) - \alpha_i\mathrm{div}(\boldsymbol{u}^{k-1}) - c_{p_i}p_i^{k-1}, \quad i=1,\ldots,n.
\end{aligned}$$

Later, the static problem (4) is considered and, for convenience, the superscript for the time-step functions is skipped, that is, $\boldsymbol{u}^k, \boldsymbol{v}_i^k$, and p_i^k will be denoted by $\boldsymbol{u}, \boldsymbol{v}_i$, and p_i, respectively.

As usual, let $L^2(\Omega)$ be the space of square Lebesgue integrable functions equipped with the standard L^2 norm $\|\cdot\|$; $H^1(\Omega)^d$ denotes the space of vector-valued H^1-functions equipped with the norm $\|\cdot\|_1$ for which $\|\boldsymbol{u}\|_1^2 := \|\boldsymbol{u}\|^2 + \|\nabla\boldsymbol{u}\|^2$; $H(\mathrm{div};\Omega) := \{\boldsymbol{v}\in L^2(\Omega)^d \,:\, \mathrm{div}\,\boldsymbol{v}\in L^2(\Omega)\}$with norm $\|\cdot\|_{\mathrm{div}}$ defined by $\|\boldsymbol{v}\|_{\mathrm{div}}^2 := \|\boldsymbol{v}\|^2 + \|\mathrm{div}\boldsymbol{v}\|^2$. When the case $\Gamma_{\boldsymbol{u},D} = \Gamma_{p_i,N} = \Gamma$ and $\boldsymbol{u}_D = \boldsymbol{0}$, $q_{i,N}=0$ is considered, the notations $\boldsymbol{U} = H_0^1(\Omega)^d$ and $\boldsymbol{V}_i = H_0(\mathrm{div},\Omega)$, $i=1,\ldots,n$ are used. To guarantee the uniqueness of the solution for the pressure variables p_i, we set $P_i = L_0^2(\Omega) := \{p\in L^2(\Omega) \,:\, \int_\Omega p\,d\boldsymbol{x}=0\}$ for $i=1,\ldots,n$.

3 | STABILITY ANALYSIS

Before presenting the stability analysis, we perform a transformation of the governing system of partial differential equations with the aim of reducing the number of model parameters. One could additionally nondimensionalize the equations,[39] which, however, would not change the range of the parameters as considered in this paper.

First, the parameter μ is eliminated from the system by dividing Equation (4) by 2μ, that is, making the substitutions

$$2\mu \to 1, \frac{\lambda}{2\mu} \to \lambda, \frac{\alpha_i}{2\mu} \to \alpha_i, \frac{\boldsymbol{f}}{2\mu} \to \boldsymbol{f}, \frac{\tau}{2\mu} \to \tau, \frac{c_{p_i}}{2\mu} \to c_{p_i}, \frac{g_i}{2\mu} \to g_i, \quad i = 1, \ldots, n,$$

Equation (4) becomes

$$-\text{div}\ \boldsymbol{\epsilon}(\boldsymbol{u}) - \lambda \nabla \text{div}\ \boldsymbol{u} + \sum_{i=1}^{n} \alpha_i \nabla p_i = \boldsymbol{f}, \tag{5a}$$

$$K_i^{-1} \boldsymbol{v}_i + \nabla p_i = \boldsymbol{0}, \qquad i = 1, \ldots, n, \tag{5b}$$

$$-\alpha_i \text{div}\ \boldsymbol{u} - \tau \text{div}\ \boldsymbol{v}_i - c_{p_i} p_i - \tau \sum_{\substack{j=1 \\ j \neq i}}^{n} \beta_{ij}(p_i - p_j) = g_i, \qquad i = 1, \ldots, n. \tag{5c}$$

Next, Equation (5b) is multiplied by α_i and Equation (5c) by α_i^{-1} so that the substitutions $\tilde{\boldsymbol{v}}_i := \frac{\tau}{\alpha_i} \boldsymbol{v}_i$, $\tilde{p}_i := \alpha_i p_i$, $\tilde{g}_i := \frac{g_i}{\alpha_i}$ yield

$$-\text{div}\, \boldsymbol{\epsilon}(\boldsymbol{u}) - \lambda \nabla \text{div}\, \boldsymbol{u} + \sum_{i=1}^{n} \nabla \tilde{p}_i = \boldsymbol{f}, \tag{6a}$$

$$\tau^{-1} K_i^{-1} \alpha_i^2 \tilde{\boldsymbol{v}}_i + \nabla \tilde{p}_i = \boldsymbol{0}, \qquad i = 1, \ldots, n, \tag{6b}$$

$$-\text{div}\, \boldsymbol{u} - \text{div}\, \tilde{\boldsymbol{v}}_i - \frac{c_{p_i}}{\alpha_i^2} \tilde{p}_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(-\frac{\tau \beta_{ij}}{\alpha_i^2} \tilde{p}_i + \frac{\tau \beta_{ij}}{\alpha_i \alpha_j} \tilde{p}_j \right) = \tilde{g}_i, \qquad i = 1, \ldots, n. \tag{6c}$$

We define

$$R_i^{-1} = \tau^{-1} K_i^{-1} \alpha_i^2, \quad \alpha_{p_i} = \frac{c_{p_i}}{\alpha_i^2}, \quad \beta_{ii} = \sum_{\substack{j=1 \\ j \neq i}}^{n} \beta_{ij}, \quad \alpha_{ij} = \frac{\tau \beta_{ij}}{\alpha_i \alpha_j}, \ \ i, j = 1, \ldots, n$$

and make the rather general and reasonable assumptions that

$$\lambda > 0, \quad R_1^{-1}, \ldots, R_n^{-1} > 0, \quad \alpha_{p_1}, \ldots, \alpha_{p_n} \geq 0, \quad \alpha_{ij} \geq 0, \ \ i, j = 1, \ldots, n. \tag{7}$$

Making use of these substitutions, and, for convenience, skipping the "tilde" symbol, system (4a) becomes

$$-\text{div}\, \boldsymbol{\epsilon}(\boldsymbol{u}) - \lambda \nabla \text{div}\, \boldsymbol{u} + \sum_{i=1}^{n} \nabla p_i = \boldsymbol{f}, \tag{8a}$$

$$R_i^{-1} \boldsymbol{v}_i + \nabla p_i = \boldsymbol{0}, \qquad i = 1, \ldots, n, \tag{8b}$$

$$-\text{div}\, \boldsymbol{u} - \text{div}\, \boldsymbol{v}_i - (\alpha_{p_i} + \alpha_{ii}) p_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} \alpha_{ij} p_j = g_i, \qquad i = 1, \ldots, n, \tag{8c}$$

or

$$\mathcal{A} \left[\boldsymbol{u}^T, \boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_n^T, p_1, \ldots, p_n \right]^T = \left[\boldsymbol{f}^T, \boldsymbol{0}^T, \ldots, \boldsymbol{0}^T, g_1, \ldots, g_n \right]^T, \tag{9}$$

where

$$\mathcal{A} := \begin{bmatrix} -\text{div}\ \boldsymbol{\epsilon} - \lambda \nabla \text{div} & 0 & \ldots & \ldots & 0 & \nabla & \ldots & \ldots & \nabla \\ 0 & R_1^{-1} I & 0 & \ldots & 0 & \nabla & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & 0 \\ 0 & 0 & \ldots & 0 & R_n^{-1} I & 0 & \ldots & 0 & \nabla \\ -\text{div} & -\text{div} & 0 & \ldots & 0 & \tilde{\alpha}_{11} I & \alpha_{12} I & \ldots & \alpha_{1n} I \\ \vdots & 0 & \ddots & & \vdots & \alpha_{21} I & \ddots & & \alpha_{2n} I \\ \vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & \vdots \\ -\text{div} & 0 & \ldots & 0 & -\text{div} & \alpha_{n1} I & \alpha_{n2} I & \ldots & \tilde{\alpha}_{nn} I \end{bmatrix} \tag{10}$$

is the scaled operator and $\tilde{\alpha}_{ii} = -\alpha_{p_i} - \alpha_{ii}, i = 1, \ldots, n$.

For convenience, let $\boldsymbol{v}^T = (\boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_n^T)$, $\boldsymbol{p}^T = (p_1, \ldots, p_n)$, $\boldsymbol{z}^T = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)$, $\boldsymbol{q}^T = (q_1, \ldots, q_n)$, and $\boldsymbol{V} = \boldsymbol{V}_1 \times \ldots \times \boldsymbol{V}_n$, $\boldsymbol{P} = P_1 \times \ldots \times P_n$. With the boundary conditions, system (8a) has the following weak formulation: Find $(\boldsymbol{u}; \boldsymbol{v}; \boldsymbol{p}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$ such that, for any $(\boldsymbol{w}; \boldsymbol{z}; \boldsymbol{q}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$, there holds

$$(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda (\text{div}\, \boldsymbol{u}, \text{div}\, \boldsymbol{w}) - \sum_{i=1}^{n} (p_i, \text{div}\, \boldsymbol{w}) = (\boldsymbol{f}, \boldsymbol{w}) \tag{11a}$$

$$(R_i^{-1}\boldsymbol{v}_i, \boldsymbol{z}_i) - (p_i, \operatorname{div}\boldsymbol{z}_i) = 0, \quad i = 1, \ldots, n, \tag{11b}$$

$$-(\operatorname{div}\boldsymbol{u}, q_i) - (\operatorname{div}\boldsymbol{v}_i, q_i) - (\alpha_{p_i} + \alpha_{ii})(p_i, q_i) + \sum_{\substack{j=1 \\ j\neq i}}^{n} \alpha_{ij}(p_j, q_i) = (g_i, q_i), \quad i = 1, \ldots, n. \tag{11c}$$

Following the work of Lipnikov,[40] we first consider the following Hilbert spaces and weighted norms:

$$\boldsymbol{U} = H_0^1(\Omega)^d, \qquad (\boldsymbol{u}, \boldsymbol{w})_{\boldsymbol{u}} = (\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{w}), \tag{12}$$

$$\boldsymbol{V}_i = H_0(\operatorname{div}, \Omega), \qquad (\boldsymbol{v}_i, \boldsymbol{z}_i)_{\boldsymbol{V}_i} = (R_i^{-1}\boldsymbol{v}_i, \boldsymbol{z}_i) + \left(R_i^{-1}\operatorname{div}\boldsymbol{v}_i, \operatorname{div}\boldsymbol{z}_i\right), \qquad i = 1, \ldots, n, \tag{13}$$

$$P_i = L_0^2(\Omega), \qquad (p_i, q_i)_{P_i} = (p_i, q_i), \qquad i = 1, \ldots, n. \tag{14}$$

System (11), however, is not uniformly stable with respect to the parameters R_i^{-1} under these norms as shown in the work of Hong et al.[23] Therefore, proper parameter-dependent norms for the spaces $\boldsymbol{U}, \boldsymbol{V}_i, P_i, i = 1, \ldots, n$, have to be introduced that allow us to establish the parameter-robust stability of the MPET model (11) for parameters in the ranges presented in (7).

From experience, we know that the largest of the values $R_i^{-1}, i = 1, \ldots, n$ is important and we note that the term $(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{w}))$ dominates in the elasticity form when $\lambda \ll 1$. Hence, we define

$$R^{-1} = \max\left\{R_1^{-1}, \ldots, R_n^{-1}\right\}, \quad \lambda_0 = \max\{1, \lambda\}. \tag{15}$$

We introduce the following $n \times n$ matrices:

$$\Lambda_1 = \begin{bmatrix} \alpha_{11} & -\alpha_{12} & \ldots & -\alpha_{1n} \\ -\alpha_{21} & \alpha_{22} & \ldots & -\alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_{n1} & -\alpha_{n2} & \ldots & \alpha_{nn} \end{bmatrix}, \qquad \Lambda_2 = \begin{bmatrix} \alpha_{p_1} & 0 & \ldots & 0 \\ 0 & \alpha_{p_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \alpha_{p_n} \end{bmatrix},$$

$$\Lambda_3 = \begin{bmatrix} R & 0 & \ldots & 0 \\ 0 & R & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & R \end{bmatrix}, \qquad \Lambda_4 = \begin{bmatrix} \frac{1}{\lambda_0} & \cdots & \cdots & \frac{1}{\lambda_0} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{1}{\lambda_0} & \cdots & \cdots & \frac{1}{\lambda_0} \end{bmatrix}$$

through which the parameter-dependent norms are to be specified and analyzed. From the definition of $\alpha_{ij} = \frac{\tau\beta_{ij}}{\alpha_i\alpha_j}$, $\beta_{ii} = \sum_{\substack{j=1 \\ j\neq i}}^{n}\beta_{ij}$, and $\beta_{ij} = \beta_{ji}$, it is obvious that Λ_1 is symmetric positive semidefinite (SPSD). Because $\alpha_{p_i} \geq 0$, we have that Λ_2 is SPSD. Noting that $R > 0$, it follows that Λ_3 is symmetric positive definite (SPD). Moreover, it is obvious that Λ_4 is a rank-one matrix with eigenvalues $\lambda_i = 0, i = 1, \ldots, n-1$, and $\lambda_n = \frac{n}{\lambda_0}$.

Remark 1. Let $\boldsymbol{g}^T = (g_1, \ldots, g_n)$. It is convenient to assume that $\int_\Omega \boldsymbol{g}dx = \boldsymbol{0}$. This assumption, however, as we explain here is not restrictive for the following reason. If $\int_\Omega \boldsymbol{g}dx \neq \boldsymbol{0}$, then the "consistency condition" $\operatorname{rank}(\Lambda_1 + \Lambda_2) = \operatorname{rank}(\Lambda_g)$ has to be satisfied where $\operatorname{rank}(X)$ denotes the rank of a matrix X and $\Lambda_g = [\Lambda_1 + \Lambda_2, \boldsymbol{g}_c]$ is the matrix obtained by augmenting $\Lambda_1 + \Lambda_2$ with the column $\boldsymbol{g}_c = \frac{1}{|\Omega|}\int_\Omega \boldsymbol{g}dx$. In this case, there exists a vector $\boldsymbol{p}_c^T = (p_{1,c}, \ldots, p_{n,c}) \in \mathbb{R}^n$ such that $(\Lambda_1 + \Lambda_2)\boldsymbol{p}_c = \boldsymbol{g}_c$ (in many applications, $\Lambda_1 + \Lambda_2$ is invertible and $\boldsymbol{p}_c = (\Lambda_1 + \Lambda_2)^{-1}\boldsymbol{g}_c$). Hence, we can decompose $\boldsymbol{g} = \boldsymbol{g}_0 + \boldsymbol{g}_c$, where $\boldsymbol{g}_0 = \boldsymbol{g} - \frac{1}{|\Omega|}\int_\Omega \boldsymbol{g}dx$, and thus, $\int_\Omega \boldsymbol{g}_0 dx = \boldsymbol{0}$. Then, the solution $(\boldsymbol{u}; \boldsymbol{v}; \boldsymbol{p})$ can be decomposed according to $(\boldsymbol{u}; \boldsymbol{v}; \boldsymbol{p}) = (\boldsymbol{u}; \boldsymbol{v}; \boldsymbol{p}_0) + (\boldsymbol{0}; \boldsymbol{0}; \boldsymbol{p}_c)$, where $\boldsymbol{p}_0^T = (p_{1,0}, \ldots, p_{n,0}) \in L_0^2(\Omega) \times \ldots \times L_0^2(\Omega)$ and $\boldsymbol{p}_c$ is a basic solution of $(\Lambda_1 + \Lambda_2)\boldsymbol{p}_c = \boldsymbol{g}_c$. This decomposition shows that we only need to consider the case when $\int_\Omega \boldsymbol{g}dx = \boldsymbol{0}$.

Now, we introduce the SPD matrix

$$\Lambda = \sum_{i=1}^{4}\Lambda_i. \tag{16}$$

As we will see, Λ plays an important role in the definition of proper norms and the splitting (16) in our analysis. Furthermore, we summarize some useful properties of the matrix Λ in the following lemma.

Lemma 1. *Let $\tilde{\Lambda} = \Lambda_3 + \Lambda_4$, $\tilde{\Lambda}^{-1} = (\tilde{b}_{ij})_{n\times n}$; then, $\tilde{\Lambda}$ is SPD and, for any n-dimensional vector $\boldsymbol{x}$, we have*

$$(\Lambda\boldsymbol{x}, \boldsymbol{x}) \geq (\tilde{\Lambda}\boldsymbol{x}, \boldsymbol{x}) \geq (\Lambda_3\boldsymbol{x}, \boldsymbol{x}), \tag{17}$$

$$(\Lambda^{-1}\boldsymbol{x}, \boldsymbol{x}) \leq (\tilde{\Lambda}^{-1}\boldsymbol{x}, \boldsymbol{x}) \leq \left(\Lambda_3^{-1}\boldsymbol{x}, \boldsymbol{x}\right) = R^{-1}(\boldsymbol{x}, \boldsymbol{x}). \tag{18}$$

In addition,

$$0 < \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{b}_{ij} \le \lambda_0. \tag{19}$$

Proof. From the definitions of Λ_3, Λ_4, noting that Λ_3 is SPD and Λ_4 is SPSD, it is obvious that $\tilde{\Lambda}$ is SPD.

From the definition of Λ, noting that Λ_1 and Λ_2 are SPSD, we infer the estimates

$$(\Lambda \boldsymbol{x}, \boldsymbol{x}) \ge (\tilde{\Lambda}\boldsymbol{x}, \boldsymbol{x}) \ge (\Lambda_3 \boldsymbol{x}, \boldsymbol{x}), \qquad (\Lambda^{-1}\boldsymbol{x}, \boldsymbol{x}) \le (\tilde{\Lambda}^{-1}\boldsymbol{x}, \boldsymbol{x}) \le \left(\Lambda_3^{-1}\boldsymbol{x}, \boldsymbol{x}\right) = R^{-1}(\boldsymbol{x}, \boldsymbol{x}).$$

Next, we show that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{b}_{ij} \le \lambda_0.$$

From the definitions of Λ_3, Λ_4, and $\tilde{\Lambda}$, we have

$$\tilde{\Lambda} = \begin{bmatrix} R + \frac{1}{\lambda_0} & \frac{1}{\lambda_0} & \cdots & \frac{1}{\lambda_0} \\ \frac{1}{\lambda_0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{\lambda_0} \\ \frac{1}{\lambda_0} & \cdots & \frac{1}{\lambda_0} & R + \frac{1}{\lambda_0} \end{bmatrix}.$$

Now, using the Sherman–Morrison–Woodbury formula, we find

$$\tilde{\Lambda}^{-1} = \left(\Lambda_3 + \tilde{\boldsymbol{\lambda}}\boldsymbol{e}^T\right)^{-1} = \Lambda_3^{-1} - \frac{\Lambda_3^{-1}\tilde{\boldsymbol{\lambda}}\boldsymbol{e}^T\Lambda_3^{-1}}{1 + \boldsymbol{e}^T\Lambda_3^{-1}\tilde{\boldsymbol{\lambda}}}, \quad \text{where} \quad \tilde{\boldsymbol{\lambda}} = \left(\underbrace{\frac{1}{\lambda_0}, \ldots, \frac{1}{\lambda_0}}_{n}\right)^T, \quad \boldsymbol{e} = \underbrace{(1, \ldots, 1)}_{n}{}^T.$$

Furthermore, noting that

$$\Lambda_3^{-1} = \begin{bmatrix} \frac{1}{R} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{1}{R} \end{bmatrix} = \frac{1}{R} I_{n\times n},$$

where $I_{n\times n}$ is the $n \times n$ identity matrix, we obtain

$$\Lambda_3^{-1}\tilde{\boldsymbol{\lambda}}\boldsymbol{e}^T\Lambda_3^{-1} = \left(\frac{1}{R}I_{n\times n}\right)\begin{bmatrix} \frac{1}{\lambda_0} & \cdots & \cdots & \frac{1}{\lambda_0} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{1}{\lambda_0} & \cdots & \cdots & \frac{1}{\lambda_0} \end{bmatrix}\left(\frac{1}{R}I_{n\times n}\right) = \begin{bmatrix} \frac{1}{R^2\lambda_0} & \cdots & \cdots & \frac{1}{R^2\lambda_0} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{1}{R^2\lambda_0} & \cdots & \cdots & \frac{1}{R^2\lambda_0} \end{bmatrix}$$

and

$$\boldsymbol{e}^T\Lambda_3^{-1}\tilde{\boldsymbol{\lambda}} = (1, \ldots, \ldots, 1)\begin{bmatrix} \frac{1}{R} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{1}{R} \end{bmatrix}\begin{bmatrix} \frac{1}{\lambda_0} \\ \vdots \\ \vdots \\ \frac{1}{\lambda_0} \end{bmatrix} = \sum_{i=1}^{n}\frac{1}{R\lambda_0} = \frac{n}{R\lambda_0},$$

which implies that

$$\frac{1}{1 + \boldsymbol{e}^T\Lambda_3^{-1}\tilde{\boldsymbol{\lambda}}} = \frac{R\lambda_0}{R\lambda_0 + n}.$$

Now, we can calculate $\tilde{\Lambda}^{-1}$ as follows:

$$\tilde{\Lambda}^{-1} = \Lambda_3^{-1} - \frac{\Lambda_3^{-1}\tilde{\boldsymbol{\lambda}}\boldsymbol{e}^T\Lambda_3^{-1}}{1+\boldsymbol{e}^T\Lambda_3^{-1}\tilde{\boldsymbol{\lambda}}} = \begin{bmatrix} \frac{1}{R} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{R} \end{bmatrix} - \frac{R\lambda_0}{R\lambda_0+n} \begin{bmatrix} \frac{1}{R^2\lambda_0} & \cdots & \cdots & \frac{1}{R^2\lambda_0} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{1}{R^2\lambda_0} & \cdots & \cdots & \frac{1}{R^2\lambda_0} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{R} - \frac{1}{R(R\lambda_0+n)} & -\frac{1}{R(R\lambda_0+n)} & \cdots & -\frac{1}{R(R\lambda_0+n)} \\ -\frac{1}{R(R\lambda_0+n)} & \frac{1}{R} - \frac{1}{R(R\lambda_0+n)} & \cdots & -\frac{1}{R(R\lambda_0+n)} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{R(R\lambda_0+n)} & -\frac{1}{R(R\lambda_0+n)} & \cdots & \frac{1}{R} - \frac{1}{R(R\lambda_0+n)} \end{bmatrix} = (\tilde{b}_{ij})_{n\times n}.$$

Finally, we conclude that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{b}_{ij} = \frac{n}{R} - \frac{n^2}{R(R\lambda_0+n)} = \frac{nR\lambda_0+n^2-n^2}{R(R\lambda_0+n)} = \frac{n\lambda_0}{(R\lambda_0+n)} \le \frac{n\lambda_0}{n} = \lambda_0.$$

□

The crucial idea here is that we equip the Hilbert spaces $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}$ with parameter-matrix-dependent norms $\|\cdot\|_{\boldsymbol{U}}$, $\|\cdot\|_{\boldsymbol{V}}$, $\|\cdot\|_{\boldsymbol{P}}$ induced by the following *inner products*:

$$(\boldsymbol{u}, \boldsymbol{w})_{\boldsymbol{u}} = (\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{w}), \tag{20a}$$

$$(\boldsymbol{v}, \boldsymbol{z})_{\boldsymbol{V}} = \sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i, \boldsymbol{z}_i\right) + (\Lambda^{-1}\operatorname{Div}\boldsymbol{v}, \operatorname{Div}\boldsymbol{z}), \tag{20b}$$

$$(\boldsymbol{p}, \boldsymbol{q})_{\boldsymbol{P}} = (\Lambda\boldsymbol{p}, \boldsymbol{q}), \tag{20c}$$

where $\boldsymbol{p}^T = (p_1, \dots, p_n)$, $\boldsymbol{v}^T = (\boldsymbol{v}_1^T, \dots, \boldsymbol{v}_n^T)$, $(\operatorname{Div}\boldsymbol{v})^T = (\operatorname{div}\boldsymbol{v}_1, \dots, \operatorname{div}\boldsymbol{v}_n)$.

It is easy to show that (20a)–(20c) are indeed inner products on $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}$, respectively. It should be noted that $\operatorname{Div}\boldsymbol{v}$, $\operatorname{Div}\boldsymbol{z}$, and $\boldsymbol{p}, \boldsymbol{q}$ are vectors and the SPD matrix Λ is used to define the norms. These novel parameter-matrix-dependent norms play a key role in the analysis of the uniform stability of the MPET model. We further point out that, for $n = 1$, the norms defined by (20) are slightly different but equivalent to the norms that were used in the work of Hong et al.[23] to establish the parameter-robust inf-sup stability of the three-field formulation of Biot's model of consolidation.

The main result of this section is a proof of the uniform well-posedness of problem (11) under the norms induced by (20). Firstly, directly related to problem (11), we introduce the bilinear form

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) = (\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{w}) - \sum_{i=1}^{n}(p_i,\operatorname{div}\boldsymbol{w}) + \sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{z}_i\right) - \sum_{i=1}^{n}(p_i,\operatorname{div}\boldsymbol{z}_i)\\ - \sum_{i=1}^{n}(\operatorname{div}\boldsymbol{u},q_i) - \sum_{i=1}^{n}(\operatorname{div}\boldsymbol{v}_i,q_i) - \sum_{i=1}^{n}(\alpha_{p_i}+\alpha_{ii})(p_i,q_i) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\alpha_{ji}(p_j,q_i),\end{aligned}$$

which, in view of the definition of the matrices Λ_1 and Λ_2, can be written in the form

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) = (\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{w}) - \left(\sum_{i=1}^{n}p_i,\operatorname{div}\boldsymbol{w}\right) + \sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{z}_i\right) - (\boldsymbol{p},\operatorname{Div}\boldsymbol{z})\\ - \left(\operatorname{div}\boldsymbol{u},\sum_{i=1}^{n}q_i\right) - (\operatorname{Div}\boldsymbol{v},\boldsymbol{q}) - ((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{q}).\end{aligned}$$

Then, the following theorem shows the boundedness of $\mathcal{A}((\cdot;\cdot;\cdot),(\cdot;\cdot;\cdot))$ in the norms induced by (20).

Theorem 1. *There exists a constant C_b independent of the parameters $\lambda, R_i^{-1}, \alpha_{p_i}, \alpha_{ij}, i, j = 1, \dots, n$ and the network scale n such that, for any $(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}) \in \boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$, $(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}) \in \boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$,*

$$|\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}))| \le C_b(\|\boldsymbol{u}\|_{\boldsymbol{U}} + \|\boldsymbol{v}\|_{\boldsymbol{V}} + \|\boldsymbol{p}\|_{\boldsymbol{P}})(\|\boldsymbol{w}\|_{\boldsymbol{U}} + \|\boldsymbol{z}\|_{\boldsymbol{V}} + \|\boldsymbol{q}\|_{\boldsymbol{P}}).$$

Proof. From the definition of the bilinear form and by using Cauchy's inequality, we obtain

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &= (\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{w}) - \left(\sum_{i=1}^{n} p_i, \operatorname{div}\boldsymbol{w}\right)\\ &\quad + \sum_{i=1}^{n}(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{z}_i) - (\boldsymbol{p},\operatorname{Div}\boldsymbol{z}) - \left(\operatorname{div}\boldsymbol{u},\sum_{i=1}^{n} q_i\right) - (\operatorname{Div}\boldsymbol{v},\boldsymbol{q}) - ((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{q})\\ &\le \|\boldsymbol{\epsilon}(\boldsymbol{u})\|\|\boldsymbol{\epsilon}(\boldsymbol{w})\| + \lambda\|\operatorname{div}\boldsymbol{u}\|\|\operatorname{div}\boldsymbol{w}\| + \frac{1}{\sqrt{\lambda_0}}\|\sum_{i=1}^{n} p_i\|\sqrt{\lambda_0}\|\operatorname{div}\boldsymbol{w}\|\\ &\quad + \sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{v}_i\right)^{\frac{1}{2}}\left(R_i^{-1}\boldsymbol{z}_i,\boldsymbol{z}_i\right)^{\frac{1}{2}} + \|\Lambda^{\frac{1}{2}}\boldsymbol{p}\|\|\Lambda^{-\frac{1}{2}}\operatorname{Div}\boldsymbol{z}\| + \sqrt{\lambda_0}\|\operatorname{div}\boldsymbol{u}\|\frac{1}{\sqrt{\lambda_0}}\|\sum_{i=1}^{n} q_i\|\\ &\quad + \|\Lambda^{-\frac{1}{2}}\operatorname{Div}\boldsymbol{v}\|\|\Lambda^{\frac{1}{2}}\boldsymbol{q}\| + \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{p}\|\|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{q}\|.\end{aligned}$$

Then, another application of Cauchy's inequality, in view of the definition of Λ_4, yields

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &\le \|\boldsymbol{\epsilon}(\boldsymbol{u})\|\|\boldsymbol{\epsilon}(\boldsymbol{w})\| + \lambda\|\operatorname{div}\boldsymbol{u}\|\|\operatorname{div}\boldsymbol{w}\| + \|\Lambda_4^{\frac{1}{2}}\boldsymbol{p}\|\sqrt{\lambda_0}\|\operatorname{div}\boldsymbol{w}\|\\ &\quad + \left(\sum_{i=1}^{n}(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{v}_i)\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n}(R_i^{-1}\boldsymbol{z}_i,\boldsymbol{z}_i)\right)^{\frac{1}{2}} + \|\Lambda^{\frac{1}{2}}\boldsymbol{p}\|\|\Lambda^{-\frac{1}{2}}\operatorname{Div}\boldsymbol{z}\|\\ &\quad + \sqrt{\lambda_0}\|\operatorname{div}\boldsymbol{u}\|\|\Lambda_4^{\frac{1}{2}}\boldsymbol{q}\| + \|\Lambda^{-\frac{1}{2}}\operatorname{Div}\boldsymbol{v}\|\|\Lambda^{\frac{1}{2}}\boldsymbol{q}\| + \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{p}\|\|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{q}\|.\end{aligned}$$

□

Before we prove the uniform inf-sup condition for the MPET problem, we recall some well-known results.[19,22]

Lemma 2. *There exists a constant $\beta_v > 0$ such that*

$$\inf_{q\in P_i}\sup_{\boldsymbol{v}\in\boldsymbol{V}_i}\frac{(div\,\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{div}\|q\|} \ge \beta_d, \quad i=1,\ldots,n. \tag{21}$$

Moreover, for any $(q_1,\ldots,q_n)\in P_1\times\ldots\times P_n$, the sum $\sum_{i=1}^{n} q_i$ is in $L_0^2(\Omega)$ and the classical Stokes inf-sup condition[19] implies the following.

Lemma 3. *There exists a constant $\beta_s > 0$ such that*

$$\inf_{(q_1,\ldots,q_n)\in P_1\times\ldots\times P_n}\sup_{\boldsymbol{u}\in\boldsymbol{U}}\frac{\left(div\,\boldsymbol{u},\sum_{i=1}^{n} q_i\right)}{\|\boldsymbol{u}\|_1\|\sum_{i=1}^{n} q_i\|} \ge \beta_s. \tag{22}$$

We are now ready to prove the uniform LBB condition for $\mathcal{A}((\cdot;\cdot;\cdot),(\cdot;\cdot;\cdot))$ in the norms induced by (20).

Theorem 2. *There exists a constant $\omega > 0$ independent of the parameters $\lambda, R_i^{-1}, \alpha_{p_i}, \alpha_{ij}$ for all $i,j\in\{1,\ldots,n\}$, and independent of the number of networks n such that*

$$\inf_{(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p})\in\boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}}\sup_{(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})\in\boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}}\frac{\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}))}{(\|\boldsymbol{u}\|_{\boldsymbol{U}}+\|\boldsymbol{v}\|_{\boldsymbol{V}}+\|\boldsymbol{p}\|_{\boldsymbol{P}})(\|\boldsymbol{w}\|_{\boldsymbol{U}}+\|\boldsymbol{z}\|_{\boldsymbol{V}}+\|\boldsymbol{q}\|_{\boldsymbol{P}})} \ge \omega.$$

Proof. For any $(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}) = (\boldsymbol{u};\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n;p_1,\ldots,p_n)\in\boldsymbol{U}\times\boldsymbol{V}_1\times\ldots\times\boldsymbol{V}_n\times P_1\times\ldots\times P_n$, by Lemma 2, there exist

$$\boldsymbol{\psi}_i\in\boldsymbol{V}_i \text{ such that } \operatorname{div}\boldsymbol{\psi}_i = \sqrt{R}p_i \text{ and } \|\boldsymbol{\psi}_i\|_{\operatorname{div}} \le \beta_d^{-1}\sqrt{R}\|p_i\|, \quad i=1,\ldots,n, \tag{23}$$

and by Lemma 3, there exists

$$\boldsymbol{u}_0\in\boldsymbol{U} \text{ such that } \operatorname{div}\boldsymbol{u}_0 = \frac{1}{\sqrt{\lambda_0}}\left(\sum_{i=1}^{n} p_i\right), \quad \|\boldsymbol{u}_0\|_1 \le \beta_s^{-1}\frac{1}{\sqrt{\lambda_0}}\left\|\sum_{i=1}^{n} p_i\right\|. \tag{24}$$

Choose

$$\boldsymbol{w} = \delta\boldsymbol{u} - \frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0,\ \ \boldsymbol{z}_i = \delta\boldsymbol{v}_i - \sqrt{R}\boldsymbol{\psi}_i,\ \ i = 1, \ldots, n,\ \ \boldsymbol{q} = -\delta\boldsymbol{p} - \Lambda^{-1}\mathrm{Div}\,\boldsymbol{v}, \tag{25}$$

where δ is a positive constant to be determined later. Now, let us verify the boundedness of $(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})$ by $(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p})$ in the combined norm. Let $\boldsymbol{\psi}^T = (\boldsymbol{\psi}_1^T, \ldots, \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{T}})$ such that $\boldsymbol{z} = \delta\boldsymbol{v} - \sqrt{R}\boldsymbol{\psi}$.

Firstly, by (24), we have

$$\begin{aligned}
\left(\frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0, \frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right)_U &= \left(\epsilon\left(\frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right), \epsilon\left(\frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right)\right) + \lambda\left(\mathrm{div}\left(\frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right), \mathrm{div}\left(\frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right)\right)\\
&\le \frac{1}{\lambda_0}(\boldsymbol{\epsilon}(\boldsymbol{u}_0), \boldsymbol{\epsilon}(\boldsymbol{u}_0)) + (\mathrm{div}\,\boldsymbol{u}_0, \mathrm{div}\,\boldsymbol{u}_0) \le \frac{1}{\lambda_0}(\boldsymbol{\epsilon}(\boldsymbol{u}_0), \boldsymbol{\epsilon}(\boldsymbol{u}_0)) + \frac{1}{\lambda_0}\left(\sum_{i=1}^n p_i, \sum_{i=1}^n p_i\right)\\
&\le \frac{1}{\lambda_0}\beta_s^{-2}\frac{1}{\lambda_0}\|\sum_{i=1}^n p_i\|^2 + \frac{1}{\lambda_0}\|\sum_{i=1}^n p_i\|^2 \le \frac{1}{\lambda_0}\left(\beta_s^{-2}\frac{1}{\lambda_0} + 1\right)\|\sum_{i=1}^n p_i\|^2\\
&\le \frac{1}{\lambda_0}\left(\beta_s^{-2} + 1\right)\|\sum_{i=1}^n p_i\|^2 = \left(\beta_s^{-2} + 1\right)(\Lambda_4\boldsymbol{p}, \boldsymbol{p}) \le \left(\beta_s^{-2} + 1\right)\|\boldsymbol{p}\|_{\boldsymbol{P}}^2,
\end{aligned}$$

which implies that

$$\|\boldsymbol{w}\|_{\boldsymbol{u}} \le \delta\|\boldsymbol{u}\|_{\boldsymbol{u}} + \sqrt{\left(\beta_s^{-2} + 1\right)}\|\boldsymbol{p}\|_{\boldsymbol{P}}. \tag{26}$$

Secondly, by (18) and (23), we have

$$\begin{aligned}
\left(\sqrt{R}\boldsymbol{\psi}, \sqrt{R}\boldsymbol{\psi}\right)_{\boldsymbol{V}} &= \sum_{i=1}^n\left(R_i^{-1}\sqrt{R}\boldsymbol{\psi}_i, \sqrt{R}\boldsymbol{\psi}_i\right) + \left(\Lambda^{-1}\mathrm{Div}\left(\sqrt{R}\boldsymbol{\psi}\right), \mathrm{Div}\left(\sqrt{R}\boldsymbol{\psi}\right)\right)\\
&\le R\sum_{i=1}^n\left(R_i^{-1}\boldsymbol{\psi}_i, \boldsymbol{\psi}_i\right) + R^{-1}\left(\mathrm{Div}\left(\sqrt{R}\boldsymbol{\psi}\right), \mathrm{Div}\left(\sqrt{R}\boldsymbol{\psi}\right)\right) \le \sum_{i=1}^n(\boldsymbol{\psi}_i, \boldsymbol{\psi}_i) + (\mathrm{Div}\,\boldsymbol{\psi}, \mathrm{Div}\,\boldsymbol{\psi})\\
&= \sum_{i=1}^n\|\boldsymbol{\psi}_i\|^2 + \sum_{i=1}^n(\mathrm{div}\,\boldsymbol{\psi}_i, \mathrm{div}\,\boldsymbol{\psi}_i) = \sum_{i=1}^n\|\boldsymbol{\psi}_i\|_{\mathrm{div}}^2 \le \sum_{i=1}^n\beta_d^{-2}R\|p_i\|^2\\
&= \beta_d^{-2}R\|\boldsymbol{p}\|^2 \le \beta_d^{-2}\|\boldsymbol{p}\|_{\boldsymbol{P}}^2,
\end{aligned}$$

which implies that

$$\|\boldsymbol{z}\|_V \le \delta\|\boldsymbol{v}\|_V + \beta_d^{-1}\|\boldsymbol{p}\|_{\boldsymbol{P}}. \tag{27}$$

Thirdly, there holds

$$\|\boldsymbol{q}\|_{\boldsymbol{P}} \le \delta\|\boldsymbol{p}\|_{\boldsymbol{P}} + \|\boldsymbol{v}\|_V \tag{28}$$

because $(\Lambda^{-1}\mathrm{Div}\,\boldsymbol{v}, \Lambda^{-1}\mathrm{Div}\,\boldsymbol{v})_{\boldsymbol{P}} = (\mathrm{Div}\,\boldsymbol{v}, \Lambda^{-1}\mathrm{Div}\,\boldsymbol{v}) \le (\boldsymbol{v}, \boldsymbol{v})_{\boldsymbol{V}}$.

Collecting the estimates (26), (27), and (28), we obtain the desired boundedness estimate

$$\|\boldsymbol{w}\|_U + \|\boldsymbol{z}\|_V + \|\boldsymbol{q}\|_{\boldsymbol{P}} \le \left(\delta + 1 + \beta_d^{-1} + \beta_s^{-1}\right)(\|\boldsymbol{u}\|_U + \|\boldsymbol{v}\|_V + \|\boldsymbol{p}\|_{\boldsymbol{P}}).$$

Next, we show the coercivity of $\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}), (\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}))$. Using the definition of $\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}), (\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}))$ and that of $(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})$ from (25), we find

$$\begin{aligned}
\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}), (\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &= (\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\mathrm{div}\,\boldsymbol{u}, \mathrm{div}\,\boldsymbol{w}) - \left(\sum_{i=1}^n p_i, \mathrm{div}\,\boldsymbol{w}\right)\\
&\quad + \sum_{i=1}^n\left(R_i^{-1}\boldsymbol{v}_i, \boldsymbol{z}_i\right) - (\boldsymbol{p}, \mathrm{Div}\,\boldsymbol{z}) - \left(\mathrm{div}\,\boldsymbol{u}, \sum_{i=1}^n q_i\right) - (\mathrm{Div}\,\boldsymbol{v}, \boldsymbol{q}) - ((\Lambda_1 + \Lambda_2)\boldsymbol{p}, \boldsymbol{q})\\
&= \left(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}\left(\delta\boldsymbol{u} - \frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right)\right) + \lambda\left(\mathrm{div}\,\boldsymbol{u}, \mathrm{div}\left(\delta\boldsymbol{u} - \frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right)\right) - \left(\sum_{i=1}^n p_i, \mathrm{div}\left(\delta\boldsymbol{u} - \frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_0\right)\right)\\
&\quad + \sum_{i=1}^n\left(R_i^{-1}\boldsymbol{v}_i, \left(\delta\boldsymbol{v}_i - \sqrt{R}\boldsymbol{\psi}_i\right)\right) - \left(\mathrm{Div}\left(\delta\boldsymbol{v} - \sqrt{R}\boldsymbol{\psi}\right), \boldsymbol{p}\right) - \left(\underbrace{(\mathrm{div}\,\boldsymbol{u}, \ldots, \mathrm{div}\,\boldsymbol{u})}_{n}{}^T, -\delta\boldsymbol{p} - \Lambda^{-1}\mathrm{Div}\,\boldsymbol{v}\right)\\
&\quad - (\mathrm{Div}\,\boldsymbol{v}, -\delta\boldsymbol{p} - \Lambda^{-1}\mathrm{Div}\,\boldsymbol{v}) - \left((\Lambda_1 + \Lambda_2)\boldsymbol{p}, (-\delta\boldsymbol{p} - \Lambda^{-1}\mathrm{Div}\,\boldsymbol{v})\right).
\end{aligned}$$

Using (23) and (24), we therefore get

$$\begin{aligned}
\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &= \delta(\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{u})) - \frac{1}{\sqrt{\lambda_0}}(\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{u}_0)) + \delta\lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}) - \frac{\lambda}{\sqrt{\lambda_0}}(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}_0) - \delta\left(\sum_{i=1}^{n} p_i, \operatorname{div}\boldsymbol{u}\right) \\
&\quad + \frac{1}{\sqrt{\lambda_0}}\left(\sum_{i=1}^{n} p_i, \operatorname{div}\boldsymbol{u}_0\right) + \delta\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{v}_i\right) - \sqrt{R}\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{\psi}_i\right) - \delta(\operatorname{Div}\boldsymbol{v},\boldsymbol{p}) + \sqrt{R}(\operatorname{Div}\boldsymbol{\psi},\boldsymbol{p}) \\
&\quad + \delta\left(\underbrace{(\operatorname{div}\boldsymbol{u},\ \ldots,\operatorname{div}\boldsymbol{u})}_{n}^{T},\boldsymbol{p}\right) + \left(\Lambda^{-1}\underbrace{(\operatorname{div}\boldsymbol{u},\ \ldots,\operatorname{div}\boldsymbol{u})}_{n}^{T},\operatorname{Div}\boldsymbol{v}\right) + \delta(\boldsymbol{p},\operatorname{Div}\boldsymbol{v}) \\
&\quad + (\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\operatorname{Div}\boldsymbol{v}) + \delta((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{p}) + \left((\Lambda_1+\Lambda_2)\boldsymbol{p},\Lambda^{-1}\operatorname{Div}\boldsymbol{v}\right) \\
&= \delta(\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{u})) - \frac{1}{\sqrt{\lambda_0}}(\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{u}_0)) + \delta\lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}) - \frac{\lambda}{\lambda_0}\left(\operatorname{div}\boldsymbol{u},\sum_{i=1}^{n} p_i\right) + \frac{1}{\lambda_0}\left(\sum_{i=1}^{n} p_i,\sum_{i=1}^{n} p_i\right) \\
&\quad + \delta\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{v}_i\right) - \sqrt{R}\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{\psi}_i\right) + R\sum_{i=1}^{n}(p_i,p_i) + (\Lambda^{-1}\left(\underbrace{(\operatorname{div}\boldsymbol{u},\ \ldots,\operatorname{div}\boldsymbol{u})}_{n}^{T},\operatorname{Div}\boldsymbol{v}\right) \\
&\quad + (\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\operatorname{Div}\boldsymbol{v}) + \delta((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{p}) + \left((\Lambda_1+\Lambda_2)\boldsymbol{p},\Lambda^{-1}\operatorname{Div}\boldsymbol{v}\right).
\end{aligned}$$

Using Young's inequality, it follows that

$$\begin{aligned}
\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &\geq \delta(\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{u})) - \frac{1}{2}\frac{1}{\sqrt{\lambda_0}}\epsilon_1(\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{u})) - \frac{1}{2}\frac{1}{\sqrt{\lambda_0}}\epsilon_1^{-1}(\boldsymbol{\epsilon}(\boldsymbol{u}_0),\boldsymbol{\epsilon}(\boldsymbol{u}_0)) + \delta\lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}) \\
&\quad - \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}) - \frac{\lambda}{4\lambda_0^2}\left(\sum_{i=1}^{n} p_i,\sum_{i=1}^{n} p_i\right) + \frac{1}{\lambda_0}\left(\sum_{i=1}^{n} p_i,\sum_{i=1}^{n} p_i\right) + \delta\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{v}_i\right) - \frac{1}{2}\epsilon_2\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{v}_i\right) \\
&\quad - \frac{1}{2}\epsilon_2^{-1}R\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{\psi}_i,\boldsymbol{\psi}_i\right) + R\sum_{i=1}^{n}(p_i,p_i) - (\Lambda^{-1}(\operatorname{div}\boldsymbol{u},\ \ldots,\operatorname{div}\boldsymbol{u})^T,(\operatorname{div}\boldsymbol{u},\ \ldots,\operatorname{div}\boldsymbol{u})^T) \\
&\quad - \frac{1}{4}(\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\operatorname{Div}\boldsymbol{v}) + (\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\operatorname{Div}\boldsymbol{v}) + \delta((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{p}) \\
&\quad - \frac{1}{4}((\Lambda_1+\Lambda_2)\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\Lambda^{-1}\operatorname{Div}\boldsymbol{v}) - ((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{p}).
\end{aligned} \tag{29}$$

From the definition of Λ and noting that both Λ_3 and Λ_4 are SPSD, we conclude that

$$\begin{aligned}
(\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\operatorname{Div}\boldsymbol{v}) - \left((\Lambda_1+\Lambda_2)\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\Lambda^{-1}\operatorname{Div}\boldsymbol{v}\right) &= (\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\Lambda\Lambda^{-1}\operatorname{Div}\boldsymbol{v}) - \left(\Lambda^{-1}\operatorname{Div}\boldsymbol{v},(\Lambda_1+\Lambda_2)\Lambda^{-1}\operatorname{Div}\boldsymbol{v}\right) \\
&= \left(\Lambda^{-1}\operatorname{Div}\boldsymbol{v},(\Lambda_3+\Lambda_4)\Lambda^{-1}\operatorname{Div}\boldsymbol{v}\right) \geq 0.
\end{aligned} \tag{30}$$

Furthermore, by (19) from Lemma 1, we have that

$$\left(\Lambda^{-1}(\operatorname{div}\boldsymbol{u},\ \ldots,\operatorname{div}\boldsymbol{u})^T,(\operatorname{div}\boldsymbol{u},\ \ldots,\operatorname{div}\boldsymbol{u})^T\right) = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{b}_{ij}\right)(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}) \leq \lambda_0(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}). \tag{31}$$

Collecting (29)–(31), the estimates from (23) and (24), and noting that $\lambda_0 = \max\{\lambda, 1\}$, the proof continues as follows:

$$\begin{aligned}
\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &\geq \left(\delta - \frac{1}{2}\frac{1}{\sqrt{\lambda_0}}\epsilon_1\right)(\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{u})) - \frac{1}{2}\frac{1}{\sqrt{\lambda_0}}\epsilon_1^{-1}\beta_s^{-2}\frac{1}{\lambda_0}\left(\sum_{i=1}^{n} p_i,\sum_{i=1}^{n} p_i\right) + (\delta-1)\lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}) \\
&\quad + \frac{3}{4\lambda_0}\left(\sum_{i=1}^{n} p_i,\sum_{i=1}^{n} p_i\right) + \left(\delta - \frac{1}{2}\epsilon_2\right)\sum_{i=1}^{n}\left(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{v}_i\right) - \frac{1}{2}\epsilon_2^{-1}\sum_{i=1}^{n}(\boldsymbol{\psi}_i,\boldsymbol{\psi}_i) + R\sum_{i=1}^{n}(p_i,p_i) \\
&\quad - (\lambda_0-\lambda+\lambda)(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{u}) + \frac{1}{2}(\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\operatorname{Div}\boldsymbol{v}) + (\delta-1)((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{p}).
\end{aligned}$$

Let $\epsilon_1 := 2\beta_s^{-2}, \epsilon_2 := 2\beta_d^{-2}$. We note that $\lambda_0 = \max\{\lambda, 1\}$ and $(\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{u}) \le 2(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{u}))$ for all $\boldsymbol{u} \in H_0^1(\Omega)^d$ to obtain

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) \ge{}& \left(\delta - \beta_s^{-2} - 2\right)(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{u})) - \frac{1}{4\lambda_0}\left(\sum_{i=1}^n p_i, \sum_{i=1}^n p_i\right) + (\delta-2)\lambda(\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{u}) + \frac{3}{4\lambda_0}\left(\sum_{i=1}^n p_i, \sum_{i=1}^n p_i\right)\\ &+ \left(\delta - \beta_d^{-2}\right)\sum_{i=1}^n \left(R_i^{-1}\boldsymbol{v}_i, \boldsymbol{v}_i\right) - \frac{1}{4}R\sum_{i=1}^n (p_i, p_i) + R\sum_{i=1}^n (p_i, p_i) + \frac{1}{2}(\Lambda^{-1}\operatorname{Div}\boldsymbol{v}, \operatorname{Div}\boldsymbol{v})\\ &+ (\delta-1)((\Lambda_1+\Lambda_2)\boldsymbol{p}, \boldsymbol{p}),\end{aligned}$$

or equivalently,

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) \ge{}& \left(\delta - \beta_s^{-2} - 2\right)(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{u})) + (\delta-2)\lambda(\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{u}) + \frac{1}{2}(\Lambda_4\boldsymbol{p}, \boldsymbol{p})\\ &+ \left(\delta - \beta_d^{-2}\right)\sum_{i=1}^n \left(R_i^{-1}\boldsymbol{v}_i, \boldsymbol{v}_i\right) + \frac{3}{4}(\Lambda_3\boldsymbol{p}, \boldsymbol{p}) + \frac{1}{2}(\Lambda^{-1}\operatorname{Div}\boldsymbol{v}, \operatorname{Div}\boldsymbol{v}) + (\delta-1)((\Lambda_1+\Lambda_2)\boldsymbol{p}, \boldsymbol{p}).\end{aligned}$$

Finally, let $\delta := \max\{\beta_s^{-2} + 2 + \frac{1}{2}, \beta_d^{-2} + \frac{1}{2}\}$. Then, using the definition of Λ, we get the desired coercivity estimate

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &= \left(\delta - \beta_s^{-2} - 2\right)(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{u})) + (\delta-2)\lambda(\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{u}) + \left(\delta - \beta_d^{-2}\right)\sum_{i=1}^n \left(R_i^{-1}\boldsymbol{v}_i, \boldsymbol{v}_i\right)\\ &\quad + \frac{1}{2}(\Lambda^{-1}\operatorname{Div}\boldsymbol{v}, \operatorname{Div}\boldsymbol{v}) + \left(\left((\delta-1)(\Lambda_1+\Lambda_2) + \frac{3}{4}\Lambda_3 + \frac{1}{2}\Lambda_4\right)\boldsymbol{p}, \boldsymbol{p}\right)\\ &\ge \frac{1}{2}\left(\|\boldsymbol{u}\|_{\boldsymbol{u}}^2 + \|\boldsymbol{v}\|_{\boldsymbol{V}}^2 + \|\boldsymbol{p}\|_{\boldsymbol{P}}^2\right).\end{aligned}$$

□

The above theorem implies the following stability result.

Corollary 1. *Let $(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}) \in \boldsymbol{U}\times\boldsymbol{V}\times P$ be the solution of (11). Then, there holds the estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{U}} + \|\boldsymbol{v}\|_{\boldsymbol{V}} + \|\boldsymbol{p}\|_{\boldsymbol{P}} \le C_1(\|\boldsymbol{f}\|_{\boldsymbol{U}^*} + \|\boldsymbol{g}\|_{\boldsymbol{P}^*}), \tag{32}$$

for some positive constant C_1 that is independent of the parameters $\lambda, R_i^{-1}, \alpha_{p_i}, \alpha_{ij}, i, j = 1, \dots, n$ and the network scale n, where $\|\boldsymbol{f}\|_{\boldsymbol{U}^*} = \sup_{\boldsymbol{w}\in\boldsymbol{U}}\frac{(\boldsymbol{f},\boldsymbol{w})}{\|\boldsymbol{w}\|_{\boldsymbol{U}}}$, $\|\boldsymbol{g}\|_{\boldsymbol{P}^*} = \sup_{\boldsymbol{q}\in\boldsymbol{P}}\frac{(\boldsymbol{g},\boldsymbol{q})}{\|\boldsymbol{q}\|_{\boldsymbol{P}}} = \|\Lambda^{-\frac{1}{2}}\boldsymbol{g}\|$.

Remark 2. We want to emphasize that the parameter ranges as specified in (7) are indeed relevant because the variations of the model parameters are quite large in many applications. For that reason, Theorem 1 and Theorem 2 are fundamental results that provide the parameter-robust stability of the model (11a)– (11c). We also point out that the matrix technique plays an interesting role for proving the uniform stability.

Remark 3. Let $\Lambda = (\gamma_{ij})_{n\times n}, \Lambda^{-1} = (\tilde{\gamma}_{ij})_{n\times n}$ and define

$$\mathscr{B} := \begin{bmatrix} \mathscr{B}_{\boldsymbol{u}}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathscr{B}_{\boldsymbol{v}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathscr{B}_{\boldsymbol{p}}^{-1} \end{bmatrix}, \tag{33}$$

where

$$\mathscr{B}_{\boldsymbol{u}} = -\operatorname{div}\boldsymbol{\epsilon} - \lambda\nabla\operatorname{div},$$

$$\mathscr{B}_{\boldsymbol{v}} = \begin{bmatrix} R_1^{-1}I & 0 & \dots & 0 \\ 0 & R_2^{-1}I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & R_n^{-1}I \end{bmatrix} - \begin{bmatrix} \tilde{\gamma}_{11}\nabla\operatorname{div} & \tilde{\gamma}_{12}\nabla\operatorname{div} & \dots & \tilde{\gamma}_{1n}\nabla\operatorname{div} \\ \tilde{\gamma}_{21}\nabla\operatorname{div} & \tilde{\gamma}_{22}\nabla\operatorname{div} & \dots & \tilde{\gamma}_{2n}\nabla\operatorname{div} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\gamma}_{n1}\nabla\operatorname{div} & \tilde{\gamma}_{n2}\nabla\operatorname{div} & \dots & \tilde{\gamma}_{nn}\nabla\operatorname{div} \end{bmatrix},$$

and

$$\mathscr{B}_{\boldsymbol{p}} = \begin{bmatrix} \gamma_{11}I & \gamma_{12}I & \dots & \gamma_{1n}I \\ \gamma_{21}I & \gamma_{22}I & \dots & \gamma_{2n}I \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1}I & \gamma_{n2}I & \dots & \gamma_{nn}I \end{bmatrix}.$$

Inferring from the theory presented in the work of Mardal et al.,[41] Theorems 1 and 2 imply that the operator $\mathscr{B}$ defined in (33) is a uniform norm-equivalent (canonical) block-diagonal preconditioner for the operator $\mathscr{A}$ in (10), robust in all model and discretization parameters, that is, $\kappa(\mathscr{B}\mathscr{A}) = \mathcal{O}(1)$.

4 | UNIFORMLY STABLE AND STRONGLY MASS-CONSERVATIVE DISCRETIZATIONS

In recent years, DG methods have been developed to solve various problems,[42–46] and some unified analysis for a finite element including DG methods has recently been presented in the works of Hong et al.[47,48] In this section, as motivated by the works of Schötzau et al.[49] and Hong et al.,[50] we propose discretizations of the MPET model problem (11). These discretizations preserve the divergence condition (namely Equation (8c)) pointwise, which results in a strong conservation of mass (see Proposition 1). Furthermore, they are also locking free when the Lamé parameter λ tends to ∞.[51]

4.1 | Preliminaries and notation

Let $\mathcal{T}_h$ be a shape-regular triangulation of mesh-size h of the domain Ω into triangles $\{T\}$ and define the set of all interior edges (or faces) of $\mathcal{T}_h$ by $\mathcal{E}_h^I$ and the set of all boundary edges (or faces) by $\mathcal{E}_h^B$. Let $\mathcal{E}_h = \mathcal{E}_h^I \cup \mathcal{E}_h^B$.

For $s \geq 1$, we introduce the spaces

$$H^s(\mathcal{T}_h) = \{\phi \in L^2(\Omega), \text{ such that } \phi|_T \in H^s(T) \text{ for all } T \in \mathcal{T}_h\}.$$

We further define some trace operators. Denote by $e = \partial T_1 \cap \partial T_2$ the common boundary (interface) of two subdomains T_1 and T_2 in $\mathcal{T}_h$, and by $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$, the unit normal vectors to e that point to the exterior of T_1 and T_2, correspondingly. For any $e \in \mathcal{E}_h^I$ and $q \in H^1(\mathcal{T}_h)$, $\boldsymbol{v} \in H^1(\mathcal{T}_h)^d$ and $\boldsymbol{\tau} \in H^1(\mathcal{T}_h)^{d\times d}$, the averages are defined as

$$\{\boldsymbol{v}\} = \frac{1}{2}(\boldsymbol{v}|_{\partial T_1 \cap e} \cdot \boldsymbol{n}_1 - \boldsymbol{v}|_{\partial T_2 \cap e} \cdot \boldsymbol{n}_2), \quad \{\boldsymbol{\tau}\} = \frac{1}{2}(\boldsymbol{\tau}|_{\partial T_1 \cap e}\boldsymbol{n}_1 - \boldsymbol{\tau}|_{\partial T_2 \cap e}\boldsymbol{n}_2),$$

and the jumps are given by

$$[q] = q|_{\partial T_1 \cap e} - q|_{\partial T_2 \cap e}, \quad [\boldsymbol{v}] = \boldsymbol{v}|_{\partial T_1 \cap e} - \boldsymbol{v}|_{\partial T_2 \cap e}.$$

When $e \in \mathcal{E}_h^B$, then the above quantities are defined as

$$\{\boldsymbol{v}\} = \boldsymbol{v}|_e \cdot \boldsymbol{n}, \quad \{\boldsymbol{\tau}\} = \boldsymbol{\tau}|_e\boldsymbol{n}, \quad [q] = q|_e, \quad [\boldsymbol{v}] = \boldsymbol{v}|_e.$$

If $\boldsymbol{n}_T$ is the outward unit normal to ∂T, it is easy to show that, for $\boldsymbol{\tau} \in H^1(\Omega)^{d\times d}$ and for all $\boldsymbol{v} \in H^1(\mathcal{T}_h)^d$, we have

$$\sum_{T\in\mathcal{T}_h} \int_{\partial T} (\boldsymbol{\tau}\boldsymbol{n}_T) \cdot \boldsymbol{v} ds = \sum_{e\in\mathcal{E}_h} \int_e \{\boldsymbol{\tau}\} \cdot [\boldsymbol{v}] ds. \tag{34}$$

4.2 | DG discretization

The finite element spaces we consider are denoted by

$$\begin{aligned}
\boldsymbol{U}_h &= \{\boldsymbol{u} \in H(\mathrm{div};\Omega) : \boldsymbol{u}|_T \in \boldsymbol{U}(T),\ T \in \mathcal{T}_h;\ \boldsymbol{u}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\},\\
\boldsymbol{V}_{i,h} &= \{\boldsymbol{v} \in H(\mathrm{div};\Omega) : \boldsymbol{v}|_T \in \boldsymbol{V}_i(T),\ T \in \mathcal{T}_h;\ \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\},\quad i = 1, \ldots, n,\\
P_{i,h} &= \left\{ q \in L^2(\Omega) : q|_T \in Q_i(K),\ T \in \mathcal{T}_h;\ \int_\Omega q dx = 0 \right\},\quad i = 1, \ldots, n.
\end{aligned}$$

The discretizations we analyze in the present context define the local spaces $\boldsymbol{U}(T)/\boldsymbol{V}_i(T)/Q_i(T)$ via the triplets $BDM_l(T)/RT_{l-1}(T)/P_{l-1}(T)$, or $BDFM_l(T)/RT_{l-1}(T)/P_{l-1}(T)$ for $l \geq 1$. Note that, for each of these choices, the important condition $\mathrm{div}\,\boldsymbol{U}(T) = \mathrm{div}\,\boldsymbol{V}_i(T) = Q_i(T)$ is satisfied.

Note that the normal component of any $\boldsymbol{u} \in \boldsymbol{U}_h$ is continuous on the internal edges and vanishes on the boundary edges. Then, for all $e \in \mathcal{E}_h$ and for all $\boldsymbol{\tau} \in H^1(\mathcal{T}_h)^d, \boldsymbol{u} \in \boldsymbol{U}_h$, it holds

$$\int_e [\boldsymbol{u}_n] \cdot \boldsymbol{\tau} ds = 0, \quad \text{implying that} \quad \int_e [\boldsymbol{u}] \cdot \boldsymbol{\tau} ds = \int_e [\boldsymbol{u}_t] \cdot \boldsymbol{\tau} ds, \tag{35}$$

where $\boldsymbol{u}_n$ and $\boldsymbol{u}_t$ denote the normal and tangential component of $\boldsymbol{u}$, respectively.

Similar to the continuous problem, we denote

$$\boldsymbol{v}_h^T = \left(\boldsymbol{v}_{1,h}^T, \dots \boldsymbol{v}_{n,h}^T\right), \quad \boldsymbol{p}_h^T = (p_{1,h}, \dots, p_{n,h}), \quad \boldsymbol{z}_h^T = \left(\boldsymbol{z}_{1,h}^T, \dots \boldsymbol{z}_{n,h}^T\right),$$

$$\boldsymbol{q}_h^T = (q_{1,h}, \dots, q_{n,h}), \quad \boldsymbol{V}_h = \boldsymbol{V}_{1,h} \times \dots \times \boldsymbol{V}_{n,h}, \quad \boldsymbol{P}_h = P_{1,h} \times \dots \times P_{n,h}.$$

With this notation at hand, the discretization of the variational problem (11) is given as follows: Find $(\boldsymbol{u}_h; \boldsymbol{v}_h; \boldsymbol{p}_h,) \in \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h$ such that, for any $(\boldsymbol{w}_h; \boldsymbol{z}_h; \boldsymbol{q}_h) \in \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h$ and $i = 1, \dots, n$,

$$a_h(\boldsymbol{u}_h, \boldsymbol{w}_h) + \lambda(\operatorname{div} \boldsymbol{u}_h, \operatorname{div} \boldsymbol{w}_h) - \sum_{i=1}^{n} (p_{i,h}, \operatorname{div} \boldsymbol{w}_h) = (\boldsymbol{f}, \boldsymbol{w}_h), \tag{36a}$$

$$\left(R_i^{-1} \boldsymbol{v}_{i,h}, \boldsymbol{z}_{i,h}\right) - (p_{i,h}, \operatorname{div} \boldsymbol{z}_{i,h}) = 0, \tag{36b}$$

$$-(\operatorname{div} \boldsymbol{u}_h, q_{i,h}) - (\operatorname{div} \boldsymbol{v}_{i,h}, q_{i,h}) + \tilde{\alpha}_{ii}(p_{i,h}, q_{i,h}) + \sum_{\substack{j=1 \\ j \neq i}}^{n} \alpha_{ij}(p_{j,h}, q_{i,h}) = (g_i, q_{i,h}), \tag{36c}$$

where

$$\begin{aligned} a_h(\boldsymbol{u}, \boldsymbol{w}) = & \sum_{K \in \mathcal{T}_h} \int_K \boldsymbol{\epsilon}(\boldsymbol{u}) : \boldsymbol{\epsilon}(\boldsymbol{w}) dx - \sum_{e \in \mathcal{E}_h} \int_e \{\boldsymbol{\epsilon}(\boldsymbol{u})\} \cdot [\boldsymbol{w}_t] ds \\ & - \sum_{e \in \mathcal{E}_h} \int_e \{\boldsymbol{\epsilon}(\boldsymbol{w})\} \cdot [\boldsymbol{u}_t] ds + \sum_{e \in \mathcal{E}_h} \int_e \eta h_e^{-1} [\boldsymbol{u}_t] \cdot [\boldsymbol{w}_t] ds, \end{aligned} \tag{37}$$

$\tilde{\alpha}_{ii} = -\alpha_{p_i} - \alpha_{ii}$, and η is a stabilization parameter independent of parameters λ, R_i^{-1}, α_{p_i}, α_{ij} for all $i, j \in \{1, \dots, n\}$, the network scale n, and the mesh size h.

Remark 4. The general rescaled boundary conditions

$$p_i = p_{i,D} \qquad \text{on } \Gamma_{p_i,D}, \ i = 1, \dots, n, \tag{38a}$$

$$\boldsymbol{v}_i \cdot \boldsymbol{n} = q_{i,N} \qquad \text{on } \Gamma_{p_i,N}, \ i = 1, \dots, n, \tag{38b}$$

$$\boldsymbol{u} = \boldsymbol{u}_D \qquad \text{on } \Gamma_{\boldsymbol{u},D}, \tag{38c}$$

$$\left(\boldsymbol{\sigma} - \sum_{i=1}^{n} p_i \boldsymbol{I}\right) \boldsymbol{n} = \boldsymbol{g}_N \qquad \text{on } \Gamma_{\boldsymbol{u},N} \tag{38d}$$

can be incorporated as explained in the work of Hong et al.[23]

Proposition 1. *Let $(\boldsymbol{u}_h; \boldsymbol{v}_h; \boldsymbol{p}_h) \in \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h$ be the solution of (36a)-(36c); then, the pointwise mass conservation equation is satisfied, that is,*

$$-\operatorname{div} \boldsymbol{u}_h - \operatorname{div} \boldsymbol{v}_{i,h} - (\alpha_{p_i} + \alpha_{ii}) p_{i,h} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \alpha_{ij} p_{j,h} = Q_{i,h} g_i, \ i = 1, \dots, n, \ \forall x \in K, \forall K \in \mathcal{T}_h, \tag{39}$$

where the L^2-projection on $P_{i,h}$ is denoted by $Q_{i,h}$. Hence, if $g_i = 0$, then $-\operatorname{div} \boldsymbol{u}_h - \operatorname{div} \boldsymbol{v}_{i,h} - (\alpha_{p_i} + \alpha_{ii}) p_{i,h} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \alpha_{ij} p_{j,h} = 0$.

For $\boldsymbol{u} \in \boldsymbol{U}_h$, we introduce the mesh-dependent norms

$$\|\boldsymbol{u}\|_h^2 = \sum_{K\in\mathcal{T}_h} \|\epsilon(\boldsymbol{u})\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\boldsymbol{u}_t]\|_{0,e}^2,$$
$$\|\boldsymbol{u}\|_{1,h}^2 = \sum_{K\in\mathcal{T}_h} \|\nabla\boldsymbol{u}\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\boldsymbol{u}_t]\|_{0,e}^2,$$

the "DG"-norm

$$\|\boldsymbol{u}\|_{DG}^2 = \sum_{K\in\mathcal{T}_h} \|\nabla\boldsymbol{u}\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\boldsymbol{u}_t]\|_{0,e}^2 + \sum_{K\in\mathcal{T}_h} h_K^2|\boldsymbol{u}|_{2,K}^2, \tag{40}$$

and finally, the mesh-dependent norm $\|\cdot\|_{\boldsymbol{U}_h}$

$$\|\boldsymbol{u}\|_{\boldsymbol{U}_h}^2 = \|\boldsymbol{u}\|_{DG}^2 + \lambda\|\operatorname{div}\boldsymbol{u}\|^2. \tag{41}$$

Regarding the well-posedness and approximation properties of the DG formulation, we refer the reader to the works of Hong et al.[50,51] Firstly, from the discrete version of Korn's inequality, the norms $\|\cdot\|_{DG}$, $\|\cdot\|_h$, and $\|\cdot\|_{1,h}$ are equivalent on $\boldsymbol{U}_h$, namely,

$$\|\boldsymbol{u}\|_{DG} \eqsim \|\boldsymbol{u}\|_h \eqsim \|\boldsymbol{u}\|_{1,h}, \text{ for all } \boldsymbol{u} \in \boldsymbol{U}_h.$$

Secondly, the bilinear form $a_h(\cdot,\cdot)$ from (37) is continuous and it is valid that

$$|a_h(\boldsymbol{u},\boldsymbol{w})| \lesssim \|\boldsymbol{u}\|_{DG}\|\boldsymbol{w}\|_{DG}, \quad \text{for all } \boldsymbol{u},\boldsymbol{w} \in H^2(\mathcal{T}_h)^d. \tag{42}$$

Thirdly, the LBB conditions

$$\begin{aligned}
&\inf_{(q_{1,h},\dots,q_{n,h})\in P_{1,h}\times\dots\times P_{n,h}} \sup_{\boldsymbol{u}_h\in\boldsymbol{U}_h} \frac{\left(\operatorname{div}\boldsymbol{u}_h, \sum_{i=1}^n q_{i,h}\right)}{\|\boldsymbol{u}_h\|_{1,h}\|\sum_{i=1}^n q_{i,h}\|} \geq \beta_{sd},\\
&\inf_{q_{i,h}\in P_{i,h}} \sup_{\boldsymbol{v}_{i,h}\in\boldsymbol{V}_{i,h}} \frac{(\operatorname{div}\boldsymbol{v}_{i,h}, q_{i,h})}{\|\boldsymbol{v}_{i,h}\|_{\operatorname{div}}\|q_{i,h}\|} \geq \beta_{dd}, \quad i = 1,\dots,n
\end{aligned} \tag{43}$$

are satisfied for our choice of the finite element spaces $\boldsymbol{U}_h$, $\boldsymbol{V}_h$, and $\boldsymbol{P}_h$; see, for example, the work of Schötzau et al.[49] Here, the positive constants β_{sd} and β_{dd} are independent of the parameters λ, R_i^{-1}, α_{p_i}, α_{ij} for all $i,j \in \{1,\dots,n\}$, the network scale n, and the mesh size h. Finally, using standard techniques, one can show that

$$a_h(\boldsymbol{u}_h,\boldsymbol{u}_h) \geq \alpha_a\|\boldsymbol{u}_h\|_h^2, \quad \text{for all} \quad \boldsymbol{u}_h \in \boldsymbol{U}_h, \tag{44}$$

where α_a is a positive constant independent of the parameters $\lambda, R_i^{-1}, \alpha_{p_i}, \alpha_{ij}, i,j = 1,\dots,n$, the network scale n, and the mesh size h.

Related to the discrete problem (36a)-(36c), and from the definition of the matrices Λ_1 and Λ_2, we define the bilinear form

$$\begin{aligned}
\mathcal{A}_h((\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h),(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h)) &= a_h(\boldsymbol{u}_h,\boldsymbol{w}_h) + \lambda(\operatorname{div}\boldsymbol{u}_h,\operatorname{div}\boldsymbol{w}_h) - \sum_{i=1}^n (p_{i,h},\operatorname{div}\boldsymbol{w}_h)\\
&+ \sum_{i=1}^n \left(R_i^{-1}\boldsymbol{v}_{i,h},\boldsymbol{z}_{i,h}\right) - (\boldsymbol{p}_h,\operatorname{Div}\boldsymbol{z}_h) - \left(\operatorname{div}\boldsymbol{u}_h, \sum_{i=1}^n q_{i,h}\right) - (\operatorname{Div}\boldsymbol{v}_h,\boldsymbol{q}_h) - ((\Lambda_1+\Lambda_2)\boldsymbol{p}_h,\boldsymbol{q}_h).
\end{aligned} \tag{45}$$

The following theorem results directly from the definitions of the norms $\|\cdot\|_{\boldsymbol{U}_h}$, $\|\cdot\|_{\boldsymbol{V}}$ and $\|\cdot\|_{\boldsymbol{P}}$.

Theorem 3. *There exists a constant C_{bd} independent of the parameters λ, R_i^{-1}, α_{p_i}, α_{ij} for all $i,j \in \{1,\dots,n\}$, the network scale n, and the mesh size h such that the inequality*

$$|\mathcal{A}_h((\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h),(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h))| \leq C_{bd}(\|\boldsymbol{u}_h\|_{\boldsymbol{U}_h} + \|\boldsymbol{v}_h\|_{\boldsymbol{V}} + \|\boldsymbol{p}_h\|_{\boldsymbol{P}})(\|\boldsymbol{w}_h\|_{\boldsymbol{U}_h} + \|\boldsymbol{z}_h\|_{\boldsymbol{V}} + \|\boldsymbol{q}_h\|_{\boldsymbol{P}})$$

is fulfilled for any $(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h) \in \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h, (\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h) \in \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h$.

The second main result of this paper is given in the following theorem.

Theorem 4. *There exits a positive constant β_0 independent of the parameters λ, R_i^{-1}, α_{p_i}, α_{ij} for all $i,j \in \{1, \dots, n\}$, the network scale n, and the mesh size h such that*

$$\inf_{(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h)\in \boldsymbol{U}_h\times\boldsymbol{V}_h\times\boldsymbol{P}_h} \sup_{(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h)\in \boldsymbol{U}_h\times\boldsymbol{V}_h\times\boldsymbol{P}_h} \frac{\mathcal{A}_h((\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h),(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h))}{(\|\boldsymbol{u}_h\|_{\boldsymbol{U}_h}+\|\boldsymbol{v}_h\|_{\boldsymbol{V}}+\|\boldsymbol{p}_h\|_{\boldsymbol{P}})(\|\boldsymbol{w}_h\|_{\boldsymbol{U}_h}+\|\boldsymbol{z}_h\|_{\boldsymbol{V}}+\|\boldsymbol{q}_h\|_{\boldsymbol{P}})} \geq \beta_0. \tag{46}$$

Proof. Noting that $a_h(\boldsymbol{u}_h,\boldsymbol{u}_h)$ is coercive and the inf-sup conditions (43) hold, the proof of this theorem uses similar arguments and follows the lines of the proof of Theorem 2. □

The following stability estimate is a consequence of the above theorem.

Corollary 2. *Let $(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h) \in \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h$ be the solution of (36a)-(36c); then, the estimate*

$$\|\boldsymbol{u}_h\|_{\boldsymbol{U}_h}+\|\boldsymbol{v}_h\|_{\boldsymbol{V}}+\|\boldsymbol{p}_h\|_{\boldsymbol{P}} \leq C_2\left(\|\boldsymbol{f}\|_{\boldsymbol{U}_h^*}+\|\boldsymbol{g}\|_{\boldsymbol{P}^*}\right) \tag{47}$$

holds where

$$\|\boldsymbol{f}\|_{\boldsymbol{U}_h^*}=\sup_{\boldsymbol{w}_h\in\boldsymbol{U}_h}\frac{(\boldsymbol{f},\boldsymbol{w}_h)}{\|\boldsymbol{w}_h\|_{\boldsymbol{U}_h}}, \quad \|\boldsymbol{g}\|_{\boldsymbol{P}^*}=\sup_{\boldsymbol{q}_h\in\boldsymbol{P}_h}\frac{(\boldsymbol{g},\boldsymbol{q}_h)}{\|\boldsymbol{q}_h\|_{\boldsymbol{P}}},$$

and C_2 is a constant independent of λ, R_i^{-1}, α_{p_i}, α_{ij} for all $i,j \in \{1, \dots, n\}$, the network scale n, and the mesh size h.

Remark 5. Let $\boldsymbol{W}_h := \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h$ be equipped with the norm $\|\cdot\|_{\boldsymbol{W}_h} := \|\cdot\|_{\boldsymbol{U}_h}+\|\cdot\|_{\boldsymbol{V}}+\|\cdot\|_{\boldsymbol{P}}$ and consider the operator

$$\mathscr{A}_h := \begin{bmatrix}
-\mathrm{div}_h\boldsymbol{\epsilon}_h-\lambda\nabla_h\mathrm{div}_h & 0 & \dots & \dots & 0 & \nabla_h & \dots & \dots & \nabla_h \\
0 & R_1^{-1}I_h & 0 & \dots & 0 & \nabla_h & 0 & \dots & 0 \\
\vdots & 0 & \ddots & & \vdots & 0 & \ddots & & \vdots \\
\vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & 0 \\
0 & 0 & \dots & 0 & R_n^{-1}I_h & 0 & \dots & 0 & \nabla_h \\
-\mathrm{div}_h & -\mathrm{div}_h & 0 & \dots & 0 & \tilde{\alpha}_{11}I_h & \alpha_{12}I_h & \dots & \alpha_{1n}I_h \\
\vdots & 0 & \ddots & & \vdots & \alpha_{21}I_h & \ddots & & \alpha_{2n}I_h \\
\vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & \vdots \\
-\mathrm{div}_h & 0 & \dots & 0 & -\mathrm{div}_h & \alpha_{n1}I_h & \alpha_{n2}I_h & \dots & \tilde{\alpha}_{nn}I_h
\end{bmatrix}, \tag{48}$$

induced by the bilinear form (45). Clearly, $\mathscr{A}_h$ is self-adjoint and indefinite on $\boldsymbol{W}_h$. Moreover, Theorems 3 and 4 imply that it is a uniform isomorphism in the sense of being bounded and having a bounded inverse with bounds independent of the mesh size, the network scale, and the model parameters. Following the framework in the study of Mardal et al.,[41] we define the self-adjoint positive definite operator

$$\mathscr{B}_h := \begin{bmatrix} \mathscr{B}_{h,\boldsymbol{u}}^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mathscr{B}_{h,\boldsymbol{v}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \mathscr{B}_{h,\boldsymbol{p}}^{-1} \end{bmatrix}, \tag{49}$$

where

$\mathscr{B}_{h,\boldsymbol{u}} = -\mathrm{div}_h\boldsymbol{\epsilon}_h-\lambda\nabla_h\mathrm{div}_h$,

$$\mathscr{B}_{h,\boldsymbol{v}} = \begin{bmatrix} R_1^{-1}I_h & 0 & \dots & 0 \\ 0 & R_2^{-1}I_h & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & R_n^{-1}I_h \end{bmatrix} - \begin{bmatrix} \tilde{\gamma}_{11}\nabla_h\mathrm{div}_h & \tilde{\gamma}_{12}\nabla_h\mathrm{div}_h & \dots & \tilde{\gamma}_{1n}\nabla_h\mathrm{div}_h \\ \tilde{\gamma}_{21}\nabla_h\mathrm{div}_h & \tilde{\gamma}_{22}\nabla_h\mathrm{div}_h & \dots & \tilde{\gamma}_{2n}\nabla_h\mathrm{div}_h \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\gamma}_{n1}\nabla_h\mathrm{div}_h & \tilde{\gamma}_{n2}\nabla_h\mathrm{div}_h & \dots & \tilde{\gamma}_{nn}\nabla_h\mathrm{div}_h \end{bmatrix}, \quad \text{and} \quad \mathscr{B}_{h,\boldsymbol{p}} = \begin{bmatrix} \gamma_{11}I_h & \gamma_{12}I_h & \dots & \gamma_{1n}I_h \\ \gamma_{21}I_h & \gamma_{22}I_h & \dots & \gamma_{2n}I_h \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1}I_h & \gamma_{n2}I_h & \dots & \gamma_{nn}I_h \end{bmatrix}.$$

It is obvious that

$$\langle\mathscr{B}_h^{-1}\boldsymbol{x}_h,\boldsymbol{x}_h\rangle \eqsim \|\boldsymbol{x}_h\|_{\boldsymbol{W}_h}^2,$$

where $\boldsymbol{x}_h = (\boldsymbol{u}_h,\boldsymbol{v}_h,\boldsymbol{p}_h) \in \boldsymbol{W}_h$; "$\eqsim$" stands for a norm equivalence, uniform with respect to model and discretization parameters; and $\langle\cdot,\cdot\rangle$ expresses the duality pairing between $\boldsymbol{W}_h$ and $\boldsymbol{W}_h^*$, that is, $\mathscr{B}_h^{-1}$ is a uniform isomorphism.

By using the properties of $\mathscr{B}_h$ and $\mathscr{A}_h$ when solving the generalized eigenvalue problem $\mathscr{A}_h\boldsymbol{x}_h = \xi\mathscr{B}_h^{-1}\boldsymbol{x}_h$, the condition number $\kappa(\mathscr{B}_h\mathscr{A}_h)$ is easily shown to be uniformly bounded with respect to the parameters λ, R_i^{-1}, α_{p_i}, α_{ij} for all $i,j \in \{1, \dots, n\}$ in the ranges specified in (7), the network scale n, and the mesh size h. Therefore, $\mathscr{B}_h$ defines a uniform preconditioner.

Remark 6. To apply the preconditioner $\mathscr{B}_h$, one has to solve an elasticity system discretized by an $H(\mathrm{div})$-conforming DG method[51] and n-coupled elliptic $H(\mathrm{div})$ problems discretized by RT elements, which can be decoupled by diagonalization as follows. Denoting $D_R^{-1} = \mathrm{diag}(R_1^{-1}, R_2^{-1}, \ldots, R_n^{-1})$ and $\mathscr{D}_R^{-1} = \mathrm{blockdiag}(R_1^{-1} I_h, \ldots, R_n^{-1} I_h) = D_R^{-1} \otimes I_h$, we have

$$(\mathscr{B}_{h,\boldsymbol{v}} \boldsymbol{v}_h, \boldsymbol{z}_h) = \left(\mathscr{D}_R^{-\frac{1}{2}} \boldsymbol{v}_h, \mathscr{D}_R^{-\frac{1}{2}} \boldsymbol{z}_h \right) + \left(\Lambda^{-1} \mathrm{Div}\, \boldsymbol{v}_h, \mathrm{Div}\, \boldsymbol{z}_h \right). \tag{50}$$

Now, by the change of variables $\hat{\boldsymbol{v}}_h = \mathscr{D}_R^{-\frac{1}{2}} \boldsymbol{v}_h, \hat{\boldsymbol{z}}_h = \mathscr{D}_R^{-\frac{1}{2}} \boldsymbol{z}_h$, we get

$$\left(\mathscr{D}_R^{\frac{1}{2}} \mathscr{B}_{h,\boldsymbol{v}} \mathscr{D}_R^{\frac{1}{2}} \hat{\boldsymbol{v}}_h, \hat{\boldsymbol{z}}_h \right) = (\hat{\boldsymbol{v}}_h, \hat{\boldsymbol{z}}_h) + \left(\Lambda^{-1} \mathrm{Div} \left(\mathscr{D}_R^{\frac{1}{2}} \hat{\boldsymbol{v}}_h \right), \mathrm{Div} \left(\mathscr{D}_R^{\frac{1}{2}} \hat{\boldsymbol{z}}_h \right) \right) = (\hat{\boldsymbol{v}}_h, \hat{\boldsymbol{z}}_h) + \left(D_R^{\frac{1}{2}} \Lambda^{-1} D_R^{\frac{1}{2}} \mathrm{Div}\, \hat{\boldsymbol{v}}_h, \mathrm{Div}\, \hat{\boldsymbol{z}}_h \right). \tag{51}$$

Next, denoting $\Lambda_R^{-1} = D_R^{\frac{1}{2}} \Lambda^{-1} D_R^{\frac{1}{2}}$, we can diagonalize Λ_R^{-1} as $D_{\boldsymbol{v}}^{-1} = Q_{\boldsymbol{v}} \Lambda_R^{-1} Q_{\boldsymbol{v}}^T$, where $Q_{\boldsymbol{v}}$ satisfying $Q_{\boldsymbol{v}} Q_{\boldsymbol{v}}^T = I_{n\times n}$ is an orthogonal matrix and $D_{\boldsymbol{v}}^{-1} = \mathrm{diag}(\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_n)$ is the diagonal matrix composed from the eigenvalues of Λ_R^{-1}. Hence, by the further substitution $\bar{\boldsymbol{v}}_h = \mathscr{Q}_{\boldsymbol{v}} \hat{\boldsymbol{v}}_h, \bar{\boldsymbol{z}}_h = \mathscr{Q}_{\boldsymbol{v}} \hat{\boldsymbol{z}}_h$, where $\mathscr{Q}_{\boldsymbol{v}} = Q_{\boldsymbol{v}} \otimes I_h$, we obtain

$$\left(\mathscr{Q}_{\boldsymbol{v}} \mathscr{D}_R^{\frac{1}{2}} \mathscr{B}_{h,\boldsymbol{v}} \mathscr{D}_R^{\frac{1}{2}} \mathscr{Q}_{\boldsymbol{v}}^T \bar{\boldsymbol{v}}_h, \bar{\boldsymbol{z}}_h \right) = \left(\mathscr{Q}_{\boldsymbol{v}}^T \bar{\boldsymbol{v}}_h, \mathscr{Q}_{\boldsymbol{v}}^T \bar{\boldsymbol{z}}_h \right) + \left(D_{\boldsymbol{v}}^{-1} \mathrm{Div}\, \bar{\boldsymbol{v}}_h, \mathrm{Div}\, \bar{\boldsymbol{z}}_h \right) = (\bar{\boldsymbol{v}}_h, \bar{\boldsymbol{z}}_h) + \left(D_{\boldsymbol{v}}^{-1} \mathrm{Div}\, \bar{\boldsymbol{v}}_h, \mathrm{Div}\, \bar{\boldsymbol{z}}_h \right). \tag{52}$$

We denote

$$\mathscr{B}_{h,\bar{\boldsymbol{v}}} := \begin{bmatrix} I_h & 0 & \ldots & 0 \\ 0 & I_h & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & I_h \end{bmatrix} - \begin{bmatrix} \bar{\mu}_1 \nabla_h \mathrm{div}_h & 0 & \ldots & 0 \\ 0 & \bar{\mu}_2 \nabla_h \mathrm{div}_h & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \bar{\mu}_n \nabla_h \mathrm{div}_h \end{bmatrix},$$

which means that we only need to solve n-decoupled elliptic $H(div)$ problems discretized by RT elements to get $\bar{\boldsymbol{v}}_h$. This task has been addressed in the work of Kraus et al.,[52] where optimal solvers for the lowest order case have been discussed. Other order-optimal multigrid methods and efficient preconditioners for this type of $H(div)$ problems can be found in other works.[53–55] Finally, we obtain the original $\boldsymbol{v}_h$ from back substitution, that is, $\boldsymbol{v}_h = \mathscr{D}_R^{\frac{1}{2}} \mathscr{Q}_{\boldsymbol{v}}^T \bar{\boldsymbol{v}}_h$. Similarly, diagonalization can also be applied to $\mathscr{B}_{h,\boldsymbol{p}}$ to obtain the diagonal preconditioner

$$\mathscr{B}_{h,\bar{\boldsymbol{p}}} := \begin{bmatrix} \mu_1 I_h & 0 & \ldots & 0 \\ 0 & \mu_2 I_h & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mu_n I_h \end{bmatrix}$$

for the system with $\bar{\boldsymbol{p}}_h = Q_{\boldsymbol{p}} \boldsymbol{p}_h$, where $D_{\boldsymbol{p}} = Q_{\boldsymbol{p}} \Lambda Q_{\boldsymbol{p}}^T$, $D_{\boldsymbol{p}} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_n)$, and μ_i, $i = 1, \ldots, n$, denote the eigenvalues of $\Lambda = (\gamma_{ij})_{n\times n}$.

4.3 | Error estimates

This subsection summarizes the error estimates that follow from the stability results presented in Section 4.2. For further details (in the case $n = 1$), we refer the reader to the work of Hong et al.[23]

Theorem 5. *For the solution $(\boldsymbol{u}; \boldsymbol{v}; \boldsymbol{p})$ of (11) and $(\boldsymbol{u}_h; \boldsymbol{v}_h; \boldsymbol{p}_h)$ of (36a)– (36c), the error estimates*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{U_h} + \|\boldsymbol{v} - \boldsymbol{v}_h\|_V \le C_{e,u} \inf_{\boldsymbol{w}_h \in \boldsymbol{U}_h, \boldsymbol{z}_h \in \boldsymbol{V}_h} \left(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{U_h} + \|\boldsymbol{v} - \boldsymbol{z}_h\|_V \right) \tag{53}$$

and

$$\|\boldsymbol{p} - \boldsymbol{p}_h\|_P \le C_{e,p} \inf_{\boldsymbol{w}_h \in \boldsymbol{U}_h, \boldsymbol{z}_h \in \boldsymbol{V}_h, \boldsymbol{q}_h \in \boldsymbol{P}_h} \left(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{U_h} + \|\boldsymbol{v} - \boldsymbol{z}_h\|_V + \|\boldsymbol{p} - \boldsymbol{q}_h\|_P \right), \tag{54}$$

hold true, where the constants $C_{e,u}, C_{e,p}$ are independent of $\lambda, R_i^{-1}, \alpha_{p_i}, \alpha_{ij}, i, j = 1, \ldots, n$, the network scale n, and the mesh size h.

Proof. The proof of this result is analogous to the proof of Theorem 5.2 in the work of Hong et al.[23] □

Remark 7. In particular, the above theorem shows that the proposed discretizations are locking free. Note that the estimate (53) controls the error in $\boldsymbol{u}$ plus the error in $\boldsymbol{v}$ by the sum of the errors of the corresponding best approximations, whereas the estimate (54) requires the best approximation errors of all three vector variables $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{p}$ to control the error in $\boldsymbol{p}$.

TABLE 1 Errors measured in parameter-dependent norms ($\alpha_{p_1} = 10^{-4}$, $\lambda = 10^4$)

		h					
R_1^{-1}		1/8	1/16	1/32	1/64	1/128	1/256
1E0	$\|\cdot\|_{\boldsymbol{P}}$	2.1E−1	1.0E−1	5.2E−2	2.6E−2	1.3E−2	6.6E−3
	$\|\cdot\|_{\boldsymbol{V}}$	1.3E1	6.6E0	3.3E0	1.7E0	8.2E−1	4.1E−1
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3
1E2	$\|\cdot\|_{\boldsymbol{P}}$	2.1E−2	1.0E−2	5.1E−3	2.6E−3	1.3E−3	6.6E−4
	$\|\cdot\|_{\boldsymbol{V}}$	1.3E0	6.6E−1	3.3E−1	1.7E−1	8.2E−2	4.1E−2
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3
1E4	$\|\cdot\|_{\boldsymbol{P}}$	2.1E−3	1.0E−3	5.1E−4	2.6E−4	1.3E−4	6.6E−5
	$\|\cdot\|_{\boldsymbol{V}}$	1.3E−1	6.6E−2	3.3E−2	1.7E−2	8.2E−3	4.1E−3
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3
1E8	$\|\cdot\|_{\boldsymbol{P}}$	2.0E−3	1.0E−3	5.1E−4	2.6E−4	1.3E−4	6.6E−5
	$\|\cdot\|_{\boldsymbol{V}}$	1.6E−4	8.3E−5	4.4E−5	2.3E−5	1.2E−5	6.1E−6
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3
1E16	$\|\cdot\|_{\boldsymbol{P}}$	2.0E−3	1.0E−3	5.2E−4	2.6E−4	1.3E−4	6.6E−5
	$\|\cdot\|_{\boldsymbol{V}}$	1.6E−8	8.3E−9	4.4E−9	2.3E−9	1.2E−9	6.1E−10
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3

5 | NUMERICAL EXPERIMENTS

The following numerical experiments are for three widely applied MPET models, namely, the one-network, two-network, and four-network models. We suppose that the domain Ω is the unit square in $\mathbb{R}^2$, and during the discretization, it has been partitioned as bisections of $2N^2$ triangles with mesh size $h = 1/N$. To discretize the pressure variables, we use discontinuous piecewise constant elements; the fluxes are discretized employing the lowest order Raviart–Thomas space and the displacement we approximate with the Brezzi-Douglas-Marini elements of lowest order. All the numerical tests included in this section have been carried out in FEniCS.[56,57] The aim of these experiments is

(i) to validate the convergence of the error estimates in the derived parameter-dependent norms and
(ii) to test the robustness of the proposed block-diagonal preconditioners by using it within the MinRes algorithm where the iterative process has been initialized with a random vector.

In these numerical experiments, we apply exactly the block-diagonal preconditioners; inexact solvers, corresponding to approximate preconditioners, are to be investigated in future work.

5.1 | The one-network model

Here, we consider the simplest case of a system with only one pressure and one flux, namely, Biot's consolidation model. We solve system (8) for

$$\boldsymbol{f} = \begin{bmatrix} -(2y^3 - 3y^2 + y)(12x^2 - 12x + 2) - (x-1)^2x^2(12y - 6) + 900(y-1)^2y^2(4x^3 - 6x^2 + 2x) \\ (2x^3 - 3x^2 + x)(12y^2 - 12y + 2) + (y-1)^2y^2(12x - 6) + 900(x-1)^2x^2(4y^3 - 6y^2 + 2y) \end{bmatrix}$$

and

$$g = R_1\left(\frac{\partial \phi_2}{\partial x} + \frac{\partial \phi_2}{\partial y}\right) - \alpha_{p_1}(\phi_2 - 1), \text{ where } \phi_1 = (x-1)^2(y-1)^2x^2y^2, \ \phi_2 = 900(x-1)^2(y-1)^2x^2y^2, \ (x,y) \in \Omega.$$

Then, the exact solution of system (8) with boundary conditions $\boldsymbol{u}|_{\partial\Omega} = 0, \boldsymbol{v}\cdot\boldsymbol{n}|_{\partial\Omega} = 0$ is given by

$$\boldsymbol{u} = \left(\frac{\partial \phi_1}{\partial y}, -\frac{\partial \phi_1}{\partial x}\right), \ p = \phi_2 - 1, \ \boldsymbol{v} = -R_1\nabla p, \text{ where } p \in L_0^2(\Omega).$$

We performed experiments with different sets of input parameters. In Tables 1 –3, we report the error of the numerical solution in the introduced parameter-dependent norms $\|\cdot\|_{\boldsymbol{P}}$, $\|\cdot\|_{\boldsymbol{V}}$, $\|\cdot\|_{\boldsymbol{U}_h}$. Additionally, we list the number of MinRes iterations n_{it} and average residual convergence factor with the proposed block-diagonal preconditioner where the stopping criterion is residual reduction by 10^8 in the norm induced by the preconditioner. The robustness of the method is validated with respect to variation of the parameters λ, R_1^{-1}, α_{p_1}, as introduced in (8), and the discretization parameter h.

TABLE 2 Errors measured in parameter-dependent norms ($\alpha_{p_1} = 0, R_1^{-1} = 10^8$)

		h					
λ		1/8	1/16	1/32	1/64	1/128	1/256
1E0	$\|\cdot\|_{\boldsymbol{P}}$	2.0E−1	1.0E−1	5.2E−2	2.6E−2	1.3E−2	6.6E−3
	$\|\cdot\|_{\boldsymbol{V}}$	1.6E−4	8.9E−5	5.7E−5	4.6E−5	4.3E−5	4.1E−5
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3
1E4	$\|\cdot\|_{\boldsymbol{P}}$	2.0E−3	1.0E−3	5.2E−4	2.6E−4	1.3E−4	6.6E−5
	$\|\cdot\|_{\boldsymbol{V}}$	1.6E−4	8.6E−5	4.5E−5	2.3E−5	1.2E−5	6.1E−6
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3
1E8	$\|\cdot\|_{\boldsymbol{P}}$	2.1E−5	1.0E−5	5.2E−6	2.6E−6	1.3E−6	6.6E−7
	$\|\cdot\|_{\boldsymbol{V}}$	1.3E−3	6.5E−4	3.3E−4	1.6E−4	8.2E−5	4.1E−5
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.3E−2	1.1E−2	5.6E−3	2.8E−3

TABLE 3 Errors measured in parameter-dependent norms ($R_1^{-1} = 10^4, \lambda = 10^0$)

		h					
α_{p_1}		1/8	1/16	1/32	1/64	1/128	1/256
1E0	$\|\cdot\|_{\boldsymbol{P}}$	2.0E−1	1.0E−1	5.2E−2	2.6E−2	1.3E−2	6.6E−3
	$\|\cdot\|_{\boldsymbol{V}}$	1.6E−2	8.1E−3	4.1E−3	2.0E−3	1.0E−3	5.1E−4
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.0E−2	4.5E−2	2.2E−2	1.1E−2	5.6E−3	2.8E−3
1E−4	$\|\cdot\|_{\boldsymbol{P}}$	2.0E−1	1.0E−1	5.2E−2	2.6E−2	1.3E−2	6.6E−3
	$\|\cdot\|_{\boldsymbol{V}}$	1.6E−2	8.3E−3	4.2E−3	2.1E−3	1.0E−3	5.1E−4
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.2E−2	1.1E−2	5.6E−3	2.8E−3
0	$\|\cdot\|_{\boldsymbol{P}}$	2.0E−1	1.0E−1	5.2E−2	2.6E−2	1.3E−2	6.6E−3
	$\|\cdot\|_{\boldsymbol{V}}$	1.6E−2	8.3E−3	4.2E−3	2.1E−3	1.0E−3	5.1E−4
	$\|\cdot\|_{\boldsymbol{U}_h}$	9.1E−2	4.5E−2	2.2E−2	1.1E−2	5.6E−3	2.8E−3

As can be seen from Tables 1 –3 the error in the considered parameter-dependent norms decreases by a factor of 2 when decreasing the mesh size by the same factor independently of the model parameters. Although the error of the velocity in Table 2 is not reduced by this factor when $\lambda = 1E0$, the previous statement remains valid and in accordance with the theoretical results. Remember that, according to Theorem 5, estimate (53) bounds the sum of the errors of the approximations of $\boldsymbol{u}$ and $\boldsymbol{v}$ and, hence, reflects the convergence of the larger of the two.

The results in Table 4 suggest that the number of MinRes iterations required to achieve a prescribed solution accuracy is bounded by a constant independent of λ, R_1^{-1}, α_{p_1}, and h, whereas the average residual reduction factor always remains smaller than 0.70. Note that, in this table, the authors have tried to present the most unfavourable setting of input parameters in order to stress test the proposed method.

5.2 | The two-network model

The governing partial differential equations of the Biot–Barenblatt model are a special case of the flux-based MPET system (1) and involve two pressures and two fluxes ($n = 2$). We consider here the cantilever bracket benchmark problem proposed by the National Agency for Finite Element Methods and Standards[58] with $\boldsymbol{f} = 0, g_1 = 0$, and $g_2 = 0$.

The boundary Γ of the domain $\Omega = [0,1]^2$ is split into Γ_1, Γ_2, Γ_3, and Γ_4 denoting the bottom, right, top, and left boundaries, respectively, and the boundary conditions $\boldsymbol{u} = 0$ on Γ_4, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I})\boldsymbol{n} = (0,0)^T$ on $\Gamma_1 \cup \Gamma_2$, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I})\boldsymbol{n} = (0, -1)^T$ on Γ_3, $p_1 = 2$ on Γ, and $p_2 = 20$ on Γ are imposed.

The base values of the model parameters, used for the numerical testing of the preconditioned MinRes algorithm in Table 6, are taken from the work of Kolesov et al.[59] and are presented in Table 5.

The numerical results in Table 6 show robust behaviour with respect to mesh refinements and variation of the parameters including high contrasts of the hydraulic conductivities. Moreover, in Table 7, we have confirmed the robustness of the proposed block-diagonal preconditioners for larger values of the transfer coefficient β, while varying the hydraulic conductivities as considerably higher values than that in the work of Kolesov et al.[59] have been reported in the work of Lee et al.[60] when modelling cardiac perfusion. With the choice of parameter ranges for K_1 and K_2, we encompassed interesting test scenarios revealing changes in the convergence properties.

TABLE 4 Number of preconditioned MinRes iterations and average residual reduction factor for residual reduction by 10^8 in the norm induced by the preconditioner when solving the Biot problem

			R_1^{-1}											
h	α_p	λ	1E0		1E2		1E3		1E4		1E8		1E16	
$\frac{1}{16}$	1E0	1E0	22	0.42	32	0.52	33	0.53	23	0.41	9	0.10	9	0.10
		1E4	10	0.14	18	0.32	19	0.38	14	0.23	4	< 0.01	3	< 0.01
		1E8	7	0.05	12	0.18	13	0.21	10	0.14	3	< 0.01	3	< 0.01
	1E−4	1E0	20	0.40	36	0.57	43	0.65	33	0.54	14	0.23	15	0.25
		1E4	12	0.20	9	0.11	15	0.26	14	0.28	25	0.44	7	0.05
		1E8	5	<0.01	6	0.03	7	0.05	9	0.09	19	0.34	5	<0.01
	1E−8	1E0	20	0.40	35	0.54	48	0.67	37	0.58	16	0.27	13	0.23
		1E4	8	0.08	10	0.12	12	0.19	12	0.19	26	0.47	7	0.05
		1E8	5	<0.01	5	<0.01	6	0.03	8	0.07	14	0.24	5	<0.01
	0	1E0	19	0.36	35	0.58	49	0.67	37	0.58	16	0.27	13	0.24
		1E4	8	0.08	10	0.12	12	0.17	12	0.17	26	0.47	7	0.05
		1E8	5	<0.01	5	<0.01	6	0.03	8	0.07	14	0.24	5	<0.01
$\frac{1}{64}$	1E0	1E0	20	0.40	33	0.54	37	0.58	30	0.52	11	0.14	11	0.14
		1E4	10	0.14	18	0.32	18	0.33	19	0.38	5	0.01	4	<0.01
		1E8	7	0.05	12	0.19	15	0.26	15	0.26	5	0.01	4	<0.01
	1E−4	1E0	20	0.40	35	0.56	49	0.68	46	0.66	17	0.32	17	0.32
		1E4	10	0.14	7	0.05	15	0.25	15	0.26	33	0.54	6	0.03
		1E8	4	<0.01	6	0.03	7	0.05	9	0.09	18	0.33	5	<0.01
	1E−8	1E0	20	0.40	37	0.58	49	0.68	46	0.65	19	0.38	11	0.19
		1E4	6	0.03	11	0.17	12	0.19	12	0.19	27	0.50	6	0.03
		1E8	4	<0.01	7	0.05	9	0.10	9	0.10	14	0.24	5	<0.01
	0	1E0	20	0.40	37	0.58	48	0.67	46	0.64	20	0.40	11	0.19
		1E4	6	0.03	11	0.17	12	0.19	12	0.19	27	0.50	6	0.03
		1E8	4	<0.01	7	0.05	8	0.07	9	0.10	15	0.28	5	<0.01
$\frac{1}{256}$	1E0	1E0	20	0.40	29	0.51	34	0.55	32	0.54	11	0.15	11	0.15
		1E4	10	0.14	18	0.32	18	0.33	20	0.40	5	0.01	4	<0.01
		1E8	7	0.05	12	0.19	15	0.26	15	0.25	5	0.01	4	<0.01
	1E−4	1E0	20	0.40	33	0.54	49	0.68	50	0.68	18	0.34	17	0.32
		1E4	10	0.14	8	0.07	14	0.23	14	0.22	35	0.56	6	0.03
		1E8	3	<0.01	6	0.03	7	0.05	9	0.09	18	0.33	4	<0.01
	1E−8	1E0	21	0.39	37	0.58	49	0.67	50	0.68	20	0.39	11	0.15
		1E4	6	0.03	11	0.17	12	0.19	12	0.19	27	0.48	6	0.04
		1E8	4	<0.01	7	0.05	9	0.10	9	0.10	13	0.23	5	<0.01
	0	1E0	20	0.38	37	0.58	48	0.68	50	0.68	20	0.39	11	0.16
		1E4	6	0.02	11	0.16	12	0.19	12	0.19	27	0.50	6	0.03
		1E8	4	<0.01	6	0.03	9	0.09	9	0.10	15	0.27	5	<0.01

TABLE 5 Base values of model parameters for the Barenblatt model

Parameter	Value	Unit
λ	4.2	MPa
μ	2.4	MPa
c_{p_1}	54	$(\text{GPa})^{-1}$
c_{p_2}	14	$(\text{GPa})^{-1}$
α_1	0.95	
α_2	0.12	
β	5	$10^{-10}\text{kg}/(\text{m}\cdot\text{s})$
	100	$10^{-10}\text{kg}/(\text{m}\cdot\text{s})$
K_1	6.18	10^{-15}m^2
K_2	27.2	10^{-15}m^2

TABLE 6 Number of preconditioned MinRes iterations and average residual reduction factor for residual reduction by 10^8 in the norm induced by the preconditioner when solving the Barenblatt problem

h	β		K_2		$K_2 \cdot 10^2$		$K_2 \cdot 10^4$		$K_2 \cdot 10^6$	
$\frac{1}{16}$	5E−10	$K_1 \cdot 10^{-2}$	13	0.23	16	0.30	24	0.45	19	0.35
		$K_1 \cdot 10^{-1}$	14	0.25	18	0.33	26	0.47	21	0.41
		K_1	15	0.28	19	0.36	29	0.51	22	0.43
	1E−8	$K_1 \cdot 10^{-2}$	13	0.23	16	0.31	24	0.45	20	0.39
		$K_1 \cdot 10^{-1}$	14	0.24	18	0.33	26	0.47	21	0.41
		K_1	15	0.28	19	0.35	30	0.52	21	0.41
$\frac{1}{64}$	5E−10	$K_1 \cdot 10^{-2}$	16	0.29	27	0.49	26	0.47	20	0.40
		$K_1 \cdot 10^{-1}$	16	0.31	29	0.51	28	0.50	21	0.41
		K_1	17	0.32	30	0.52	31	0.53	22	0.43
	1E−8	$K_1 \cdot 10^{-2}$	16	0.31	26	0.47	26	0.47	20	0.40
		$K_1 \cdot 10^{-1}$	16	0.31	29	0.51	28	0.50	21	0.41
		K_1	18	0.33	30	0.52	31	0.53	24	0.45
$\frac{1}{256}$	5E−10	$K_1 \cdot 10^{-2}$	19	0.35	30	0.52	28	0.50	21	0.41
		$K_1 \cdot 10^{-1}$	21	0.41	33	0.55	29	0.51	21	0.41
		K_1	21	0.41	35	0.56	30	0.52	22	0.43
	1E−8	$K_1 \cdot 10^{-2}$	19	0.35	29	0.51	29	0.51	20	0.39
		$K_1 \cdot 10^{-1}$	20	0.40	32	0.54	29	0.51	22	0.43
		K_1	22	0.42	35	0.56	30	0.52	23	0.44

TABLE 7 Number of preconditioned MinRes iterations and average residual reduction factor for residual reduction by 10^8 in the norm induced by the preconditioner when solving the Barenblatt problem

			$\beta = 10^{-6}$		$\beta = 10^{-3}$		$\beta = 10^{0}$		$\beta = 10^{3}$		$\beta = 10^{6}$	
h=1/256.	$K_2 \cdot 10^{-3}$	$K_1 \cdot 10^{-3}$	14	0.26	13	0.23	13	0.23	13	0.22	13	0.22
		K_1	17	0.33	17	0.33	17	0.32	15	0.29	15	0.28
		$K_1 \cdot 10^3$	29	0.51	29	0.51	27	0.45	25	0.43	23	0.42
		$K_1 \cdot 10^6$	27	0.49	27	0.48	26	0.46	25	0.43	23	0.42
	K_2	$K_1 \cdot 10^{-3}$	16	0.31	16	0.31	15	0.28	15	0.29	14	0.26
		K_1	22	0.42	23	0.42	20	0.40	15	0.29	16	0.30
		$K_1 \cdot 10^3$	36	0.57	39	0.62	33	0.54	30	0.54	31	0.55
		$K_1 \cdot 10^6$	35	0.56	35	0.57	34	0.55	32	0.52	30	0.53
	$K_2 \cdot 10^3$	$K_1 \cdot 10^{-3}$	26	0.46	26	0.46	25	0.44	22	0.42	22	0.42
		K_1	35	0.56	35	0.56	32	0.52	30	0.54	30	0.53
		$K_1 \cdot 10^3$	44	0.64	46	0.66	45	0.65	38	0.61	37	0.60
		$K_1 \cdot 10^6$	41	0.62	41	0.62	41	0.62	33	0.53	33	0.53

5.3 | The four-network model

In this example, we consider the four-network MPET model. The boundary Γ of Ω is split as in the previous example, that is, $\Gamma = \bar{\Gamma}_1 \cup \bar{\Gamma}_2 \cup \bar{\Gamma}_3 \cup \bar{\Gamma}_4$ with $\Gamma_i \cap \Gamma_j = \emptyset$ for $i \neq j$ and $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$ denoting the bottom, right, top, and left boundaries, respectively. Then, the boundary conditions are as follows: $\boldsymbol{u} = 0$ on Γ_4, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I} - p_3\boldsymbol{I} - p_4\boldsymbol{I})\boldsymbol{n} = (0,0)^T$ on $\Gamma_1 \cup \Gamma_2$, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I} - p_3\boldsymbol{I} - p_4\boldsymbol{I})\boldsymbol{n} = (0,-1)^T$ on Γ_3, $p_1 = 2$ on Γ, $p_2 = 20$ on Γ, $p_3 = 30$ on Γ, and $p_4 = 40$ on Γ. The right-hand sides in (8) are chosen to be $\boldsymbol{f} = 0$, $g_1 = 0$, $g_2 = 0$, $g_3 = 0$, and $g_4 = 0$.

The base values of the parameters for numerical testing are given in Table 8 and taken from the work of Vardakis et al.,[7] where the four-network MPET model has been used to simulate fluid flow in the human brain.

Table 9 shows robust behaviour of the block-diagonal preconditioner (49) as the number of MinRes iterations remains uniformly bounded for large variations of the coefficients λ, K_3 and $K = K_1 = K_2 = K_4$. Note that, in all three examples, that is, for the one-network problem, the two-network-problem, and the four-network problem, the observed average residual reduction factors were always below 0.7 and did not increase as the number of networks was increased, which is in accordance with the theoretical findings. Moreover, the authors have tried to perform the numerical tests for the parameter ranges leading to the worst results.

TABLE 8 Base values of model parameters for the four-network MPET model

Parameter	Value	Unit
λ	505	Nm^{-2}
μ	216	Nm^{-2}
$c_{p_1} = c_{p_2} = c_{p_3} = c_{p_4}$	$4.5 \cdot 10^{-10}$	$\mathrm{m}^2\mathrm{N}^{-1}$
$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$	0.99	
$\beta_{12} = \beta_{24}$	$1.5 \cdot 10^{-19}$	$\mathrm{m}^2\mathrm{N}^{-1}\mathrm{s}^{-1}$
β_{23}	$2.0 \cdot 10^{-19}$	$\mathrm{m}^2\mathrm{N}^{-1}\mathrm{s}^{-1}$
β_{34}	$1.0 \cdot 10^{-13}$	$\mathrm{m}^2\mathrm{N}^{-1}\mathrm{s}^{-1}$
$K_1 = K_2 = K_4 = K$	$(1.0 \cdot 10^{-10})/(2.67 \cdot 10^{-3})$	$\mathrm{m}^2/\mathrm{Nsm}^{-2}$
K_3	$(1.4 \cdot 10^{-14})/(8.9 \cdot 10^{-4})$	$\mathrm{m}^2/\mathrm{Nsm}^{-2}$

Note. MPET = multiple-network poroelastic theory.

TABLE 9 Number of preconditioned MinRes iterations and average residual reduction factor for residual reduction by 10^8 in the norm induced by the preconditioner when solving the four-network MPET problem

h			$K_3 \cdot 10^{-2}$		K_3		$K_3 \cdot 10^2$		$K_3 \cdot 10^4$		$K_3 \cdot 10^6$		$K_3 \cdot 10^{10}$	
$\frac{1}{32}$	λ	$K \cdot 10^{-2}$	32	0.52	34	0.55	29	0.50	29	0.50	29	0.51	29	0.51
		K	17	0.32	20	0.39	23	0.42	19	0.38	18	0.33	22	0.42
		$K \cdot 10^2$	15	0.28	17	0.33	20	0.38	24	0.45	35	0.58	35	0.57
	$\lambda \cdot 10^4$	$K \cdot 10^{-2}$	24	0.44	33	0.53	38	0.61	38	0.60	38	0.60	38	0.61
		K	15	0.27	26	0.47	38	0.60	31	0.52	32	0.52	31	0.51
		$K \cdot 10^2$	13	0.22	24	0.44	39	0.61	22	0.42	16	0.31	16	0.30
	$\lambda \cdot 10^8$	$K \cdot 10^{-2}$	25	0.43	27	0.45	17	0.31	17	0.31	17	0.32	17	0.32
		K	26	0.48	25	0.44	13	0.23	10	0.15	11	0.16	10	0.15
		$K \cdot 10^2$	26	0.48	24	0.45	13	0.22	14	0.26	14	0.25	15	0.27
$\frac{1}{64}$	λ	$K \cdot 10^{-2}$	32	0.53	32	0.53	28	0.51	28	0.50	28	0.51	28	0.51
		K	20	0.39	21	0.40	21	0.40	20	0.39	21	0.40	23	0.42
		$K \cdot 10^2$	18	0.33	16	0.31	20	0.39	26	0.48	34	0.55	34	0.56
	$\lambda \cdot 10^4$	$K \cdot 10^{-2}$	24	0.45	37	0.59	39	0.62	39	0.62	39	0.61	39	0.61
		K	18	0.34	29	0.51	38	0.61	28	0.51	28	0.51	28	0.50
		$K \cdot 10^2$	14	0.25	32	0.54	38	0.60	21	0.40	14	0.25	14	0.25
	$\lambda \cdot 10^8$	$K \cdot 10^{-2}$	27	0.46	27	0.47	17	0.32	17	0.32	17	0.34	17	0.34
		K	26	0.49	25	0.44	12	0.20	9	0.12	10	0.14	9	0.11
		$K \cdot 10^2$	25	0.43	24	0.44	12	0.19	16	0.31	17	0.32	13	0.22
$\frac{1}{128}$	λ	$K \cdot 10^{-2}$	33	0.53	34	0.56	29	0.52	29	0.52	29	0.53	29	0.53
		K	20	0.40	21	0.41	22	0.42	21	0.41	22	0.41	22	0.41
		$K \cdot 10^2$	17	0.33	17	0.32	20	0.40	27	0.47	35	0.56	35	0.57
	$\lambda \cdot 10^4$	$K \cdot 10^{-2}$	23	0.42	37	0.58	40	0.62	40	0.63	40	0.63	40	0.62
		K	19	0.38	29	0.51	34	0.56	28	0.51	27	0.47	28	0.51
		$K \cdot 10^2$	15	0.27	33	0.54	34	0.56	21	0.41	15	0.28	14	0.26
	$\lambda \cdot 10^8$	$K \cdot 10^{-2}$	27	0.48	27	0.48	18	0.34	18	0.33	16	0.29	16	0.30
		K	25	0.44	25	0.44	13	0.23	9	0.12	11	0.16	11	0.17
		$K \cdot 10^2$	26	0.49	25	0.46	12	0.21	17	0.34	18	0.34	13	0.23

Note. MPET = multiple-network poroelastic theory.

6 | CONCLUSIONS

In this paper, as motivated by the approach recently presented by Hong et al.[23] for the Biot model, we establish the uniform stability and design stable discretizations and parameter-robust preconditioners for flux-based formulations of multiple-network poroelasticity systems. Novel proper parameter-matrix-dependent norms that provide the key for establishing uniform inf-sup stability of the continuous problems are introduced. The stability results that could be obtained using the presented *matrix technique* are uniform not only with respect to the Lamé parameter λ but also to all the other model parameters such as small or large permeability coefficients K_i, arbitrary small or even vanishing storage

coefficients c_{p_i}, arbitrary small or even vanishing network transfer coefficients $\beta_{ij}, i,j = 1, \ldots, n$, the scale of the networks n, and the time step size τ.

Moreover, strongly mass-conservative and uniformly stable discretizations are proposed and corresponding uniform and optimal error estimates proved, which are also independent of the Lamé parameter λ; the permeability coefficients K_i; the storage coefficients c_{p_i}; the network transfer coefficients $\beta_{ij}, i,j = 1, \ldots, n$; the scale of the networks n; the time step size τ; and the mesh size h. The transfer of the canonical (norm-equivalent) operator preconditioners from the continuous to the discrete level lays the foundation for optimal and fully robust iterative solution methods. Numerical experiments motivated by practical applications are presented. These confirm both the uniform and optimal convergence of the proposed finite element methods and the uniform robustness of the norm-equivalent preconditioners.

CONFLICT OF INTEREST

This work does not have any conflicts of interest.

ORCID

Qingguo Hong https://orcid.org/0000-0003-2420-9653
Johannes Kraus https://orcid.org/0000-0003-2261-599X

REFERENCES

1. Barenblatt GI, Zheltov IP, Kochina IN. Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata]. J Appl Math Mech. 1960;24(5):1286–1303.
2. Bai M, Elsworth D, Roegiers J-C. Multiporosity/multipermeability approach to the simulation of naturally fractured reservoirs. Water Resour Res. 1993;29(6):1621–1633.
3. Biot MA. General theory of three-dimensional consolidation. J Appl Phys. 1941;12(2):155–164.
4. Biot MA. Theory of elasticity and consolidation for a porous anisotropic solid. J Appl Phys. 1955;26(2):182–185.
5. Tully B, Ventikos Y. Cerebral water transport using multiple-network poroelastic theory: application to normal pressure hydrocephalus. J Fluid Mech. 2011;667:188–215.
6. Vardakis JC, Tully BJ, Ventikos Y. Exploring the efficacy of endoscopic ventriculostomy for hydrocephalus treatment via a multicompartmental poroelastic model of CSF transport: a computational perspective. PLoS ONE. 2013;8(12):e84577.
7. Vardakis JC, Chou D, Tully BJ, et al. Investigating cerebral oedema using poroelasticity. Med Eng Phys. 2016;38(1):48–57.
8. Chou D, Vardakis JC, Guo L, Tully BJ, Ventikos Y. A fully dynamic multi-compartmental poroelastic system: application to aqueductal stenosis. J Biomech. 2016;49:2306–2312.
9. Guo L, Vardakis JC, Lassila T, et al. Subject-specific multi-poroelastic model for exploring the risk factors associated with the early stages of Alzheimer's disease. Interface Focus. 2018;8(1):20170019.
10. Wang HF. Theory of linear poroelasticity with applications to geomechanics and hydrogeology. Princeton, NJ: Princeton University Press; 2000.
11. Lee JJ, Mardal K-A, Winther R. Parameter-robust discretization and preconditioning of Biot's consolidation model. SIAM J Sci Comput. 2017;39(1):A1–A24.
12. Coussy O. Poromechanics. West Sussex, UK: John Wiley & Sons; 2004.
13. Smith JH, Humphrey JAC. Interstitial transport and transvascular fluid exchange during infusion into brain and tumor tissue. Microvascular Research. 2007;73(1):58–73.
14. Støverud KH, Alnæs M, Langtangen HP, Haughton V, Mardal K-A. Poro-elastic modeling of Syringomyelia - a systematic study of the effects of pia mater, central canal, median fissure, white and gray matter on pressure wave propagation and fluid movement within the cervical spinal cord. Comput Methods Biomech Biomed Engin. 2016;19(6):686–698.
15. Axelsson O, Blaheta R, Byczanski P. Stable discretization of poroelasticity problems and efficient preconditioners for arising saddle point type matrices. Comput Vis Sci. 2012;15(4):191–207.
16. Gaspar FJ, Lisbona FJ, Vabishchevich PN. A finite difference analysis of Biot's consolidation model. Appl Numer Math. 2003;44(4):487–506.
17. Gaspar FJ, Lisbona FJ, Vabishchevich PN. Staggered grid discretizations for the quasi-static Biot's consolidation problem. Appl Numer Math. 2006;56(6):888–898.
18. Nordbotten JM. Stable cell-centered finite volume discretization for Biot equations. SIAM J Numer Anal. 2016;54(2):942–968.
19. Boffi D, Brezzi F, Fortin M. Mixed finite element methods and applications. Berlin, Germany: Springer-Verlag Berlin Heidelberg; 2013. Springer series in computational mathematics, No. 44.

20. Ern A, Guermond J-L. Theory and practice of finite elements. New York, NY: Springer-Verlag; 2004. Applied mathematical sciences, No. 159.
21. Babuška I. Error-bounds for finite element method. Numerische Mathematik. 1971;16:322–333.
22. Brezzi F. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. Rev Française Automat Informat Recherche Opérationnelle Sér Rouge. 1974;8(R-2):129–151.
23. Hong Q, Kraus J. Parameter-robust stability of classical three-field formulation of Biot's consolidation model. Electron Trans Numer Anal. 2018;48:202–226.
24. Boffi D, Botti M, Di Pietro DA. A nonconforming high-order method for the Biot problem on general meshes. SIAM J Sci Comput. 2016;38(3):A1508–A1537.
25. Adler JH, Gaspar FJ, Hu X, Rodrigo C, Zikatanov LT. Robust block preconditioners for Biot's model. arXiv:1705.08842v1 [mathNA]. 2017.
26. Oyarzúa R, Ruiz-Baier R. Locking-free finite element methods for poroelasticity. SIAM J Numer Anal. 2016;54(5):2951–2973.
27. Lee JJ. Robust error analysis of coupled mixed methods for Biot's consolidation model. J Sci Comput. 2016;69(2):610–632.
28. Bærland T, Lee JJ, Mardal K-A, Winther R. Weakly imposed symmetry and robust preconditioners for Biot's consolidation model. Comput Methods Appl Math. 2017;17(3):377–396.
29. Lee JJ, Piersanti E, Mardal KA, Rognes ME. A mixed finite element method for nearly incompressible multiple-network poroelasticity. arXiv preprint arXiv:1804.07568. 2018.
30. Hansbo P, Larson MG. Discontinuous Galerkin and the Crouzeix-Raviart element: application to elasticity. Esaim Math Model Numer Anal. 2003;37(01):63–72.
31. Hu X, Rodrigo C, Gaspar FJ, Zikatanov LT. A nonconforming finite element method for the Biot's consolidation model in poroelasticity. J Comput Appl Math. 2017;310:143–154.
32. Fortin M, Soulie M. A non-conforming piecewise quadratic finite element on triangles. Int J Numer Meth Eng. 1983;19(4):505–520.
33. Rodrigo C, Hu X, Ohm P, Adler JH, Gaspar FJ, Zikatanov LT. New stabilized discretizations for poroelasticity and the Stokes' equations. 2018;341:467-484.
34. Lee JJ. Robust three-field finite element methods for Biot's consolidation model in poroelasticity. BIT Numer Math. 2018;58(2):347–372.
35. Phillips PJ, Wheeler MF. A coupling of mixed and continuous Galerkin finite element methods for poroelasticity. I: the continuous in time case. Computational Geosciences. 2007;11(2):131–144.
36. Phillips PJ, Wheeler MF. A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity. Computational Geosciences. 2008;12(4):417–435.
37. Kanschat G, Riviere B. A finite element method with strong mass conservation for Biot's linear consolidation model. J Sci Comput. 2018;77(3):1762–1779.
38. Showalter RE. Poroelastic filtration coupled to stokes flow. In: Control theory of partial differential equations. Boca Raton, FL: Chapman & Hall/CRC, 2010; p. 229–241. Lecture notes in pure and applied mathematics, No. 242.
39. Barry SI, Mercer GN. Exact solutions for two-dimensional time-dependent flow and deformation within a poroelastic medium. J Appl Mech. 1999;66(2):536–540.
40. Lipnikov K. Numerical methods for the Biot model in poroelasticity. Houston, TX: University of Houston; 2002.
41. Mardal K-A, Winther R. Preconditioning discretizations of systems of partial differential equations. Numer Linear Algebra Appl. 2011;18(1):1–40.
42. Arnold DN. An interior penalty finite element method with discontinuous elements. SIAM J Numer Anal. 1982;19(4):742–760.
43. Brezzi F, Manzini G, Marini D, Pietra P, Russo A. Discontinuous Galerkin approximations for elliptic problems. Numer Methods Partial Differ Equ. 2000;16(4):365–378.
44. Arnold DN, Brezzi F, Cockburn B, Marini LD. Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J Numer Anal. 2002;39:1749–1779.
45. Cockburn B, Kanschat G, Schötzau D. A note on discontinuous Galerkin divergence-free solutions of the Navier–Stokes equations. J Sci Comput. 2007;31(1):61–73.
46. Hong Q, Hu J, Shu S, Xu J. A discontinuous Galerkin method for the fourth-order curl problem. J Comput Math. 2012;30(6):565–578.
47. Hong Q, Wang F, Wu S, Xu J. A unified study of continuous and discontinuous Galerkin methods. Sci China Math. 2019;62(1):1–32.
48. Hong Q, Xu J. Uniform stability and error analysis for some discontinuous Galerkin methods. arXiv preprint arXiv:180509670. 2018.
49. Schötzau D, Schwab C, Toselli A. Mixed *hp*-DGFEM for incompressible flows. SIAM J Numer Anal. 2002;40(6):2171–2194.
50. Hong Q, Kraus J. Uniformly stable discontinuous Galerkin discretization and robust iterative solution methods for the Brinkman problem. SIAM J Numer Anal. 2016;54(5):2750–2774.
51. Hong Q, Kraus J, Xu J, Zikatanov L. A robust multigrid method for discontinuous Galerkin discretizations of Stokes and linear elasticity equations. Numerische Mathematik. 2016;132(1):23–49.
52. Kraus J, Lazarov R, Lymbery M, Margenov S, Zikatanov L. Preconditioning heterogeneous $H(div)$ problems by additive Schur complement approximation and applications. SIAM J Sci Comput. 2016;38(2):A875–A898.
53. Vassilevski PS, Lazarov RD. Preconditioning mixed finite element saddle-point elliptic problems. Numer Linear Algebra Appl. 1996;3(1):1–20.
54. Arnold DN, Falk RS, Winther R. Preconditioning in $H(\mathrm{div})$ and applications. Math Comput. 1997;66:957–984.
55. Hiptmair R, Xu J. Nodal auxiliary space preconditioning in $H(\mathrm{curl})$ and $H(\mathrm{div})$ spaces. SIAM J Numer Anal. 2007;45(6):2483–2509.

56. Alnæs MS, Blechta J, Hake J, et al. The FEniCS Project Version 1.5. Arch Numer Softw. 2015;3(100):9-23.
57. Logg A, Mardal K-A, Wells GN, editors. Automated solution of differential equations by the finite element method. New York, NY: Springer; 2012.
58. National Agency for Finite Element Methods & Standards (Great Britain). The standard NAFEMS benchmarks. NAFEMS: Glasgow, UK; 1990.
59. Kolesov AE, Vabishchevich PN. Splitting schemes with respect to physical processes for double-porosity poroelasticity problems. Russ J Numer Anal Math Model. 2017;32(2):99-113.
60. Lee J, Cookson A, Chabiniok R, et al. Multiscale modelling of cardiac perfusion. In: Modeling the heart and the circulatory system. Cham, Switzerland: Springer International Publishing, 2015; p. 51–96.

How to cite this article: Hong Q, Kraus J, Lymbery M, Philo F. Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models. *Numer Linear Algebra Appl*. 2019;e2242. https://doi.org/10.1002/nla.2242

PARAMETER-ROBUST CONVERGENCE ANALYSIS OF FIXED-STRESS SPLIT ITERATIVE METHOD FOR MULTIPLE-PERMEABILITY POROELASTICITY SYSTEMS

MULTISCALE MODEL. SIMUL.
Vol. 18, No. 2, pp. 916–941

PARAMETER-ROBUST CONVERGENCE ANALYSIS OF FIXED-STRESS SPLIT ITERATIVE METHOD FOR MULTIPLE-PERMEABILITY POROELASTICITY SYSTEMS*

QINGGUO HONG†, JOHANNES KRAUS‡, MARIA LYMBERY‡, AND MARY F. WHEELER§

Abstract. We consider flux-based multiple-porosity/multiple-permeability poroelasticity systems describing mulitple-network flow and deformation in a poroelastic medium, also referred to as MPET models. The focus of the paper is on the convergence analysis of the fixed-stress split iteration, a commonly used coupling technique for the flow and mechanics equations defining poromechanical systems. We formulate the fixed-stress split method in this context and prove its linear convergence. The contraction rate of this fixed-point iteration does not depend on any of the physical parameters appearing in the model. This is confirmed by numerical results which further demonstrate the advantage of the fixed-stress split scheme over a preconditioned MinRes solver accelerated by norm-equivalent preconditioning.

Key words. multiple-porosity/multiple-permeability poroelasiticity, MPET system, fixed-stress split iterative coupling, convergence analysis

AMS subject classifications. 65M12, 65M60, 65F10, 65N22, 35Q92

DOI. 10.1137/19M1253988

1. Introduction.

Double-porosity poroelasticity models have been used to describe the motion of liquids in fissured rocks as early as in [5]. As a generalization of Biot's theory of consolidation [7, 8], they have been further extended in the framework of multiple-network poroelastic theory (MPET) where the deformable elastic matrix is permeated by two or more fluid networks with differing porosities and permeabilities. The latter find important applications in biophysics and medicine; see [40, 15, 21, 41].

In a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, the mathematical model is described by the MPET system:

$$-\mathrm{div}\,\boldsymbol{\sigma} + \sum_{i=1}^{n} \alpha_i \nabla p_i = \boldsymbol{f} \quad \text{in} \quad \Omega \times (0,T), \tag{1.1a}$$

$$\boldsymbol{v}_i = -K_i \nabla p_i \quad \text{in} \quad \Omega \times (0,T), \tag{1.1b}$$

$$-\alpha_i \mathrm{div}\,\dot{\boldsymbol{u}} - \mathrm{div}\,\boldsymbol{v}_i - c_{p_i}\dot{p}_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} \beta_{ij}(p_i - p_j) = g_i \quad \text{in} \quad \Omega \times (0,T), \tag{1.1c}$$

where (1.1a) and (1.1b) are for $i = 1, \ldots, n$. Here

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\epsilon}(\boldsymbol{u}) + \lambda \mathrm{div}(\boldsymbol{u})\boldsymbol{I} \quad \text{and} \quad \boldsymbol{\epsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T), \tag{1.2}$$

*Received by the editors April 2, 2019; accepted for publication (in revised form) March 30, 2020; published electronically May 26, 2020.
https://doi.org/10.1137/19M1253988

†Department of Mathematics, Pennsylvania State University, State College, PA 16802 (huq11@psu.edu).

‡Faculty of Mathematics, University of Duisburg-Essen, Essen, 45127, Germany (johannes.kraus@uni-due.de, maria.lymbery@uni-due.de).

§Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX 78712 (mfw@ices.utexas.edu).

denote the effective stress and the strain tensor, respectively, and the Lamé parameters λ and μ are expressed in terms of the modulus of elasticity E and the Poisson ratio $\nu \in [0, 1/2)$ by $\lambda := \frac{\nu E}{(1+\nu)(1-2\nu)}$, $\mu := \frac{E}{2(1+\nu)}$. The displacement $\boldsymbol{u}$, fluxes $\boldsymbol{v}_i$, and corresponding pressures p_i, $i = 1, \ldots, n$, are the unknown physical quantities.

The constants α_i in (1.1a) are known as Biot–Willis parameters, while $\boldsymbol{f}$ represents the body force density. The hydraulic conductivity tensors K_i in (1.1b) are defined as the permeability divided by the viscosity of the ith network. The constants c_{p_i} in (1.1c) denote the constrained specific storage coefficients; see, e.g., [38]. The network transfer coefficients β_{ij} couple the network pressures and hence $\beta_{ij} = \beta_{ji}$. Fluid extractions or injections enter the system via the source terms g_i in (1.1c).

The system (1.1) is well-posed under proper boundary and initial conditions. For stability reasons, this system is discretized in time by an implicit method. This creates a coupled static problem in each time step. The latter can be solved fully implicit, using a loose or explicit coupling, or an iterative coupling. In general, the loosely or explicitly coupled approach is less accurate than the fully implicit one which, however, is normally more computationally expensive. Iterative coupling is a commonly used alternative to avoid the disadvantages of the aforementioned approaches. The most popular procedures in this category are the undrained split, the fixed-stress split, the drained split, and the fixed-strain split iterative methods. As shown in [27], in contrast to the drained split and the fixed-strain split methods, the undrained split and fixed-stress split methods are unconditionally stable.

Convergence estimates and the rate of convergence for the latter methods have been derived in [35] for the quasi-static Biot system. The convergence and error analysis of an iterative coupling scheme for solving a fully discretized Biot system based on the fixed-stress split has been provided in [3]. Linear convergence in energy norms of a variant of the fixed-stress split iteration applied to heterogenous media has been shown in [13] for linearized Biot's equations.

Other variants of the fixed-stress split iterative scheme include a two-grid algorithm in which the flow subproblem of the Biot system is solved on a fine grid whereas the poromechanics subproblem is solved on a coarse grid (see [16]) or the multirate fixed-stress split iterative scheme which exploits different time scales for the mechanics and flow problems by taking several finer time steps for flow within one coarse time step for the mechanics of the system (see [2]).

The fixed-stress split scheme has also been successfully applied and proved convergent for space-time finite element approximations of the quasi-static Biot system; cf. [6]. In the context of unsaturated materials, it can be used for linearization of nonlinear poromechanics problems [10, 11]. When combined with Anderson acceleration, as shown in [12], this yields a highly efficient method. Other applications include fractured porous media [20] and fracture propagation [31]. The optimization of the stabilization parameter that serves the acceleration of the fixed-stress iterative method is considered for the Biot problem in the two-field formulation in [39].

In this paper we propose a fixed-stress split method for the MPET system. We prove its linear convergence and, furthermore, show with a proper choice for the stabilization parameter that the rate of convergence is independent of the physical parameters in the model. These theoretical findings are also tested computationally. The obtained numerical results support the proven convergence rate estimate and demonstrate the precedence of the fixed-stress split iterative method over the MinRes algorithm with norm-equivalent preconditioning.

The remainder of the paper is structured as follows. In section 2 we introduce notation and recall some important stability properties of the flux-based MPET system

(see [24], and also [23] for the special case of Biot's system), which are to be used later. Section 3 contains the main contribution of the paper. There, the fixed-stress algorithm for the MPET system is formulated and a parameter-robust convergence rate estimate proven. Section 4 discusses a discrete MPET model to which the main results from section 3 are then transferred in section 5. Numerical tests for the proposed fixed-stress split iterative coupling scheme are presented in section 6. Section 7 gives concluding remarks.

2. Properties of the flux-based MPET problem. It will be convenient for our exposition to rescale the MPET equations (1.1) and to represent the time-discrete problem in operator form. We start with imposing the following boundary and initial conditions that guarantee the well-posedness of system (1.1):

$$
\begin{aligned}
p_i(\boldsymbol{x},t) &= p_{i,D}(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x}\in\Gamma_{p_i,D}, \quad t>0, \quad i=1,\dots,n,\\
\boldsymbol{v}_i(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x}) &= q_{i,N}(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x}\in\Gamma_{p_i,N}, \quad t>0, \quad i=1,\dots,n,\\
\boldsymbol{u}(\boldsymbol{x},t) &= \boldsymbol{u}_D(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x}\in\Gamma_{\boldsymbol{u},D}, \quad t>0,\\
\left(\boldsymbol{\sigma}(\boldsymbol{x},t)-\sum_{i=1}^{n}\alpha_i p_i\boldsymbol{I}\right)\boldsymbol{n}(\boldsymbol{x}) &= \boldsymbol{g}_N(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x}\in\Gamma_{\boldsymbol{u},N}, \quad t>0,\\
p_i(\boldsymbol{x},0) &= p_{i,0}(\boldsymbol{x}), \qquad \boldsymbol{x}\in\Omega, \quad i=1,\dots,n,\\
\boldsymbol{u}(\boldsymbol{x},0) &= \boldsymbol{u}_0(\boldsymbol{x}), \qquad \boldsymbol{x}\in\Omega,
\end{aligned}
$$

where $\Gamma_{p_i,D}\cap\Gamma_{p_i,N}=\emptyset$, $\overline{\Gamma}_{p_i,D}\cup\overline{\Gamma}_{p_i,N}=\Gamma=\partial\Omega$ for $i=1,\dots,n$, $\Gamma_{\boldsymbol{u},D}\cap\Gamma_{\boldsymbol{u},N}=\emptyset$, and $\overline{\Gamma}_{\boldsymbol{u},D}\cup\overline{\Gamma}_{\boldsymbol{u},N}=\Gamma$ are fulfilled.

Using the backward Euler method for time discretization (see also [24]), one has to solve a static problem of the form

$$
-2\mu\operatorname{div}\boldsymbol{\epsilon}(\boldsymbol{u}^k)-\lambda\nabla\operatorname{div}\boldsymbol{u}^k+\sum_{i=1}^{n}\alpha_i\nabla p_i^k=\boldsymbol{f}^k, \tag{2.1a}
$$

$$
K_i^{-1}\boldsymbol{v}_i^k+\nabla p_i^k=\boldsymbol{0}, \qquad i=1,\dots,n, \tag{2.1b}
$$

$$
-\alpha_i\operatorname{div}\boldsymbol{u}^k-\tau\operatorname{div}\boldsymbol{v}_i^k-c_{p_i}p_i^k-\tau\sum_{\substack{j=1\\ j\neq i}}^{n}\beta_{ij}(p_i^k-p_j^k)=g_i^k, \qquad i=1,\dots,n, \tag{2.1c}
$$

at every time moment $t_k=t_{k-1}+\tau$, $k=1,2,\dots$. In system (2.1), the unknown time-step functions $\boldsymbol{u}^k$, $\boldsymbol{v}_i^k$, p_i^k approximate $\boldsymbol{u}$, $\boldsymbol{v}_i$, p_i at $t=t_k$, the right-hand sides are given by $\boldsymbol{f}^k=\boldsymbol{f}(x,t_k)$, $g_i^k=-\tau g_i(x,t_k)-\alpha_i\operatorname{div}(\boldsymbol{u}^{k-1})-c_{p_i}p_i^{k-1}$ for $i=1,\dots,n$, and τ denotes the time step size. We divide (2.1) by 2μ and, for convenience, denote

$$
\frac{\lambda}{2\mu}\to\lambda,\ \frac{\alpha_i}{2\mu}\to\alpha_i,\ \frac{\boldsymbol{f}^k}{2\mu}\to\boldsymbol{f}^k,\ \frac{\tau}{2\mu}\to\tau,\ \frac{c_{p_i}}{2\mu}\to c_{p_i},\ \frac{g_i^k}{2\mu}\to g_i^k, \quad i=1,\dots,n.
$$

Next, we multiply (2.1b) by α_i and (2.1c) by α_i^{-1} and introduce the new variables

$$
\boldsymbol{v}_i:=\frac{\tau}{\alpha_i}\boldsymbol{v}_i^k, \quad p_i:=\alpha_i p_i^k, \quad \boldsymbol{u}:=\boldsymbol{u}^k, \quad \boldsymbol{f}:=\boldsymbol{f}^k, \quad g_i:=\frac{g_i^k}{\alpha_i}, \quad i=1,\dots,n.
$$

Then system (2.1) takes the form

$$-\operatorname{div}\boldsymbol{\epsilon}(\boldsymbol{u}) - \lambda\nabla\operatorname{div}\boldsymbol{u} + \sum_{i=1}^{n}\nabla p_i = \boldsymbol{f}, \tag{2.2a}$$

$$\tau^{-1}K_i^{-1}\alpha_i^2\boldsymbol{v}_i + \nabla p_i = \boldsymbol{0}, \qquad i = 1,\dots,n, \tag{2.2b}$$

$$-\operatorname{div}\boldsymbol{u} - \operatorname{div}\boldsymbol{v}_i - \frac{c_{p_i}}{\alpha_i^2}p_i + \sum_{\substack{j=1\\ j\neq i}}^{n}\left(-\frac{\tau\beta_{ij}}{\alpha_i^2}p_i + \frac{\tau\beta_{ij}}{\alpha_i\alpha_j}p_j\right) = g_i, \qquad i = 1,\dots,n, \tag{2.2c}$$

in the new variables. Finally, using the parameter substitutions

$$R_i^{-1} := \tau^{-1}K_i^{-1}\alpha_i^2, \quad \alpha_{p_i} := \frac{c_{p_i}}{\alpha_i^2}, \quad \beta_{ii} := \sum_{\substack{j=1\\ j\neq i}}^{n}\beta_{ij}, \quad \alpha_{ij} := \frac{\tau\beta_{ij}}{\alpha_i\alpha_j}, \quad \tilde{\alpha}_{ii} := -\alpha_{p_i} - \alpha_{ii}$$

for $i, j = 1,\dots,n$, system (2.2) can be represented in operator notation by

$$\mathscr{A}\left[\boldsymbol{u}^T, \boldsymbol{v}_1^T, \dots, \boldsymbol{v}_n^T, p_1, \dots, p_n\right]^T = \left[\boldsymbol{f}^T, \boldsymbol{0}^T, \dots, \boldsymbol{0}^T, g_1, \dots, g_n\right]^T, \tag{2.3}$$

where

$$\mathscr{A} := \begin{bmatrix}
-\operatorname{div}\epsilon - \lambda\nabla\operatorname{div} & 0 & \dots & \dots & 0 & \nabla & \dots & \dots & \nabla \\
0 & R_1^{-1}I & 0 & \dots & 0 & \nabla & 0 & \dots & 0 \\
\vdots & 0 & \ddots & & \vdots & 0 & \ddots & & \vdots \\
\vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & 0 \\
0 & 0 & \dots & 0 & R_n^{-1}I & 0 & \dots & 0 & \nabla \\
-\operatorname{div} & -\operatorname{div} & 0 & \dots & 0 & \tilde{\alpha}_{11}I & \alpha_{12}I & \dots & \alpha_{1n}I \\
\vdots & 0 & \ddots & & \vdots & \alpha_{21}I & \ddots & & \alpha_{2n}I \\
\vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & \vdots \\
-\operatorname{div} & 0 & \dots & 0 & -\operatorname{div} & \alpha_{n1}I & \alpha_{n2}I & \dots & \tilde{\alpha}_{nn}I
\end{bmatrix} \tag{2.4}$$

is the rescaled operator. For the scaled parameters we make the following plausible and quite nonrestrictive assumptions, namely,

$$\lambda > 0, \quad R_1^{-1},\dots,R_n^{-1} > 0, \quad \alpha_{p_1},\dots,\alpha_{p_n} \geq 0, \quad \alpha_{ij} \geq 0, \quad i,j = 1,\dots,n. \tag{2.5}$$

2.1. Preliminaries and notation. Let us denote $\boldsymbol{v}^T := (\boldsymbol{v}_1^T,\dots,\boldsymbol{v}_n^T)$, $\boldsymbol{z}^T := (\boldsymbol{z}_1^T,\dots,\boldsymbol{z}_n^T)$, $\boldsymbol{p}^T := (p_1,\dots,p_n)$, $\boldsymbol{q}^T := (q_1,\dots,q_n)$ where $\boldsymbol{v},\boldsymbol{z} \in \boldsymbol{V} = \boldsymbol{V}_1 \times \cdots \times \boldsymbol{V}_n$, $\boldsymbol{p},\boldsymbol{q} \in \boldsymbol{P} = P_1 \times \cdots \times P_n$ and $\boldsymbol{U} := \{\boldsymbol{u} \in H^1(\Omega)^d : \boldsymbol{u} = \boldsymbol{0} \text{ on } \Gamma_{\boldsymbol{u},D}\}$, $\boldsymbol{V}_i := \{\boldsymbol{v}_i \in H(\operatorname{div},\Omega) : \boldsymbol{v}_i\cdot\boldsymbol{n} = 0 \text{ on } \Gamma_{p_i,N}\}$, $P_i := L^2(\Omega)$, and $P_i := L_0^2(\Omega)$ if $\Gamma_{\boldsymbol{u},D} = \Gamma = \partial\Omega$.

The weak formulation of system (2.3) reads: find $(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}) \in \boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$, such that for any $(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}) \in \boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$ there hold

$$(\epsilon(\boldsymbol{u}),\epsilon(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{w}) - \sum_{i=1}^{n}(p_i,\operatorname{div}\boldsymbol{w}) = (\boldsymbol{f},\boldsymbol{w}), \tag{2.6a}$$

$$(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{z}_i) - (p_i,\operatorname{div}\boldsymbol{z}_i) = 0, \qquad i = 1,\dots,n, \tag{2.6b}$$

$$-(\operatorname{div}\boldsymbol{u},q_i) - (\operatorname{div}\boldsymbol{v}_i,q_i) + \tilde{\alpha}_{ii}(p_i,q_i) + \sum_{\substack{j=1\\ j\neq i}}^{n}\alpha_{ij}(p_j,q_i) = (g_i,q_i), \quad i = 1,\dots,n, \tag{2.6c}$$

or, equivalentely, $\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) = F(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})$ for all $(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}) \in \boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$, where

$$F(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}) = (\boldsymbol{f},\boldsymbol{w}) + \sum_{i=1}^{n}(g_i,q_i) \quad \text{and}$$

$$\begin{aligned}\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})) &= (\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{w}) - \sum_{i=1}^{n}(p_i,\operatorname{div}\boldsymbol{w})\\ &\quad + \sum_{i=1}^{n}(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{z}_i) - (\boldsymbol{p},\operatorname{Div}\boldsymbol{z}) - \sum_{i=1}^{n}(\operatorname{div}\boldsymbol{u},q_i)\\ &\quad - (\operatorname{Div}\boldsymbol{v},\boldsymbol{q}) - \sum_{i=1}^{n}(\alpha_{p_i}+\alpha_{ii})(p_i,q_i) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\alpha_{ji}(p_j,q_i)\\ &= (\boldsymbol{\epsilon}(\boldsymbol{u}),\boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{w}) - \sum_{i=1}^{n}(p_i,\operatorname{div}\boldsymbol{w})\\ &\quad + \sum_{i=1}^{n}(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{z}_i) - (\boldsymbol{p},\operatorname{Div}\boldsymbol{z}) - \sum_{i=1}^{n}(\operatorname{div}\boldsymbol{u},q_i)\\ &\quad - (\operatorname{Div}\boldsymbol{v},\boldsymbol{q}) - ((\Lambda_1+\Lambda_2)\boldsymbol{p},\boldsymbol{q}).\end{aligned}$$

Here we have denoted $(\operatorname{Div}\boldsymbol{y})^T := (\operatorname{div}\boldsymbol{y}_1,\ldots,\operatorname{div}\boldsymbol{y}_n)$ for $\boldsymbol{y}\in\boldsymbol{V}$ and

$$\Lambda_1 := \begin{bmatrix}\alpha_{11} & -\alpha_{12} & \ldots & -\alpha_{1n}\\ -\alpha_{21} & \alpha_{22} & \ldots & -\alpha_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ -\alpha_{n1} & -\alpha_{n2} & \ldots & \alpha_{nn}\end{bmatrix}, \quad \Lambda_2 := \begin{bmatrix}\alpha_{p_1} & 0 & \ldots & 0\\ 0 & \alpha_{p_2} & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \alpha_{p_n}\end{bmatrix}.$$

Furthermore, define $R^{-1} := \max\{R_1^{-1},\ldots,R_n^{-1}\}$, $\lambda_0 := \max\{1,\lambda\}$, and the $n\times n$ matrices

$$\Lambda_3 := \begin{bmatrix}R & 0 & \ldots & 0\\ 0 & R & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & R\end{bmatrix}, \quad \Lambda_4 := \begin{bmatrix}\frac{1}{\lambda_0} & \cdots & \cdots & \frac{1}{\lambda_0}\\ \vdots & & & \vdots\\ \vdots & & & \vdots\\ \frac{1}{\lambda_0} & \cdots & \cdots & \frac{1}{\lambda_0}\end{bmatrix}$$

that are used later in the convergence analysis of the fixed-stress split iterative method. It is easy to show that Λ_i are symmetric positive semidefinite for $i=1,2,4$, while Λ_3 is symmetric positive definite (SPD).

Moreover, we denote $\Lambda := \sum_{i=1}^{4}\Lambda_i$, which obviously is an SPD matrix and therefore can be used to define the parameter-matrix-dependent norms $\|\cdot\|_{\boldsymbol{U}}$, $\|\cdot\|_{\boldsymbol{V}}$, $\|\cdot\|_{\boldsymbol{P}}$ induced by the inner products:

$$(\boldsymbol{u},\boldsymbol{w})_{\boldsymbol{U}} = (\epsilon(\boldsymbol{u}),\epsilon(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{w}), \tag{2.7a}$$

$$(\boldsymbol{v},\boldsymbol{z})_{\boldsymbol{V}} = \sum_{i=1}^{n}(R_i^{-1}\boldsymbol{v}_i,\boldsymbol{z}_i) + (\Lambda^{-1}\operatorname{Div}\boldsymbol{v},\operatorname{Div}\boldsymbol{z}), \tag{2.7b}$$

$$(\boldsymbol{p},\boldsymbol{q})_{\boldsymbol{P}} = (\Lambda\boldsymbol{p},\boldsymbol{q}). \tag{2.7c}$$

As shown in [24], these norms are crucial to show the parameter-robust stability of the MPET system.

2.2. Stability properties. The following inf-sup conditions for the spaces $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{P}$ are assumed to be fulfilled in the analysis presented in this paper:

$$\inf_{q\in P_i}\sup_{\boldsymbol{v}\in\boldsymbol{V}_i}\frac{(\mathrm{div}\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\mathrm{div}}\|q\|}\geq\beta_d,\quad i=1,\dots,n, \tag{2.8}$$

$$\inf_{(q_1,\cdots,q_n)\in P_1\times\cdots\times P_n}\sup_{\boldsymbol{u}\in\boldsymbol{U}}\frac{\left(\mathrm{div}\boldsymbol{u},\sum_{i=1}^{n}q_i\right)}{\|\boldsymbol{u}\|_1\left\|\sum_{i=1}^{n}q_i\right\|}\geq\beta_s \tag{2.9}$$

for some constants $\beta_d>0$ and $\beta_s>0$; see [14, 9]. Then from [24], we know that the MPET problem (2.6) is uniformly well-posed, namely the three assertions in Theorem 2.1 hold.

THEOREM 2.1.

(i) *There exists a positive constant C_b independent of the parameters λ, R_i^{-1}, α_{p_i}, α_{ij}, $i,j\in\{1,\dots,n\}$ and the network scale n such that the inequality*

$$|\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}))|\leq C_b(\|\boldsymbol{u}|_{\boldsymbol{U}}+\|\boldsymbol{v}\|_{\boldsymbol{V}}+\|\boldsymbol{p}\|_{\boldsymbol{P}})(\|\boldsymbol{w}\|_{\boldsymbol{U}}+\|\boldsymbol{z}\|_{\boldsymbol{V}}+\|\boldsymbol{q}\|_{\boldsymbol{P}})$$

holds true for any $(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p})\in\boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$, $(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})\in\boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$.

(ii) *There is a constant $\omega>0$ independent of the parameters $\lambda, R_i^{-1},\alpha_{p_i},\alpha_{ij}$, $i,j\in\{1,\dots,n\}$ and the number of networks n such that*

$$\inf_{(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p})\in\boldsymbol{X}}\sup_{(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q})\in\boldsymbol{X}}\frac{\mathcal{A}((\boldsymbol{u};\boldsymbol{v};\boldsymbol{p}),(\boldsymbol{w};\boldsymbol{z};\boldsymbol{q}))}{(\|\boldsymbol{u}\|_{\boldsymbol{U}}+\|\boldsymbol{v}\|_{\boldsymbol{V}}+\|\boldsymbol{p}\|_{\boldsymbol{P}})(\|\boldsymbol{w}\|_{\boldsymbol{U}}+\|\boldsymbol{z}\|_{\boldsymbol{V}}+\|\boldsymbol{q}\|_{\boldsymbol{P}})}\geq\omega, \tag{2.10}$$

where $\boldsymbol{X}:=\boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$.

(iii) *The MPET system* (2.6) *has a unique solution $(\boldsymbol{u};\boldsymbol{v};\boldsymbol{p})\in\boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$ and the following stability estimate holds:*

$$\|\boldsymbol{u}\|_{\boldsymbol{U}}+\|\boldsymbol{v}\|_{\boldsymbol{V}}+\|\boldsymbol{p}\|_{\boldsymbol{P}}\leq C_1(\|\boldsymbol{f}\|_{\boldsymbol{U}^*}+\|\boldsymbol{g}\|_{\boldsymbol{P}^*}), \tag{2.11}$$

where C_1 is a positive constant independent of the parameters λ, R_i^{-1}, α_{p_i}, $\alpha_{ij}, i,j\in\{1,\dots,n\}$ and the network scale n, and $\|\boldsymbol{f}\|_{\boldsymbol{U}^*}=\sup_{\boldsymbol{w}\in\boldsymbol{U}}\frac{(\boldsymbol{f},\boldsymbol{w})}{\|\boldsymbol{w}\|_{\boldsymbol{U}}}$, $\|\boldsymbol{g}\|_{\boldsymbol{P}^*}=\sup_{\boldsymbol{q}\in\boldsymbol{P}}\frac{(\boldsymbol{g},\boldsymbol{q})}{\|\boldsymbol{q}\|_{\boldsymbol{P}}}=\|\Lambda^{-\frac{1}{2}}\boldsymbol{g}\|$.

2.3. A norm-equivalent and a field-of-values-equivalent preconditioner. Consider the block-diagonal operator

$$\mathscr{B}_N:=\begin{bmatrix}\mathcal{A}_{\boldsymbol{u}}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathcal{S}_{\boldsymbol{v}}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathcal{S}_{\boldsymbol{p}}\end{bmatrix}^{-1}, \tag{2.12}$$

where

$$\mathcal{A}_{\boldsymbol{u}}=-\mathrm{div}\,\boldsymbol{\epsilon}-\lambda\nabla\mathrm{div}\,, \tag{2.13}$$

$$\mathcal{S}_{\boldsymbol{v}}=\begin{bmatrix}R_1^{-1}I&0&\dots&0\\ 0&R_2^{-1}I&\dots&0\\ \vdots&\vdots&\ddots&\vdots\\ 0&0&\dots&R_n^{-1}I\end{bmatrix}-\begin{bmatrix}\tilde{\gamma}_{11}\nabla\mathrm{div}&\tilde{\gamma}_{12}\nabla\mathrm{div}&\dots&\tilde{\gamma}_{1n}\nabla\mathrm{div}\\ \tilde{\gamma}_{21}\nabla\mathrm{div}&\tilde{\gamma}_{22}\nabla\mathrm{div}&\dots&\tilde{\gamma}_{2n}\nabla\mathrm{div}\\ \vdots&\vdots&\ddots&\vdots\\ \tilde{\gamma}_{n1}\nabla\mathrm{div}&\tilde{\gamma}_{n2}\nabla\mathrm{div}&\dots&\tilde{\gamma}_{nn}\nabla\mathrm{div}\end{bmatrix}, \tag{2.14}$$

and

$$\mathcal{S}_{\boldsymbol{p}} = \begin{bmatrix} \gamma_{11}I & \gamma_{12}I & \dots & \gamma_{1n}I \\ \gamma_{21}I & \gamma_{22}I & \dots & \gamma_{2n}I \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1}I & \gamma_{n2}I & \dots & \gamma_{nn}I \end{bmatrix}. \tag{2.15}$$

Here, γ_{ij}, $\tilde{\gamma}_{ij}$, $i,j = 1,\dots,n$, are the entries of Λ and Λ^{-1}, respectively.

As substantiated in [24, 34], the stability results for the operator $\mathscr{A}$ imply that the operator $\mathscr{B}_N$ defined in (2.12) is a uniform norm-equivalent (canonical) block-diagonal preconditioner that is robust with respect to all model and discretization parameters. An analogous uniform block-diagonal preconditioner exists also on the discrete level if discrete inf-sup conditions analogous to (2.8) and (2.9) are satisfied; cf. [24].

In addition, we can also consider the field-of-values-equivalent (FOV-equivalent) preconditioner

$$\mathscr{B}_F := \begin{bmatrix} \mathcal{A}_{\boldsymbol{u}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{S}_{\boldsymbol{v}} & \mathbf{0} \\ -\mathcal{B}_{\boldsymbol{u}} & \mathcal{B}_{\boldsymbol{v}} & \mathcal{S}_{\boldsymbol{p}} \end{bmatrix}^{-1}, \tag{2.16}$$

where

$$\mathcal{B}_{\boldsymbol{u}} := \begin{bmatrix} -\mathrm{div} \\ \vdots \\ \vdots \\ -\mathrm{div} \end{bmatrix}, \quad \mathcal{B}_{\boldsymbol{v}} := \begin{bmatrix} -\mathrm{div} & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & -\mathrm{div} \end{bmatrix} \tag{2.17}$$

and $\mathcal{A}_{\boldsymbol{u}}$, $\mathcal{S}_{\boldsymbol{v}}$, $\mathcal{S}_{\boldsymbol{p}}$ are defined in (2.13), (2.14), (2.15).

The field-of-values equivalence of the block-triangular preconditioner (2.16) can be proven following the theory presented in [17, 18, 33]. Related results on parameter-robust block-preconditioners for Biot's consolidation model can be found in [1, 29] and in [24, 30] in the context of the MPET model.

3. Fixed-stress method for MPET model. In the proposed fixed-stress split iterative coupling scheme for the MPET system, and as for Biot's equations, we first solve the flow and then the mechanics problem where, in order to avoid instabilities, a stabilization term is added to the flow equation. Note that generalizing the fixed-stress iteration from the Biot to the (flux-based) MPET model is not straightforward due to the involvement of n pressures p_i and n fluxes $\boldsymbol{v}_i$. Our formulation suggests a stabilization that employs the sum of the pressures which later shows itself to be vital for the convergence properties of the scheme.

In order to elucidate our approach, we present the fixed-stress split iterative scheme for the continuous problem first. Let $\boldsymbol{u}^k$, $\boldsymbol{v}_i^k$, and p_i^k denote the kth fixed-stress iterates for $\boldsymbol{u}$, $\boldsymbol{v}_i$, and p_i, respectively, $i = 1,\dots,n$. The single rate fixed-stress split iterative method is given by the following algorithm.

Algorithm 3.1: Fixed-stress coupling iteration for the MPET system

Step a: Given $\boldsymbol{u}^m$, we solve for $\boldsymbol{v}_i^{m+1}$ and p_i^{m+1}

$$(-\operatorname{div}\boldsymbol{v}_i^{m+1}, q_i) - ((\alpha_{p_i}+\alpha_{ii})p_i^{m+1}, q_i) + \left(\sum_{\substack{j=1\\ j\neq i}}^{n}\alpha_{ij}p_j^{m+1}, q_i\right) - L\left(\sum_{j=1}^{n}p_j^{m+1}, q_i\right)$$
$$= (g_i, q_i) - L\left(\sum_{j=1}^{n}p_j^{m}, q_i\right) + (\operatorname{div}\boldsymbol{u}^m, q_i), \quad 1\le i\le n,$$

and

$$(R_i^{-1}\boldsymbol{v}_i^{m+1}, \boldsymbol{z}_i) - (p_i^{m+1}, \operatorname{div}\boldsymbol{z}_i) = \boldsymbol{0}, \quad 1\le i\le n.$$

Step b: Given $\boldsymbol{v}_i^{m+1}$ and p_i^{m+1}, we solve for $\boldsymbol{u}^{m+1}$

$$(\boldsymbol{\epsilon}(\boldsymbol{u}^{m+1}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{u}^{m+1}, \operatorname{div}\boldsymbol{w}) = (\boldsymbol{f}, \boldsymbol{w}) + \sum_{i=1}^{n}(p_i^{m+1}, \operatorname{div}\boldsymbol{w}).$$

Our main result is formulated in terms of the following quantities:

$$\boldsymbol{e}_u^k = \boldsymbol{u}^k - \boldsymbol{u} \in \boldsymbol{U}, \tag{3.1a}$$
$$\boldsymbol{e}_{v_i}^k = \boldsymbol{v}_i^k - \boldsymbol{v}_i \in \boldsymbol{V}_i, \quad i = 1,\dots,n, \tag{3.1b}$$
$$e_{p_i}^k = p_i^k - p_i \in P_i, \quad i = 1,\dots,n, \tag{3.1c}$$

denoting the errors of the kth iterates $\boldsymbol{u}^k$, $\boldsymbol{v}_i^k$, p_i^k, $i = 1,\dots,n$, generated by Algorithm 3.1. The error block-vectors $\boldsymbol{e}_v^k$ and $\boldsymbol{e}_p^k$ are given by $(\boldsymbol{e}_v^k)^T = ((\boldsymbol{e}_{v_1}^k)^T,\dots,(\boldsymbol{e}_{v_n}^k))^T$ and $(\boldsymbol{e}_p^k)^T = (e_{p_1}^k,\dots,e_{p_n}^k)$. Since $\boldsymbol{u}$, $\boldsymbol{v}_i$, p_i, $i = 1,\dots,n$, are the exact solutions of (2.6), the error equations

$$(-\mathrm{Div}\boldsymbol{e}_v^{m+1}, \boldsymbol{q}) - ((\Lambda_1+\Lambda_2)\boldsymbol{e}_p^{m+1}, \boldsymbol{q}) - L\left(\sum_{i=1}^{n}e_{p_i}^{m+1}, \sum_{i=1}^{n}q_i\right) = -L\left(\sum_{i=1}^{n}e_{p_i}^{m}, \sum_{i=1}^{n}q_i\right)$$
$$+\left(\operatorname{div}\boldsymbol{e}_u^m, \sum_{i=1}^{n}q_i\right), \tag{3.2a}$$
$$(R^{-1}\boldsymbol{e}_v^{m+1}, \boldsymbol{z}) - (\boldsymbol{e}_p^{m+1}, \mathrm{Div}\boldsymbol{z}) = 0, \tag{3.2b}$$
$$(\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{e}_u^{m+1}, \operatorname{div}\boldsymbol{w}) = \left(\sum_{i=1}^{n}e_{p_i}^{m+1}, \operatorname{div}\boldsymbol{w}\right) \tag{3.2c}$$

hold, the latter of which plays a key role in the presented convergence analysis.

Note that in the following we do not make any further restrictive assumptions on the parameters in (2.6) but consider the general situation in which only (2.5) needs to be satisfied. Useful for deriving and defining the tuning parameter L is the constant c_K in the estimate

$$\|\boldsymbol{\epsilon}(\boldsymbol{w})\| \geq c_K \|\mathrm{div}\,(\boldsymbol{w})\| \quad \text{for all } \boldsymbol{w} \in \boldsymbol{U} \tag{3.3}$$

which is used for $\boldsymbol{w} = \boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^m$ in the proof of the next Lemma.[1]

We perform the convergence analysis in two steps. The first one is the proof of the following lemma.

LEMMA 3.1. *The errors* $\boldsymbol{e}_u^{m+1}$, $\boldsymbol{e}_v^{m+1}$, *and* $\boldsymbol{e}_p^{m+1}$ *of the* $(m+1)$*st fixed-stress iterate generated by Algorithm* 3.1 *for* $L \geq \frac{1}{\lambda + c_K^2}$ *satisfy the estimate*

$$\begin{aligned} &\frac{1}{2}\Big(\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2 + \lambda\|\mathrm{div}\,\boldsymbol{e}_u^{m+1}\|^2\Big) + \|R^{-1/2}\boldsymbol{e}_v^{m+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{1/2}\boldsymbol{e}_p^{m+1}\|^2 \\ &\quad + \frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|^2 \leq \frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^{m}\right\|^2, \quad m = 0, 1, 2, \ldots. \end{aligned} \tag{3.4}$$

Proof. Setting $\boldsymbol{z} = \boldsymbol{e}_v^{m+1}, \boldsymbol{q} = -\boldsymbol{e}_p^{m+1}, \boldsymbol{w} = \boldsymbol{e}_u^{m+1}$ in (3.2a)–(3.2c), it follows that

$$\begin{aligned} &\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2 + \lambda\|\mathrm{div}\,\boldsymbol{e}_u^{m+1}\|^2 + \|R^{-1/2}\boldsymbol{e}_v^{m+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{1/2}\boldsymbol{e}_p^{m+1}\|^2 \\ &\quad + L\left(\sum_{i=1}^n (e_{p_i}^{m+1} - e_{p_i}^m), \sum_{i=1}^n e_{p_i}^{m+1}\right) = \left(\mathrm{div}\,(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^m), \sum_{i=1}^n e_{p_i}^{m+1}\right). \end{aligned} \tag{3.5}$$

Using the identity

$$\begin{aligned} &\left(\sum_{i=1}^n (e_{p_i}^{m+1} - e_{p_i}^m), \sum_{i=1}^n e_{p_i}^{m+1}\right) \\ &\quad = \frac{1}{2}\left(\left\|\sum_{i=1}^n e_{p_i}^{m+1} - \sum_{i=1}^n e_{p_i}^m\right\|^2 + \left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|^2 - \left\|\sum_{i=1}^n e_{p_i}^m\right\|^2\right), \end{aligned}$$

(3.5) can be rewritten as

$$\begin{aligned} &\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2 + \lambda\|\mathrm{div}\,\boldsymbol{e}_u^{m+1}\|^2 + \|R^{-1/2}\boldsymbol{e}_v^{m+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{1/2}\boldsymbol{e}_p^{m+1}\|^2 \\ &\quad + \frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|^2 + \frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^{m+1} - \sum_{i=1}^n e_{p_i}^m\right\|^2 \\ &= \frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^m\right\|^2 + \left(\mathrm{div}\,(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^m), \sum_{i=1}^n e_{p_i}^{m+1}\right). \end{aligned} \tag{3.6}$$

Now, taking $\boldsymbol{w} = \boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^m$ in (3.2c) we obtain

$$\begin{aligned} &\left(\mathrm{div}\,(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^m), \sum_{i=1}^n e_{p_i}^{m+1}\right) \\ &\quad = (\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1}), \boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^m)) + \lambda(\mathrm{div}\,\boldsymbol{e}_u^{m+1}, \mathrm{div}\,(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^m)) \end{aligned} \tag{3.7}$$

[1]It can easily be shown that estimate (3.3) holds true for $c_K = 1/\sqrt{d}$, where d is the space dimension.

and, substituting (3.7) in (3.6), conclude that

$$\begin{aligned}
&\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2 + \lambda\|\operatorname{div}\boldsymbol{e}_u^{m+1}\|^2 + \|R^{-1/2}\boldsymbol{e}_v^{m+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{1/2}\boldsymbol{e}_p^{m+1}\|^2 \\
&\quad + \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2 + \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1} - \sum_{i=1}^{n} e_{p_i}^{m}\right\|^2 \\
&= \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|^2 + (\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1}), \boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})) + \lambda(\operatorname{div}\boldsymbol{e}_u^{m+1}, \operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})) \\
&\leq \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|^2 + \frac{1}{2}\Big(\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2 + \lambda\|\operatorname{div}\boldsymbol{e}_u^{m+1}\|^2\Big) \\
&\quad + \frac{1}{2}\Big(\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2 + \lambda\|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2\Big).
\end{aligned}$$

The latter inequality can be expressed equivalently in the form

$$\begin{aligned}
&\frac{1}{2}\Big(\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2 + \lambda\|\operatorname{div}\boldsymbol{e}_u^{m+1}\|^2\Big) + \|R^{-1/2}\boldsymbol{e}_v^{m+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{1/2}\boldsymbol{e}_p^{m+1}\|^2 \\
&\quad + \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2 + \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1} - \sum_{i=1}^{n} e_{p_i}^{m}\right\|^2 \\
&\leq \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|^2 + \frac{1}{2}\Big(\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2 + \lambda\|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2\Big).
\end{aligned} \tag{3.8}$$

To estimate the last term in (3.8) consider (3.2c) again. Subtracting the mth error from the $(m+1)$st, choosing $\boldsymbol{w} = \boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m}$, and applying Cauchy's inequality yields

$$\begin{aligned}
&\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2 + \lambda\|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2 \\
&\qquad \leq \|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\| \left\|\sum_{i=1}^{n}(e_{p_i}^{m+1} - e_{p_i}^{m})\right\|.
\end{aligned} \tag{3.9}$$

Next, from (3.3) we have that $\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\| \geq c_K\|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|$, which implies

$$(c_K^2 + \lambda)\|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\| \leq \left\|\sum_{i=1}^{n} e_{p_i}^{m+1} - \sum_{i=1}^{n} e_{p_i}^{m}\right\|,$$

that is,

$$\|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\| \leq \frac{1}{\lambda + c_K^2}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1} - \sum_{i=1}^{n} e_{p_i}^{m}\right\|. \tag{3.10}$$

Hence

$$\begin{aligned}
&\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2 + \lambda\|\operatorname{div}(\boldsymbol{e}_u^{m+1} - \boldsymbol{e}_u^{m})\|^2 \\
&\qquad \leq \frac{1}{\lambda + c_K^2}\left\|\sum_{i=1}^{n}(e_{p_i}^{m+1} - e_{p_i}^{m})\right\|^2 \leq L\left\|\sum_{i=1}^{n}(e_{p_i}^{m+1} - e_{p_i}^{m})\right\|^2.
\end{aligned} \tag{3.11}$$

Therefore, using (3.11) in (3.8), we obtain

$$\begin{aligned}
&\frac{1}{2}\Big(\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2+\lambda\|\operatorname{div}\boldsymbol{e}_u^{m+1}\|^2\Big)+\|R^{-1/2}\boldsymbol{e}_v^{m+1}\|^2+\|(\Lambda_1+\Lambda_2)^{1/2}\boldsymbol{e}_p^{m+1}\|^2\\
&\qquad+\frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|^2+\frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^{m+1}-\sum_{i=1}^n e_{p_i}^{m}\right\|^2\\
&\le\frac{L}{2}\left\|\sum_{i=1}^n e_{p_i}^{m}\right\|^2+\frac{L}{2}\left\|\sum_{i=1}^n (e_{p_i}^{m+1}-e_{p_i}^{m})\right\|^2,
\end{aligned}$$

which completes the proof. ∎

Using (3.4), we can prove that $\sum_{i=1}^n e_{p_i}^m \overset{m\to\infty}{\longrightarrow} 0$, which is stated in the following theorem.

THEOREM 3.2. *Let c_K and β_s denote the constants in* (3.3) *and* (2.9)*, respectively. The single rate fixed-stress iterative method for the static MPET problem* (2.6) *defined in Algorithm* 3.1 *is a contraction that converges linearly for any $L \ge 1/(\lambda+c_K^2)$ independent of the model parameters and the time step size τ. The errors $\boldsymbol{e}_p^m$ in this case satisfy the inequality*

$$\left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|^2 \le \operatorname{rate}^2(\lambda)\left\|\sum_{i=1}^n e_{p_i}^{m}\right\|^2 \tag{3.12}$$

with

$$\operatorname{rate}^2(\lambda) \le \frac{1}{\frac{L^{-1}}{\beta_s^{-2}+\lambda}+1}. \tag{3.13}$$

For $L=\frac{1}{\lambda+c_K^2}$, the convergence factor in (3.12) *can be estimated by*

$$\operatorname{rate}^2(\lambda) \le \frac{1}{\frac{\lambda+c_K^2}{\beta_s^{-2}+\lambda}+1} \le \max\left\{\frac{\beta_s^{-2}}{c_K^2+\beta_s^{-2}},\frac{1}{2}\right\}. \tag{3.14}$$

Proof. By the Stokes inf-sup condition, we have that for any $\sum_{i=1}^n e_{p_i}^{m+1}$ there exists $\boldsymbol{w}_p \in \boldsymbol{U}$ such that

$$\operatorname{div}\boldsymbol{w}_p=\sum_{i=1}^n e_{p_i}^{m+1} \quad\text{and}\quad \|\boldsymbol{\epsilon}(\boldsymbol{w}_p)\| \le \beta_s^{-1}\left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|, \tag{3.15}$$

where β_s is the Stokes inf-sup constant in (2.9). Hence,

$$\|\boldsymbol{\epsilon}(\boldsymbol{w}_p)\|^2+\lambda\|\operatorname{div}\boldsymbol{w}_p\|^2 \le (\beta_s^{-2}+\lambda)\left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|^2.$$

Taking $\boldsymbol{w}=\boldsymbol{w}_p$ in (3.2c) and using (3.15) yields

$$\left\|\sum_{i=1}^n e_{p_i}^{m+1}\right\|^2 = (\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1}),\boldsymbol{\epsilon}(\boldsymbol{w}_p))+\lambda(\operatorname{div}\boldsymbol{e}_u^{m+1},\operatorname{div}\boldsymbol{w}_p). \tag{3.16}$$

Now, applying Cauchy's inequality, we obtain

$$\begin{aligned}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2 &\le (\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2+\lambda\|\operatorname{div}\boldsymbol{e}_u^{m+1}\|^2)^{\frac{1}{2}}(\|\boldsymbol{\epsilon}(\boldsymbol{w}_p)\|^2+\lambda\|\operatorname{div}\boldsymbol{w}_p\|^2)^{\frac{1}{2}}\\ &\le (\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2+\lambda\|\operatorname{div}\boldsymbol{e}_u^{m+1}\|^2)^{\frac{1}{2}}(\beta_s^{-2}+\lambda)^{\frac{1}{2}}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|,\end{aligned} \tag{3.17}$$

which implies

$$(\beta_s^{-2}+\lambda)^{-1}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2 \le \|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2+\lambda\|\operatorname{div}\boldsymbol{e}_u^{m+1}\|^2. \tag{3.18}$$

Given Lemma 3.1 and (3.18), we therefore obtain

$$\begin{aligned}&\frac{1}{2}(\beta_s^{-2}+\lambda)^{-1}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2+\|R^{-1/2}\boldsymbol{e}_v^{m+1}\|^2+\|(\Lambda_1+\Lambda_2)^{1/2}\boldsymbol{e}_p^{m+1}\|^2\\ &\quad+\frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2 \le \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|^2\end{aligned}$$

and hence

$$\left(\frac{1}{2\beta_s^{-2}+2\lambda}+\frac{L}{2}\right)\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2 \le \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|^2,$$

or, equivalently,

$$\left(\frac{L^{-1}}{\beta_s^{-2}+\lambda}+1\right)\left\|\sum_{i=1}^{n} e_{p_i}^{m+1}\right\|^2 \le \left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|^2,$$

which proves (3.12)–(3.13). Finally, (3.14) follows from (3.13) by choosing $L=\frac{1}{\lambda+c_K^2}$ and noting that $\frac{1}{\frac{\lambda+c_K^2}{\beta_s^{-2}+\lambda}+1}$ is a monotone function for $\lambda>0$. ∎

Note that $\left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|$ only defines a seminorm of $\boldsymbol{e}_p^m$ and Theorem 3.2 indicates the convergence rate of $\boldsymbol{e}_p$ in this seminorm. It still remains at this point unclear whether $\left\|\sum_{i=1}^{n} e_{p_i}^{m}\right\|\to 0$ guarantees that $\boldsymbol{e}_p^m$ converges to $\mathbf{0}$.

Theorem 3.4, as stated later, clarifies this and demonstrates the uniform convergence of $\boldsymbol{e}_u^m, \boldsymbol{e}_v^m$, and $\boldsymbol{e}_p^m$ for the fixed-stress iterative method utilizing the uniform stability results from [24]. Before we present Theorem 3.4, we introduce the matrices

$$\Lambda_L := \begin{bmatrix} L & \dots & \dots & L\\ \vdots & & & \vdots\\ \vdots & & & \vdots\\ L & \dots & \dots & L\end{bmatrix} \quad\text{and}\quad \Lambda_e := \Lambda+\Lambda_L. \tag{3.19}$$

Analogous to the assertion of Lemma 1 in [24], the properties of Λ_e are as follows.

LEMMA 3.3. *Letting* $\tilde{\Lambda}=\Lambda_3+\Lambda_4+\Lambda_L, \tilde{\Lambda}^{-1}=(\tilde{b}_{ij})_{n\times n}$, *then* $\tilde{\Lambda}$ *is SPD and for any n-dimensional vector* $\boldsymbol{x}$, *and we have*

$$(\Lambda_e\boldsymbol{x},\boldsymbol{x}) \ge (\tilde{\Lambda}\boldsymbol{x},\boldsymbol{x}) \ge (\Lambda_3\boldsymbol{x},\boldsymbol{x}), \tag{3.20}$$

$$(\Lambda_e^{-1}\boldsymbol{x},\boldsymbol{x}) \le (\tilde{\Lambda}^{-1}\boldsymbol{x},\boldsymbol{x}) \le (\Lambda_3^{-1}\boldsymbol{x},\boldsymbol{x}) = R^{-1}(\boldsymbol{x},\boldsymbol{x}). \tag{3.21}$$

Also,

$$0 < \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{b}_{ij} \leq \left(\frac{1}{\lambda_0} + L\right)^{-1}. \tag{3.22}$$

Subsequently, we can use Λ_e to define the following parameter-dependent norms:

$$(\boldsymbol{u}, \boldsymbol{w})_{\boldsymbol{U}} = (\epsilon(\boldsymbol{u}), \epsilon(\boldsymbol{w})) + \lambda(\operatorname{div} \boldsymbol{u}, \operatorname{div} \boldsymbol{w}), \tag{3.23a}$$

$$(\boldsymbol{v}, \boldsymbol{z})_{\boldsymbol{V}_e} = \sum_{i=1}^{n} (R_i^{-1}\boldsymbol{v}_i, \boldsymbol{z}_i) + (\Lambda_e^{-1}\operatorname{Div} \boldsymbol{v}, \operatorname{Div} \boldsymbol{z}), \tag{3.23b}$$

$$(\boldsymbol{p}, \boldsymbol{q})_{\boldsymbol{P}_e} = (\Lambda_e \boldsymbol{p}, \boldsymbol{q}). \tag{3.23c}$$

As stated in the following theorem, the fixed-stress split iterative method for the MPET system converges uniformly.

THEOREM 3.4. *Consider the fixed-stress split iterative method as defined in Algorithm* 3.1 *and assume that* $L \geq 1/(\lambda + c_K^2)$. *Then the errors* $\boldsymbol{e}_u^m$, $\boldsymbol{e}_v^m$, *and* $\boldsymbol{e}_p^m$ *given in* (3.1), *measured in the norms induced by* (3.23), *satisfy the estimates*

$$\|\boldsymbol{e}_u^m\|_{\boldsymbol{U}} \leq C_u [\operatorname{rate}(\lambda)]^m \left\| \sum_{i=1}^{n} e_{p_i}^0 \right\|, \tag{3.24}$$

$$\|\boldsymbol{e}_v^m\|_{\boldsymbol{V}_e} + \|\boldsymbol{e}_p^m\|_{\boldsymbol{P}_e} \leq C_{vp} [\operatorname{rate}(\lambda)]^m \left\| \sum_{i=1}^{n} e_{p_i}^0 \right\|, \tag{3.25}$$

where the constants C_u *and* C_{vp} *are independent of the model parameters and the time step size* τ. *Furthermore, the convergence rate* $\operatorname{rate}(\lambda)$ *satisfies* (3.13).

Proof. In the same manner as we derived (3.11) we find

$$\|\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1})\|^2 + \lambda \|\operatorname{div} \boldsymbol{e}_u^{m+1}\|^2 \leq \left(\frac{1}{c_K^2 + \lambda}\right) \left\| \sum_{i=1}^{n} e_{p_i}^{m+1} \right\|^2,$$

which shows (3.24). Moreover, rewriting the error equations (3.2a)–(3.2c) and using the definition of Λ_L we deduce the variational problem

$$(\boldsymbol{\epsilon}(\boldsymbol{e}_u^{m+1}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div} \boldsymbol{e}_u^{m+1}, \operatorname{div} \boldsymbol{w}) - \left(\sum_{i=1}^{n} e_{p_i}^{m+1}, \operatorname{div} \boldsymbol{w}\right) = 0,$$

$$(R^{-1}\boldsymbol{e}_v^{m+1}, \boldsymbol{z}) - (\boldsymbol{e}_p^{m+1}, \operatorname{Div} \boldsymbol{z}) = 0,$$

$$\begin{aligned} &- \left(\operatorname{div} \boldsymbol{e}_u^{m+1}, \sum_{i=1}^{n} q_i\right) - (\operatorname{Div} \boldsymbol{e}_v^{m+1}, \boldsymbol{q}) - ((\Lambda_1 + \Lambda_2 + \Lambda_L)\boldsymbol{e}_p^{m+1}, \boldsymbol{q}) \\ &\quad = -L \left(\sum_{i=1}^{n} e_{p_i}^m, \sum_{i=1}^{n} q_i\right) + \left(\operatorname{div} \boldsymbol{e}_u^m - \operatorname{div} \boldsymbol{e}_u^{m+1}, \sum_{i=1}^{n} q_i\right). \end{aligned} \tag{3.26}$$

Denote

$$g_e = -L \sum_{i=1}^{n} e_{p_i}^m + \operatorname{div} \boldsymbol{e}_u^m - \operatorname{div} \boldsymbol{e}_u^{m+1};$$

then by the triangle inequality, (3.10), and the contraction estimate (3.12), it follows that

$$
\begin{aligned}
\|g_e\| &= \left\| -L \sum_{i=1}^{n} e_{p_i}^m + \operatorname{div} \boldsymbol{e}_u^m - \operatorname{div} \boldsymbol{e}_u^{m+1} \right\| \\
&\le L \left\| \sum_{i=1}^{n} e_{p_i}^m \right\| + \frac{1}{\lambda + c_K^2} \left\| \sum_{i=1}^{n} e_{p_i}^m - \sum_{i=1}^{n} e_{p_i}^{m+1} \right\| \\
&\le L \left\| \sum_{i=1}^{n} e_{p_i}^m \right\| + \frac{2}{\lambda + c_K^2} \left\| \sum_{i=1}^{n} e_{p_i}^m \right\| \\
&\le 3L \left\| \sum_{i=1}^{n} e_{p_i}^m \right\| \\
&\le 3L [\mathrm{rate}(\lambda)]^m \left\| \sum_{i=1}^{n} e_{p_i}^0 \right\|.
\end{aligned} \tag{3.27}
$$

Next, by taking $\boldsymbol{f} = \boldsymbol{0}$, $\boldsymbol{g} = (g_e, g_e, \ldots, g_e)^T$ and replacing $\Lambda_1 + \Lambda_2$ by $\Lambda_1 + \Lambda_2 + \Lambda_L$ in (2.6) and using the uniform stability estimate (2.11) with Λ replaced by Λ_e, we obtain

$$
\begin{aligned}
\|\boldsymbol{e}_u^{m+1}\|_{\boldsymbol{U}} + \|\boldsymbol{e}_v^{m+1}\|_{\boldsymbol{V}_e} + \|\boldsymbol{e}_p^{m+1}\|_{\boldsymbol{P}_e} &\le C_1 \|\boldsymbol{g}\|_{\boldsymbol{P}_e^*} \\
&= C_1 \|\Lambda_e^{-\frac{1}{2}} \boldsymbol{g}\| = C_1 (\Lambda_e^{-1} \boldsymbol{g}, \boldsymbol{g})^{\frac{1}{2}}.
\end{aligned} \tag{3.28}
$$

Further, by Lemma 3.3 and (3.27), we have

$$
\begin{aligned}
(\Lambda_e^{-1} \boldsymbol{g}, \boldsymbol{g}) \le (\tilde{\Lambda}^{-1} \boldsymbol{g}, \boldsymbol{g}) &= (\tilde{\Lambda}^{-1} (\underbrace{g_e, g_e \ldots, g_e}_{n})^T, (\underbrace{g_e, g_e \ldots, g_e}_{n})^T) \\
&= \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{b}_{ij} \right) (g_e, g_e) \le \left(\frac{1}{\lambda_0} + L \right)^{-1} (g_e, g_e) \\
&\le 9 \left(\frac{1}{\lambda_0} + L \right)^{-1} L^2 [\mathrm{rate}(\lambda)]^{2m} \left\| \sum_{i=1}^{n} e_{p_i}^0 \right\|^2 \\
&\le 9L [\mathrm{rate}(\lambda)]^{2m} \left\| \sum_{i=1}^{n} e_{p_i}^0 \right\|^2.
\end{aligned} \tag{3.29}
$$

Combining (3.28) and (3.29) then implies (3.24) and (3.25). ∎

4. Discrete MPET problem. In this section, mass conservative discretizations of the MPET model are discussed; cf. [23, 24]. The analysis here can be similarly used for other stable discretizations of the three-field formulation of Biot's consolidation or the MPET model; see, e.g., [26, 36].

4.1. Notation. We consider a shape-regular triangulation $\mathcal{T}_h$ of Ω into triangles/tetrahedrons. Here, the subscript h indicates the mesh size. The set of all interior edges/faces and the set of all boundary edges/faces of $\mathcal{T}_h$ are denoted by $\mathcal{E}_h^I$ and $\mathcal{E}_h^B$, respectively, and their union by $\mathcal{E}_h$.

We define the broken Sobolev spaces

$$
H^s(\mathcal{T}_h) = \{\phi \in L^2(\Omega), \text{ such that } \phi|_T \in H^s(T) \text{ for all } T \in \mathcal{T}_h\}
$$

for $s \ge 1$.

We next introduce the notion of jumps $[\cdot]$ and averages $\{\cdot\}$. Let T_1 and T_2 be two elements from the triangulation sharing an edge or face e and let $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ be the corresponding unit normal vectors to e pointing to the exterior of T_1 and T_2. Then for $q \in H^1(\mathcal{T}_h)$, $\boldsymbol{v} \in H^1(\mathcal{T}_h)^d$, and $\boldsymbol{\tau} \in H^1(\mathcal{T}_h)^{d\times d}$ and any $e \in \mathcal{E}_h^I$ we define

$$[q] = q|_{\partial T_1 \cap e} - q|_{\partial T_2 \cap e}, \quad [\boldsymbol{v}] = \boldsymbol{v}|_{\partial T_1 \cap e} - \boldsymbol{v}|_{\partial T_2 \cap e}$$

and

$$\{\boldsymbol{v}\} = \frac{1}{2}(\boldsymbol{v}|_{\partial T_1 \cap e} \cdot \boldsymbol{n}_1 - \boldsymbol{v}|_{\partial T_2 \cap e} \cdot \boldsymbol{n}_2), \quad \{\boldsymbol{\tau}\} = \frac{1}{2}(\boldsymbol{\tau}|_{\partial T_1 \cap e}\boldsymbol{n}_1 - \boldsymbol{\tau}|_{\partial T_2 \cap e}\boldsymbol{n}_2),$$

while for $e \in \mathcal{E}_h^B$,

$$[q] = q|_e, \quad [\boldsymbol{v}] = \boldsymbol{v}|_e, \quad \{\boldsymbol{v}\} = \boldsymbol{v}|_e \cdot \boldsymbol{n}, \quad \{\boldsymbol{\tau}\} = \boldsymbol{\tau}|_e \boldsymbol{n}.$$

4.2. Mixed finite element spaces and discrete formulation. In order to discretize the flow equations, we use a mixed finite element method to approximate the fluxes and pressures, whereas for the mechanics problem we apply a discontinuous Galerkin method to approximate the displacement. The considered finite element spaces are denoted by

$$\begin{aligned}
\boldsymbol{U}_h &= \{\boldsymbol{u} \in H(\mathrm{div};\Omega) : \boldsymbol{u}|_T \in \boldsymbol{U}(T),\ T \in \mathcal{T}_h;\ \boldsymbol{u}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\},\\
\boldsymbol{V}_{i,h} &= \{\boldsymbol{v} \in H(\mathrm{div};\Omega) : \boldsymbol{v}|_T \in \boldsymbol{V}_i(T),\ T \in \mathcal{T}_h;\ \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\}, \quad i = 1,\dots,n,\\
P_{i,h} &= \left\{q \in L^2(\Omega) : q|_T \in Q_i(T),\ T \in \mathcal{T}_h;\ \int_\Omega q dx = 0\right\}, \quad i = 1,\dots,n,
\end{aligned}$$

where $\boldsymbol{V}_i(T)/Q_i(T) = \mathrm{RT}_{l-1}(T)/\mathrm{P}_{l-1}(T)$, $\boldsymbol{U}(T) = \mathrm{BDM}_l(T)$, or $\boldsymbol{U}(T) = \mathrm{BDFM}_l(T)$ for $l \geq 1$. For each of these choices, we would like to point out that $\operatorname{div}\boldsymbol{U}(T) = \operatorname{div}\boldsymbol{V}_i(T) = Q_i(T)$ is satisfied.

As pointed out also in [23, 24], for all $\boldsymbol{u} \in \boldsymbol{U}_h$ it holds that

$$[\boldsymbol{u}_n] = 0, \quad \text{from which it follows} \quad [\boldsymbol{u}] = [\boldsymbol{u}_t], \tag{4.1}$$

where $\boldsymbol{u}_n$ and $\boldsymbol{u}_t$ denote the normal and tangential components of $\boldsymbol{u}$, respectively.

Using the notation

$$\begin{gathered}
\boldsymbol{v}_h^T = (\boldsymbol{v}_{1,h}^T, \dots, \boldsymbol{v}_{n,h}^T), \qquad \boldsymbol{p}_h^T = (p_{1,h}, \dots, p_{n,h}),\\
\boldsymbol{z}_h^T = (\boldsymbol{z}_{1,h}^T, \dots, \boldsymbol{z}_{n,h}^T), \qquad \boldsymbol{q}_h^T = (q_{1,h}, \dots, q_{n,h}),\\
\boldsymbol{V}_h = \boldsymbol{V}_{1,h} \times \cdots \times \boldsymbol{V}_{n,h}, \quad \boldsymbol{P}_h = P_{1,h} \times \cdots \times P_{n,h}, \quad \boldsymbol{X}_h = \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h,
\end{gathered}$$

the discretization of problem (2.6) can be expressed as: find $(\boldsymbol{u}_h; \boldsymbol{v}_h; \boldsymbol{p}_h) \in \boldsymbol{X}_h$, such that for any $(\boldsymbol{w}_h; \boldsymbol{z}_h; \boldsymbol{q}_h) \in \boldsymbol{X}_h$ and $i = 1, \dots, n$

$$a_h(\boldsymbol{u}_h, \boldsymbol{w}_h) + \lambda(\operatorname{div}\boldsymbol{u}_h, \operatorname{div}\boldsymbol{w}_h) - \sum_{i=1}^n (p_{i,h}, \operatorname{div}\boldsymbol{w}_h) = (\boldsymbol{f}, \boldsymbol{w}_h), \tag{4.2a}$$

$$(R_i^{-1}\boldsymbol{v}_{i,h}, \boldsymbol{z}_{i,h}) - (p_{i,h}, \operatorname{div}\boldsymbol{z}_{i,h}) = 0, \tag{4.2b}$$

$$-(\operatorname{div}\boldsymbol{u}_h, q_{i,h}) - (\operatorname{div}\boldsymbol{v}_{i,h}, q_{i.h}) + \tilde{\alpha}_{ii}(p_{i,h}, q_{i,h}) + \sum_{\substack{j=1\\ j\neq i}}^n \alpha_{ij}(p_{j,h}, q_{i,h}) = (g_i, q_{i,h}), \tag{4.2c}$$

where

$$a_h(\boldsymbol{u},\boldsymbol{w}) = \sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{\epsilon}(\boldsymbol{u}):\boldsymbol{\epsilon}(\boldsymbol{w})dx - \sum_{e\in\mathcal{E}_h}\int_e \{\boldsymbol{\epsilon}(\boldsymbol{u})\}\cdot[\boldsymbol{w}_t]ds \\ - \sum_{e\in\mathcal{E}_h}\int_e \{\boldsymbol{\epsilon}(\boldsymbol{w})\}\cdot[\boldsymbol{u}_t]ds + \sum_{e\in\mathcal{E}_h}\int_e \eta h_e^{-1}[\boldsymbol{u}_t]\cdot[\boldsymbol{w}_t]ds, \tag{4.3}$$

$\tilde{\alpha}_{ii} = -\alpha_{p_i} - \alpha_{ii}$, and η is a stabilization parameter independent of the parameters λ, R_i^{-1}, α_{p_i}, α_{ij}, where $i,j\in\{1,\dots,n\}$, the network scale n, and the mesh size h.

The discrete variational problem (4.2) corresponds to the weak formulation (2.6) with homogeneous boundary conditions. The DG discretizations for general rescaled boundary conditions can be found in [24, 23].

4.3. Stability properties. Let $\boldsymbol{u}$ be a function from $\boldsymbol{U}_h$ and consider the mesh dependent norms

$$\|\boldsymbol{u}\|_h^2 = \sum_{K\in\mathcal{T}_h}\|\boldsymbol{\epsilon}(\boldsymbol{u})\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\boldsymbol{u}_t]\|_{0,e}^2,$$
$$\|\boldsymbol{u}\|_{1,h}^2 = \sum_{K\in\mathcal{T}_h}\|\nabla\boldsymbol{u}\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\boldsymbol{u}_t]\|_{0,e}^2,$$

$$\|\boldsymbol{u}\|_{DG}^2 = \sum_{K\in\mathcal{T}_h}\|\nabla\boldsymbol{u}\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\boldsymbol{u}_t]\|_{0,e}^2 + \sum_{K\in\mathcal{T}_h} h_K^2|\boldsymbol{u}|_{2,K}^2, \tag{4.4}$$

and

$$\|\boldsymbol{u}\|_{\boldsymbol{U}_h}^2 = \|\boldsymbol{u}\|_{DG}^2 + \lambda\|\operatorname{div}\boldsymbol{u}\|^2. \tag{4.5}$$

The well-posedness and approximation properties of the DG formulation are detailed in [25, 22]. Here we briefly present some important results:

- $\|\cdot\|_{DG}$, $\|\cdot\|_h$, and $\|\cdot\|_{1,h}$ are equivalent on $\boldsymbol{U}_h$; that is,

$$\|\boldsymbol{u}\|_{DG} \eqsim \|\boldsymbol{u}\|_h \eqsim \|\boldsymbol{u}\|_{1,h} \quad \text{for all } \boldsymbol{u}\in\boldsymbol{U}_h.$$

- $a_h(\cdot,\cdot)$ from (4.3) is continuous and it holds true that

$$|a_h(\boldsymbol{u},\boldsymbol{w})| \lesssim \|\boldsymbol{u}\|_{DG}\|\boldsymbol{w}\|_{DG} \quad \text{for all} \quad \boldsymbol{u},\ \boldsymbol{w}\in H^2(\mathcal{T}_h)^d. \tag{4.6}$$

- The inf-sup conditions

$$\inf_{(q_{1,h},\cdots,q_{n,h})\in P_{1,h}\times\cdots\times P_{n,h}}\ \sup_{\boldsymbol{u}_h\in\boldsymbol{U}_h}\ \frac{(\operatorname{div}\boldsymbol{u}_h, \sum_{i=1}^n q_{i,h})}{\|\boldsymbol{u}_h\|_{1,h}\|\sum_{i=1}^n q_{i,h}\|} \geq \beta_{sd}, \\ \inf_{q_{i,h}\in P_{i,h}}\ \sup_{\boldsymbol{v}_{i,h}\in\boldsymbol{V}_{i,h}}\ \frac{(\operatorname{div}\boldsymbol{v}_{i,h}, q_{i,h})}{\|\boldsymbol{v}_{i,h}\|_{\operatorname{div}}\|q_{i,h}\|} \geq \beta_{dd}, \quad i=1,\dots,n, \tag{4.7}$$

  are valid for our choice of $\boldsymbol{U}_h$, $\boldsymbol{V}_h$, and $\boldsymbol{P}_h$ (see [37]), and the positive constants β_{sd} and β_{dd} are independent of λ, R_i^{-1}, α_{p_i}, α_{ij} for $i,j\in\{1,\dots,n\}$, the network scale n, and the mesh size h.

- $a_h(\cdot,\cdot)$ is coercive, namely

$$a_h(\boldsymbol{u}_h,\boldsymbol{u}_h) \geq \alpha_a \|\boldsymbol{u}_h\|_h^2 \quad \text{for all} \quad \boldsymbol{u}_h \in \boldsymbol{U}_h, \tag{4.8}$$

  where $\alpha_a > 0$ is a constant independent of the model and discretization parameters $\lambda, R_i^{-1}, \alpha_{p_i}, \alpha_{ij}, i,j \in \{1,\dots,n\}$, n, and h.

Using the definition of the matrices Λ_1 and Λ_2, we define the bilinear form

$$\begin{aligned}\mathcal{A}_h((\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h),(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h)) &= a_h(\boldsymbol{u}_h,\boldsymbol{w}_h) + \lambda(\operatorname{div}\boldsymbol{u}_h,\operatorname{div}\boldsymbol{w}_h)\\ &- \sum_{i=1}^{n}(p_{i,h},\operatorname{div}\boldsymbol{w}_h) + \sum_{i=1}^{n}(R_i^{-1}\boldsymbol{v}_{i,h},\boldsymbol{z}_{i,h}) - (\boldsymbol{p}_h,\operatorname{Div}\boldsymbol{z}_h)\\ &- \left(\operatorname{div}\boldsymbol{u}_h,\sum_{i=1}^{n}q_{i,h}\right) - (\operatorname{Div}\boldsymbol{v}_h,\boldsymbol{q}_h) - ((\Lambda_1+\Lambda_2)\boldsymbol{p}_h,\boldsymbol{q}_h)\end{aligned} \tag{4.9}$$

related to problem (4.2a)–(4.2c).

Let the space $\boldsymbol{X}_h$ be equipped with the norm $\|(\cdot;\cdot;\cdot)\|_{\boldsymbol{X}_h} := \|\cdot\|_{\boldsymbol{U}_h} + \|\cdot\|_{\boldsymbol{V}} + \|\cdot\|_{\boldsymbol{P}}$. Similar to Theorem 2.1, the following uniform stability results can be found as in [24].

THEOREM 4.1.

(i) *For any* $(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h) \in \boldsymbol{X}_h, (\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h) \in \boldsymbol{X}_h$ *there exists a positive constant* C_{bd} *independent of the parameters* λ, R_i^{-1}, α_{p_i}, α_{ij}, $i,j \in \{1,\dots,n\}$, *the network scale* n, *and the mesh size* h *such that the inequality*

$$|\mathcal{A}_h((\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h),(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h))| \leq C_{bd}\|(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h)\|_{\boldsymbol{X}_h}\|(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h)\|_{\boldsymbol{X}_h}$$

holds true.

(ii) *There exists a constant* $\beta_0 > 0$ *independent of the model and discretization parameters* λ, R_i^{-1}, α_{p_i}, α_{ij}, $i,j \in \{1,\dots,n\}$, n, *and* h, *such that*

$$\inf_{(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h)\in\boldsymbol{X}_h}\sup_{(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h)\in\boldsymbol{X}_h}\frac{\mathcal{A}_h((\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h),(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h))}{\|(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h)\|_{\boldsymbol{X}_h}\|(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h)\|_{\boldsymbol{X}_h}} \geq \beta_0. \tag{4.10}$$

(iii) *Let* $(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h) \in \boldsymbol{X}_h$ *solve* (4.2a)–(4.2c) *and*

$$\|\boldsymbol{f}\|_{\boldsymbol{U}_h^*} = \sup_{\boldsymbol{w}_h\in\boldsymbol{U}_h}\frac{(\boldsymbol{f},\boldsymbol{w}_h)}{\|\boldsymbol{w}_h\|_{\boldsymbol{U}_h}}, \quad \|\boldsymbol{g}\|_{\boldsymbol{P}^*} = \sup_{\boldsymbol{q}_h\in\boldsymbol{P}_h}\frac{(\boldsymbol{g},\boldsymbol{q}_h)}{\|\boldsymbol{q}_h\|_{\boldsymbol{P}}}.$$

Then the estimate

$$\|\boldsymbol{u}_h\|_{\boldsymbol{U}_h} + \|\boldsymbol{v}_h\|_{\boldsymbol{V}} + \|\boldsymbol{p}_h\|_{\boldsymbol{P}} \leq C_2(\|\boldsymbol{f}\|_{\boldsymbol{U}_h^*} + \|\boldsymbol{g}\|_{\boldsymbol{P}^*}) \tag{4.11}$$

holds with a constant C_2 *independent of the network scale* n, *the mesh size* h, *and the parameters* λ, R_i^{-1}, α_{p_i}, α_{ij}, $i,j \in \{1,\dots,n\}$.

5. Fixed-stress method for the discrete MPET model. In the manner of Algorithm 3.1, we formulate the fixed-stress method for the mixed continuous-discontinuous Galerkin finite element method (4.2):

Algorithm 5.1: Fixed-stress method for the discrete MPET problem

Step a: Given $\boldsymbol{u}_h^m$, we solve for $\boldsymbol{v}_{i,h}^{m+1}$ and $p_{i,h}^{m+1}$

$$(-\operatorname{div}\boldsymbol{v}_{i,h}^{m+1}, q_{i,h}) - ((\alpha_{p_{i,h}} + \alpha_{ii})p_{i,h}^{m+1}, q_{i,h}) + \left(\sum_{\substack{j=1\\ j\neq i}}^{n} \alpha_{ij} p_{j,h}^{m+1}, q_{i,h}\right)$$
$$-L\left(\sum_{j=1}^{n} p_{j,h}^{m+1}, q_{i,h}\right) = (g_i, q_{i,h}) - L\left(\sum_{j=1}^{n} p_{j,h}^{m}, q_{i,h}\right) + (\operatorname{div}\boldsymbol{u}^m, q_{i,h}), \quad 1 \le i \le n,$$

and

$$(R_i^{-1}\boldsymbol{v}_{i,h}^{m+1}, \boldsymbol{z}_{i,h}) - (p_{i,h}^{m+1}, \operatorname{div}\boldsymbol{z}_{i,h}) = 0, \quad 1 \le i \le n.$$

Step b: Given $\boldsymbol{v}_{i,h}^{m+1}$ and $p_{i,h}^{m+1}$, we solve for $\boldsymbol{u}_h^{m+1}$

$$a_h(\boldsymbol{u}_h^{m+1}, \boldsymbol{w}_h) + \lambda(\operatorname{div}\boldsymbol{u}_h^{m+1}, \operatorname{div}\boldsymbol{w}_h) = (\boldsymbol{f}, \boldsymbol{w}_h) + \sum_{i=1}^{n}(p_{i,h}^{m+1}, \operatorname{div}\boldsymbol{w}_h).$$

The main convergence result for Algorithm 5.1 is formulated in terms of the following quantities corresponding to the discrete case:

$$\boldsymbol{e}_{u_h}^k = \boldsymbol{u}_h^k - \boldsymbol{u}_h \in \boldsymbol{U}_h, \tag{5.1a}$$

$$\boldsymbol{e}_{v_{i,h}}^k = \boldsymbol{v}_{i,h}^k - \boldsymbol{v}_{i,h} \in \boldsymbol{V}_{i,h}, \quad i = 1, \dots, n, \tag{5.1b}$$

$$e_{p_{i,h}}^k = p_{i,h}^k - p_{i,h} \in P_{i,h}, \quad i = 1, \dots, n, \tag{5.1c}$$

denoting the errors of the kth iterates $\boldsymbol{u}_h^k$, $\boldsymbol{v}_{i,h}^k$, $p_{i,h}^k$, $i = 1, \dots, n$, generated by Algrorithm 5.1. In the discrete case, the useful constant for defining the tuning parameter L is the constant c_{K_d} from the estimate

$$a_h(\boldsymbol{w}_h, \boldsymbol{w}_h) \ge c_{K_d}^2 \|\operatorname{div}\boldsymbol{w}_h\|^2 \quad \text{for all } \boldsymbol{w}_h \in \boldsymbol{U}_h. \tag{5.2}$$

Note that c_{K_d} is strictly positive and independent of the mesh size h.

Using the approach applied to proving Lemma 3.1, for the continuous MPET model we obtain the corresponding lemma for the discrete case as follows.

LEMMA 5.1. *The errors $\boldsymbol{e}_{u_h}^{m+1}$, $\boldsymbol{e}_{v_h}^{m+1}$, and $\boldsymbol{e}_{p_h}^{m+1}$ of the $(m+1)$st fixed-stress iterate generated by Algorithm* 5.1 *for $L \ge \frac{1}{\lambda + c_{K_d}^2}$ satisfy the estimate*

$$\frac{1}{2}\Big(a_h(\boldsymbol{e}_{u_h}^{m+1}, \boldsymbol{e}_{u_h}^{m+1}) + \lambda\|\operatorname{div}\boldsymbol{e}_{u_h}^{m+1}\|^2\Big) + \|R^{-1/2}\boldsymbol{e}_{v_h}^{m+1}\|^2 + \|(\Lambda_1 + \Lambda_2)^{1/2}\boldsymbol{e}_{p_h}^{m+1}\|^2$$
$$+\frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_{i,h}}^{m+1}\right\|^2 \le \frac{L}{2}\left\|\sum_{i=1}^{n} e_{p_{i,h}}^{m}\right\|^2, \quad m = 0, 1, 2, \dots.$$

By Lemma 5.1, again following the proof of Theorem 3.2 for the continuous MPET model, we obtain the corresponding statements, Theorem 5.2, for the discrete case.

THEOREM 5.2. *Let c_{K_d} and β_{sd} denote the constants in* (5.2) *and* (4.7)*, respectively. The single rate fixed-stress iterative method for the discrete static MPET problem* (4.2) *defined in Algorithm* 5.1 *is a contraction that converges linearly for any*

$L \geq 1/(\lambda + c_{K_d}^2)$ independent of the model parameters, the time step size τ, and the mesh size h. The errors $\boldsymbol{e}_{p_h}^m$ in this case satisfy the inequality

$$\left\| \sum_{i=1}^{n} e_{p_{i,h}}^{m+1} \right\|^2 \leq \mathrm{rate}_d{}^2(\lambda) \left\| \sum_{i=1}^{n} e_{p_{i,h}}^{m} \right\|^2, \tag{5.3}$$

where

$$\mathrm{rate}_d{}^2(\lambda) \leq \frac{1}{\frac{L^{-1}}{\beta_{sd}^{-2}+\lambda} + 1}. \tag{5.4}$$

For $L = 1/(\lambda + c_{K_d}^2)$, the convergence factor in (5.3) *can be estimated by*

$$\mathrm{rate}_d{}^2(\lambda) \leq \frac{1}{\frac{\lambda + c_{K_d}^2}{\beta_{sd}^{-2}+\lambda} + 1} \leq \max\left\{ \frac{\beta_{sd}^{-2}}{c_{K_d}^2 + \beta_{sd}^{-2}}, \frac{1}{2} \right\}. \tag{5.5}$$

Note that Theorem 5.2 only gives the convergence rate of $\boldsymbol{e}_{p_h}^m$ in the seminorm $\| \sum_{i=1}^{n} e_{p_{i,h}}^m \|$. However, we can combine the estimates in Theorem 5.2 with the uniform stability result presented in Theorem 4.1 and follow the proof of Theorem 3.4 to obtain the following convergence results for $\boldsymbol{e}_{u_h}^m, \boldsymbol{e}_{v_h}^m$, and $\boldsymbol{e}_{p_h}^m$ in their respective parameter-dependent full norms.

THEOREM 5.3. *The errors $\boldsymbol{e}_{u_h}^m$, $\boldsymbol{e}_{v_h}^m$, and $\boldsymbol{e}_{p_h}^m$ defined in* (5.1) *measured in the norms induced by* (4.5) *and* (2.7) *satisfy the estimates*

$$\|\boldsymbol{e}_{u_h}^m\|_{\boldsymbol{U}_h}^2 \leq C_{ud}[\mathrm{rate}_d(\lambda)]^{2m}, \tag{5.6}$$

$$\|\boldsymbol{e}_{v_h}^m\|_{\boldsymbol{V}_e}^2 + \|\boldsymbol{e}_{p_h}^m\|_{\boldsymbol{P}_e}^2 \leq C_{vpd}[\mathrm{rate}_d(\lambda)]^{2m}, \tag{5.7}$$

where the constants C_{ud} and C_{vpd} are independent of the model parameters, the time step size, and the mesh size.

6. Numerical results. In our numerical test setup, we assume that

- $\Omega = [0, 1]$ is partitioned into $2N^2$ right-angled triangles with catheti of length $h = 1/N$;
- problem (2.6) is discretized by a strongly conservative discontinuous Galerkin method which is based on a mixed finite element space formed by the triplet of $\mathrm{BDM}_1/\mathrm{RT}_0/\mathrm{P}_0^{dc}$ elements;
- the iterative process is terminated when residual reduction by a factor 10^8 in the combined norm induced by the inner products (2.7) (the norm induced by the inverse of the preconditioner) is reached.

Numerical experiments have been performed in FEniCS, [4, 32], and their aim was

(i) to validate the theoretical estimates for the convergence of the fixed-stress split iterative method;
(ii) to compare the performance of the latter with the preconditioned MinRes algorithm using the norm-equivalent preconditioner proposed in [24] and also with the preconditioned GMRES algorithm using the FOV-equivalent preconditioner as defined in (2.16).

In all numerical tests we have used a direct solver based on LU decomposition to solve the velocity-pressure and the displacement problems arising in the fixed-stress split iteration. The stabilization parameter has been chosen to be $L = 1/(1 + \lambda)$ in all test cases except the one presented in Table 5, where the performance of the fixed-stress algorithm is compared for different values of L.

6.1. The two-network model. The Biot–Barenblatt model involves two pressures and two fluxes. In our notation, it has the following formulation:

$$-\mathrm{div}(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I}) = \boldsymbol{f}, \tag{6.1a}$$

$$R_i^{-1}\boldsymbol{v}_i + \nabla p_i = 0, \quad i = 1, 2, \tag{6.1b}$$

$$-\mathrm{div}\boldsymbol{u} - \mathrm{div}\boldsymbol{v}_i - \alpha_{p_i} p_i + \sum_{\substack{j=1 \\ j\neq i}}^{2} \alpha_{ij} p_j = g_i, \quad i = 1, 2. \tag{6.1c}$$

Specifically, the subject of numerical study in this subsection is the cantilever bracket benchmark problem (see [19]), for which $\boldsymbol{f} = \boldsymbol{0}$, $g_1 = g_2 = 0$. The boundary Γ of the domain $\Omega = [0, 1]^2$ is split into bottom, right, top, and left boundaries denoted by Γ_1, Γ_2, Γ_3, and Γ_4, respectively, and

$$\begin{aligned}
(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I})\boldsymbol{n} &= (0, 0)^T && \text{on } \Gamma_1 \cup \Gamma_2, \\
(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I})\boldsymbol{n} &= (0, -1)^T && \text{on } \Gamma_3, \\
\boldsymbol{u} &= \boldsymbol{0} && \text{on } \Gamma_4, \\
p_1 &= 2 && \text{on } \Gamma, \\
p_2 &= 20 && \text{on } \Gamma.
\end{aligned}$$

Table 1 gives the base values of the model parameters as taken from [28]. We have varied the parameter K_2 over a wider range than K_1 since, at least, for the MinRes iteration it happened to be the more interesting case. The results in Tables 2–4 show very clearly the robust behavior of the fixed-stress split iteration with respect to mesh refinements and variation of the hydraulic conductivities K_1 and K_2, and also λ. Furthermore, they demonstrate its advantage over the MinRes and GMRES methods with regard to the rate of convergence.

The purpose of Table 5 is to illustrate the convergence behavior of the fixed-stress split method for different choices of the parameter L. We have run the algorithm with $L = L_1 := 1/(0.1 + \lambda)$, $L = L_2 := 1/(0.5 + \lambda)$, $L = L_3 := 1/(1 + \lambda)$, and $L = L_4 := 1/(10 + \lambda)$. We should mention again that estimate (3.3) is valid for $c_K^2 = 0.5$ in two dimensions. In all tests, we have observed that the fixed-stress scheme diverges for $L = L_4$ and we do not report this in Table 5. Moreover, the computational results coincide with the theory presented in Lemma 5.1 and Theorems 5.2–5.3; note also that the latter were proven for the discrete constant c_{K_d}.

In Table 6, we list the elapsed times in seconds for the iterative solution of the matrix equation arising from (4.2) in one time step of the implicit Euler method

TABLE 1
Base values of model parameters for a Barenblatt model.

Parameter	Value	Unit
λ_0	4.2	MPa
μ	2.4	MPa
c_{p_1}	54	$(\mathrm{GPa})^{-1}$
c_{p_2}	14	$(\mathrm{GPa})^{-1}$
α_1	0.95	
α_2	0.12	
β	5	$10^{-10}\mathrm{kg/(m{\cdot}s)}$
	100	$10^{-10}\mathrm{kg/(m{\cdot}s)}$
K_1	6.18	$10^{-15}\mathrm{m}^2$
K_2	27.2	$10^{-15}\mathrm{m}^2$

TABLE 2
Number of preconditioned MinRes (n_M), preconditioned GMRES (n_G), and fixed-stress split (n_F) iterations: parameters from Table 1, $\lambda = \lambda_0$.

h	β		K_2			$K_2 \cdot 10^2$			$K_2 \cdot 10^4$			$K_2 \cdot 10^6$		
			n_M	n_G	n_F	n_M	n_G	n_F	n_M	n_G	n_F	n_M	n_G	n_F
$\frac{1}{16}$	5E–10	$K_1 \cdot 10^{-2}$	16	9	8	21	10	8	37	12	8	29	12	8
		$K_1 \cdot 10^{-1}$	16	9	8	21	10	8	37	12	8	29	12	8
		K_1	16	9	8	21	10	8	37	12	8	29	12	8
	1E-8	$K_1 \cdot 10^{-2}$	16	9	8	21	10	8	37	12	8	29	12	8
		$K_1 \cdot 10^{-1}$	16	9	8	21	10	8	37	12	8	29	12	8
		K_1	16	9	8	21	10	8	37	12	8	29	12	8
$\frac{1}{32}$	5E–10	$K_1 \cdot 10^{-2}$	16	9	8	26	10	8	38	11	8	27	11	8
		$K_1 \cdot 10^{-1}$	16	9	8	26	10	8	38	11	8	27	11	8
		K_1	16	9	8	26	10	8	38	11	8	27	11	8
	1E-8	$K_1 \cdot 10^{-2}$	16	9	8	26	10	8	38	11	8	27	11	8
		$K_1 \cdot 10^{-1}$	16	9	8	26	10	8	38	11	8	27	11	8
		K_1	16	9	8	26	10	8	38	11	8	27	11	8
$\frac{1}{64}$	5E–10	$K_1 \cdot 10^{-2}$	18	9	8	32	10	8	38	11	8	27	11	8
		$K_1 \cdot 10^{-1}$	18	9	8	32	10	8	38	11	8	27	11	8
		K_1	18	9	8	32	10	8	38	11	8	27	11	8
	1E-8	$K_1 \cdot 10^{-2}$	18	9	8	32	10	8	38	11	8	27	11	8
		$K_1 \cdot 10^{-1}$	18	9	8	32	10	8	38	11	8	27	11	8
		K_1	18	9	8	32	10	8	38	11	8	27	11	8

TABLE 3
Number of preconditioned MinRes (n_M), preconditioned GMRES (n_G), and fixed-stress split (n_F) iterations: parameters from Table 1, except $\lambda := 10^{-2} \cdot \lambda_0$.

h	β		K_2			$K_2 \cdot 10^2$			$K_2 \cdot 10^4$			$K_2 \cdot 10^6$		
			n_M	n_G	n_F	n_M	n_G	n_F	n_M	n_G	n_F	n_M	n_G	n_F
$\frac{1}{16}$	5E–10	$K_1 \cdot 10^{-2}$	24	13	11	38	15	11	71	16	11	42	15	11
		$K_1 \cdot 10^{-1}$	24	13	11	38	15	11	71	16	11	42	15	11
		K_1	24	13	11	38	15	11	71	16	11	42	15	11
	1E-8	$K_1 \cdot 10^{-2}$	24	13	11	38	15	11	71	16	11	42	15	11
		$K_1 \cdot 10^{-1}$	24	13	11	38	15	11	71	16	11	42	15	11
		K_1	24	13	11	38	15	11	71	16	11	42	15	11
$\frac{1}{32}$	5E–10	$K_1 \cdot 10^{-2}$	25	14	10	45	15	10	66	16	10	38	15	10
		$K_1 \cdot 10^{-1}$	25	14	10	45	15	10	66	16	10	38	15	10
		K_1	25	14	10	45	15	10	66	16	10	38	15	10
	1E-8	$K_1 \cdot 10^{-2}$	25	14	10	45	15	10	66	16	10	38	15	10
		$K_1 \cdot 10^{-1}$	25	14	10	45	15	10	66	16	10	38	15	10
		K_1	25	14	10	45	15	10	66	16	10	38	15	10
$\frac{1}{64}$	5E–10	$K_1 \cdot 10^{-2}$	25	14	10	57	15	10	66	16	10	38	15	10
		$K_1 \cdot 10^{-1}$	25	14	10	57	15	10	66	16	10	38	15	10
		K_1	25	14	10	57	15	10	66	16	10	38	15	10
	1E-8	$K_1 \cdot 10^{-2}$	25	14	10	57	15	10	66	16	10	38	15	10
		$K_1 \cdot 10^{-1}$	25	14	10	57	15	10	66	16	10	38	15	10
		K_1	25	14	10	57	15	10	66	16	10	38	15	10

using the preconditioned MinRes, GMRES, and fixed-stress split methods. The time for the setup phase is excluded since matrix factorizations typically are reused many times when solving a quasi-static problem. Over the tested parameter sets the fixed-stress method reaches the stopping criterion faster than preconditioned MinRes and performes similarly to the GMRES algorithm with FOV-equivalent preconditioning.

TABLE 4

Number of preconditioned MinRes (n_M), preconditioned GMRES (n_G), and fixed-stress split (n_F) iterations: parameters from Table 1, except $\lambda := 10^2 \cdot \lambda_0$.

h	β		K_2			$K_2 \cdot 10^2$			$K_2 \cdot 10^4$			$K_2 \cdot 10^6$		
			n_M	n_G	n_F	n_M	n_G	n_F	n_M	n_G	n_F	n_M	n_G	n_F
$\frac{1}{16}$	5E–10	$K_1 \cdot 10^{-2}$	4	4	2	8	5	2	16	5	2	14	5	2
		$K_1 \cdot 10^{-1}$	4	4	2	8	5	2	16	5	2	14	5	2
		K_1	4	4	2	8	5	2	16	5	2	14	5	2
	1E-8	$K_1 \cdot 10^{-2}$	4	4	2	8	5	2	16	5	2	14	5	2
		$K_1 \cdot 10^{-1}$	4	4	2	8	5	2	16	5	2	14	5	2
		K_1	4	4	2	8	5	2	16	5	2	14	5	2
$\frac{1}{32}$	5E–10	$K_1 \cdot 10^{-2}$	6	3	2	12	5	2	20	5	2	14	5	2
		$K_1 \cdot 10^{-1}$	6	3	2	12	5	2	20	5	2	14	5	2
		K_1	6	3	2	12	5	2	20	5	2	14	5	2
	1E-8	$K_1 \cdot 10^{-2}$	6	3	2	12	5	2	20	5	2	14	5	2
		$K_1 \cdot 10^{-1}$	6	3	2	12	5	2	20	5	2	14	5	2
		K_1	6	3	2	12	5	2	20	5	2	14	5	2
$\frac{1}{64}$	5E–10	$K_1 \cdot 10^{-2}$	7	3	2	16	5	2	21	5	2	14	5	2
		$K_1 \cdot 10^{-1}$	7	3	2	16	5	2	21	5	2	14	5	2
		K_1	7	3	2	16	5	2	21	5	2	14	5	2
	1E-8	$K_1 \cdot 10^{-2}$	7	3	2	16	5	2	21	5	2	14	5	2
		$K_1 \cdot 10^{-1}$	7	3	2	16	5	2	21	5	2	14	5	2
		K_1	7	3	2	16	5	2	21	5	2	14	5	2

TABLE 5

Number of fixed-stress split iterations for Barenblatt problem: parameters from Table 1, $L = L_1 := 1/(0.1 + \lambda)$, $L = L_2 := 1/(0.5 + \lambda)$, and $L = L_3 := 1/(1 + \lambda)$.

h	β		K_2			$K_2 \cdot 10^2$			$K_2 \cdot 10^4$			$K_2 \cdot 10^6$		
			L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3
$\frac{1}{16}$	5E–10	$K_1 \cdot 10^{-2}$	20	13	8	20	13	8	20	13	8	19	13	8
		$K_1 \cdot 10^{-1}$	20	13	8	20	13	8	20	13	8	19	13	8
		K_1	20	13	8	20	13	8	20	13	8	19	13	8
	1E-8	$K_1 \cdot 10^{-2}$	20	13	8	20	13	8	20	13	8	19	13	8
		$K_1 \cdot 10^{-1}$	20	13	8	20	13	8	20	13	8	19	13	8
		K_1	20	13	8	20	13	8	20	13	8	19	13	8
$\frac{1}{32}$	5E–10	$K_1 \cdot 10^{-2}$	20	13	8	19	13	8	19	13	8	19	13	8
		$K_1 \cdot 10^{-1}$	20	13	8	19	13	8	19	13	8	19	13	8
		K_1	20	13	8	19	13	8	19	13	8	19	13	8
	1E-8	$K_1 \cdot 10^{-2}$	20	13	8	19	13	8	19	13	8	19	13	8
		$K_1 \cdot 10^{-1}$	20	13	8	19	13	8	19	13	8	19	13	8
		K_1	20	13	8	19	13	8	19	13	8	19	13	8
$\frac{1}{64}$	5E–10	$K_1 \cdot 10^{-2}$	19	13	8	19	13	8	19	13	8	19	13	8
		$K_1 \cdot 10^{-1}$	19	13	8	19	13	8	19	13	8	19	13	8
		K_1	19	13	8	19	13	8	19	13	8	19	13	8
	1E-8	$K_1 \cdot 10^{-2}$	19	13	8	19	13	8	19	13	8	19	13	8
		$K_1 \cdot 10^{-1}$	19	13	8	19	13	8	19	13	8	19	13	8
		K_1	19	13	8	19	13	8	19	13	8	19	13	8

6.2. The four-network model. This subsection is devoted to the four-network MPET model. As with the previous example, the boundary Γ of Ω is split into bottom (Γ_1), right (Γ_2), top (Γ_3), and left (Γ_4) boundaries. The considered boundary conditions are chosen as

TABLE 6

Elapsed times in seconds for the preconditioned MinRes (t_M), preconditioned GMRES (t_G), and fixed-stress split (t_F) algorithms to reach residual reduction by a factor 10^8 in the norm induced by the preconditioner: Barenblatt problem, $h = 1/64$.

	β		K_2			$K_2 \cdot 10^2$			$K_2 \cdot 10^4$			$K_2 \cdot 10^6$		
			t_M	t_G	t_F	t_M	t_G	t_F	n_M	t_G	t_F	t_M	t_G	t_F
$\lambda \cdot 10^{-2}$	5E–10	$K_1 \cdot 10^{-2}$	1.4	0.8	0.8	3.2	0.9	0.8	3.7	0.9	0.8	2.1	0.9	0.9
		$K_1 \cdot 10^{-1}$	1.4	0.8	0.8	3.2	0.9	0.8	3.7	0.9	0.8	2.1	0.9	0.9
		K_1	1.4	0.8	0.8	3.2	0.9	0.8	3.7	0.9	0.8	2.1	0.9	0.9
	1E-8	$K_1 \cdot 10^{-2}$	1.4	0.8	0.8	3.2	0.9	0.8	3.7	0.9	0.8	2.1	0.9	0.9
		$K_1 \cdot 10^{-1}$	1.4	0.8	0.8	3.2	0.9	0.8	3.7	0.9	0.8	2.1	0.9	0.9
		K_1	1.4	0.8	0.8	3.2	0.9	0.8	3.7	0.9	0.8	2.1	0.9	0.9
λ	5E–10	$K_1 \cdot 10^{-2}$	1.0	0.5	0.6	1.8	0.6	0.6	2.1	0.6	0.6	1.5	0.6	0.7
		$K_1 \cdot 10^{-1}$	1.0	0.5	0.6	1.8	0.6	0.6	2.1	0.6	0.6	1.5	0.6	0.7
		K_1	1.0	0.5	0.6	1.8	0.6	0.6	2.1	0.6	0.6	1.5	0.6	0.7
	1E-8	$K_1 \cdot 10^{-2}$	1.0	0.5	0.6	1.8	0.6	0.6	2.1	0.6	0.6	1.5	0.6	0.7
		$K_1 \cdot 10^{-1}$	1.0	0.5	0.6	1.8	0.6	0.6	2.1	0.6	0.6	1.5	0.6	0.7
		K_1	1.0	0.5	0.6	1.8	0.6	0.6	2.1	0.6	0.6	1.5	0.6	0.7
$\lambda \cdot 10^2$	5E–10	$K_1 \cdot 10^{-2}$	0.4	0.2	0.2	1.0	0.3	0.2	1.2	0.3	0.2	0.8	0.3	0.2
		$K_1 \cdot 10^{-1}$	0.4	0.2	0.2	1.0	0.3	0.2	1.2	0.3	0.2	0.8	0.3	0.2
		K_1	0.4	0.2	0.2	1.0	0.3	0.2	1.2	0.3	0.2	0.8	0.3	0.2
	1E-8	$K_1 \cdot 10^{-2}$	0.4	0.2	0.2	1.0	0.3	0.2	1.2	0.3	0.2	0.8	0.3	0.2
		$K_1 \cdot 10^{-1}$	0.4	0.2	0.2	1.0	0.3	0.2	1.2	0.3	0.2	0.8	0.3	0.2
		K_1	0.4	0.2	0.2	1.0	0.3	0.2	1.2	0.3	0.2	0.8	0.3	0.2

TABLE 7

Base values of model parameters for a four-network MPET model.

Parameter	Value	Unit
λ	505	Nm^{-2}
μ	216	Nm^{-2}
$c_{p_1} = c_{p_2} = c_{p_3} = c_{p_4}$	$4.5 \cdot 10^{-10}$	$\mathrm{m}^2\mathrm{N}^{-1}$
$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$	0.99	
$\beta_{12} = \beta_{24}$	$1.5 \cdot 10^{-19}$	$\mathrm{m}^2\mathrm{N}^{-1}\mathrm{s}^{-1}$
β_{23}	$2.0 \cdot 10^{-19}$	$\mathrm{m}^2\mathrm{N}^{-1}\mathrm{s}^{-1}$
β_{34}	$1.0 \cdot 10^{-13}$	$\mathrm{m}^2\mathrm{N}^{-1}\mathrm{s}^{-1}$
$K_1 = K_2 = K_4 = K$	$(1.0 \cdot 10^{-10})/(2.67 \cdot 10^{-3})$	$\mathrm{m}^2/\mathrm{Nsm}^{-2}$
K_3	$(1.4 \cdot 10^{-14})/(8.9 \cdot 10^{-4})$	$\mathrm{m}^2/\mathrm{Nsm}^{-2}$

$$
\begin{aligned}
(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I} - p_3\boldsymbol{I} - p_4\boldsymbol{I})\boldsymbol{n} &= (0,0)^T && \text{on } \Gamma_1 \cup \Gamma_2,\\
(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I} - p_3\boldsymbol{I} - p_4\boldsymbol{I})\boldsymbol{n} &= (0,-1)^T && \text{on } \Gamma_3,\\
\boldsymbol{u} &= \boldsymbol{0} && \text{on } \Gamma_4,\\
p_1 &= 2 && \text{on } \Gamma,\\
p_2 &= 20 && \text{on } \Gamma,\\
p_3 &= 30 && \text{on } \Gamma,\\
p_4 &= 40 && \text{on } \Gamma,
\end{aligned}
$$

whereas the right-hand sides are $\boldsymbol{f} = \boldsymbol{0}$, $g_1 = g_2 = g_3 = g_4 = 0$.

Table 7 shows the base values of the parameters which have been taken from [41]. The presented numerical results in Table 8 demonstrate the superiority of the fixed-stress split iterative method over the preconditioned MinRes algorithm and its robustness with respect to large variations of the coefficients λ, K_3 and $K = K_1 = K_2 = K_4$.

TABLE 8

Number of preconditioned MinRes (n_M) and fixed-stress split (n_F) iterations: parameters from Table 7.

h			$K_3 \cdot 10^{-2}$		K_3		$K_3 \cdot 10^{2}$		$K_3 \cdot 10^{4}$		$K_3 \cdot 10^{6}$		$K_3 \cdot 10^{10}$	
			n_M	n_F	n_M	n_F	n_M	n_F	n_M	n_F	n_M	n_F	n_M	n_F
$\frac{1}{16}$	λ	$K \cdot 10^{-2}$	34	10	34	10	26	10	23	10	21	10	21	10
		K	24	10	24	10	24	10	22	10	21	10	19	10
		$K \cdot 10^{2}$	21	10	21	10	23	10	23	10	31	10	30	10
	$\lambda \cdot 10^{4}$	$K \cdot 10^{-2}$	18	2	23	2	24	2	34	2	34	2	34	2
		K	11	2	17	2	34	2	31	2	31	2	31	2
		$K \cdot 10^{2}$	9	2	14	2	32	2	21	2	14	2	14	2
	$\lambda \cdot 10^{8}$	$K \cdot 10^{-2}$	14	2	14	2	12	2	12	2	12	2	12	2
		K	11	2	14	2	9	2	7	2	7	2	7	2
		$K \cdot 10^{2}$	9	2	14	2	9	2	5	2	5	2	5	2
$\frac{1}{32}$	λ	$K \cdot 10^{-2}$	34	10	32	10	26	10	23	10	19	10	19	10
		K	24	10	24	10	24	10	22	10	21	10	20	10
		$K \cdot 10^{2}$	21	10	21	10	21	10	26	10	41	10	39	10
	$\lambda \cdot 10^{4}$	$K \cdot 10^{-2}$	18	2	25	2	30	2	34	2	34	2	34	2
		K	12	2	20	2	35	2	31	2	31	2	31	2
		$K \cdot 10^{2}$	9	2	18	2	34	2	21	2	14	2	14	2
	$\lambda \cdot 10^{8}$	$K \cdot 10^{-2}$	14	2	14	2	12	2	12	2	12	2	12	2
		K	12	2	14	2	9	2	7	2	7	2	7	2
		$K \cdot 10^{2}$	11	2	14	2	9	2	6	2	5	2	5	2
$\frac{1}{64}$	λ	$K \cdot 10^{-2}$	34	10	32	10	26	10	21	10	19	10	19	10
		K	24	10	24	10	24	10	23	10	22	10	21	10
		$K \cdot 10^{2}$	21	10	21	10	21	10	36	10	45	10	45	10
	$\lambda \cdot 10^{4}$	$K \cdot 10^{-2}$	20	2	28	2	34	2	34	2	34	2	34	2
		K	13	2	25	2	36	2	31	2	31	2	31	2
		$K \cdot 10^{2}$	6	2	25	2	36	2	21	2	14	2	14	2
	$\lambda \cdot 10^{8}$	$K \cdot 10^{-2}$	14	2	14	2	12	2	12	2	12	2	12	2
		K	12	2	14	2	9	2	7	2	7	2	7	2
		$K \cdot 10^{2}$	12	2	14	2	9	2	6	2	5	2	5	2

7. Concluding remarks. To the best of our knowledge, this paper is the first example of a proposed and analyzed fixed-stress split iterative scheme for a three-field formulation of the MPET model. Fundamental to the linear convergence of the evolved algorithm is the incorporation of stabilization that employs the sum of all pressures. By applying the stability results proven in [24], we have demonstrated that the contraction rate of this fixed-point iteration is independent of any model physical parameters. Furthermore, the performed numerical experiments have confirmed the theoretical findings and have clearly demonstrated the efficiency of the presented fixed-stress scheme.

REFERENCES

[1] J. ADLER, F. GASPAR, X. HU, C. RODRIGO, AND L. ZIKATANOV, *Robust block preconditioners for Biot's model*, in Domain Decomposition Methods in Science and Engineering XXIV, P. Bjørstad, ed., Lect. Notes Comput. Sci. Eng. 125, Springer, New York, 2019, pp. 3–16.

[2] T. ALMANI, K. KUMAR, A. DOGRU, G. SINGH, AND M. WHEELER, *Convergence analysis of multirate fixed-stress split iterative schemes for coupling flow with geomechanics*, Comput. Methods Appl. Mech. Engrg., 311 (2016), pp. 180–207.

[3] T. ALMANI, K. KUMAR, AND M. WHEELER, *Convergence and error analysis of fully discrete iterative coupling schemes for coupling flow with geomechanics*, Comput. Geosci., 21 (2017), pp. 1157–1172.

[4] M. ALNÆS, J. BLECHTA, J. HAKE, A. JOHANSSON, B. KEHLET, A. LOGG, C. RICHARDSON, J. RING, M. ROGNES, AND G. WELLS, *The FeniCS project version* 1.5, Arch. Numer. Software, 3 (2015).

[5] G. Barenblatt, G. Zheltov, and I. Kochina, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata]*, J. Appl. Math. Mech., 24 (1960), pp. 1286–1303.
[6] M. Bause, F. Radu, and U. Köcher, *Space-time finite element approximation of the Biot poroelasticity system with iterative coupling*, Comput. Methods Appl. Mech. Engrg., 320 (2017), pp. 745–768.
[7] M. Biot, *General theory of three-dimensional consolidation*, J. Appl. Phys., 12 (1941), pp. 155–164.
[8] M. Biot, *Theory of elasticity and consolidation for a porous anisotropic solid*, J. Appl. Phys., 26 (1955), pp. 182–185.
[9] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Ser. Comput. Math. 44, Springer, Heidelberg, 2013.
[10] M. Borregales, F. Radu, K. Kumar, and J. Nordbotten, *Robust iterative schemes for non-linear poromechanics*, Comput. Geosci., 22 (2018), pp. 1021–1038.
[11] J. Both, K. Kumar, J. Nordbotten, and F. Radu, *Iterative methods for coupled flow and geomechanics in unsaturated porous media*, in Proceedings of Poromechanics VI, 2017, pp. 411–418.
[12] J. Both, K. Kumar, J. Nordbotten, and F. Radu, *Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media*, Comput. Math. Appl., 77 (2019), pp. 1499–1502.
[13] J. W. Both, M. Borregales, J. Nordbotten, K. Kumar, and F. Radu, *Robust fixed stress splitting for Biot's equations in heterogeneous media*, Appl. Math. Lett., 68 (2017), pp. 101–108.
[14] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129–151.
[15] D. Chou, J. Vardakis, L. Guo, B. Tully, and Y. Ventikos, *A fully dynamic multi-compartmental poroelastic system: Application to aqueductal stenosis*, J. Biomech., 49 (2016), pp. 2306–2312.
[16] S. Dana and M. Wheeler, *Convergence analysis of two-grid fixed stress iterative scheme for coupled flow and deformation in heterogeneous poroelastic media*, Comput. Methods Appl. Mech. Engrg., 341 (2018).
[17] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for non-symmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.
[18] H. C. Elman, *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations*, Ph.D. thesis, Yale University, New Haven, CT, 1982.
[19] *The Standard NAFEMS Benchmarks*, NAFEMS, Glasgow, 1990.
[20] V. Girault, K. Kumar, and M. Wheeler, *Convergence of iterative coupling of geomechanics with flow in a fractured poroelastic medium*, Comput. Geosci., 20 (2016), pp. 997–1011.
[21] L. Guo, J. Vardakis, T. Lassila, M. Mitolo, N. Ravikumar, D. Chou, M. Lange, A. Sarrami-Foroushani, B. Tully, Z. Taylor, S. Varma, A. Venneri, A. Frangi, and Y. Ventikos, *Subject-specific multi-poroelastic model for exploring the risk factors associated with the early stages of alzheimer's disease*, Interface Focus, 8 (2018), 20170019.
[22] Q. Hong and J. Kraus, *Uniformly stable discontinuous Galerkin discretization and robust iterative solution methods for the Brinkman problem*, SIAM J. Numer. Anal., 54 (2016), pp. 2750–2774.
[23] Q. Hong and J. Kraus, *Parameter-robust stability of classical three-field formulation of Biot's consolidation model*, Electron. Trans. Numer. Anal., 48 (2018), pp. 202–226.
[24] Q. Hong, J. Kraus, M. Lymbery, and F. Philo, *Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models*, Numer. Linear Algebra Appl., 26 (2019), e2242.
[25] Q. Hong, J. Kraus, J. Xu, and L. Zikatanov, *A robust multigrid method for discontinuous Galerkin discretizations of Stokes and linear elasticity equations*, Numer. Math., 132 (2016), pp. 23–49.
[26] X. Hu, C. Rodrigo, F. Gaspar, and L. Zikatanov, *A nonconforming finite element method for the Biot's consolidation model in poroelasticity*, J. Comput. Appl. Math., 310 (2017), pp. 143–154.
[27] J. Kim, H. Tchelepi, and R. Juanes, *Stability, accuracy and efficiency of sequential methods for coupled flow and geomechanics*, SPE J., 16 (2011).
[28] A. Kolesov and P. Vabishchevich, *Splitting schemes with respect to physical processes for double-porosity poroelasticity problems*, Russ. J. Numer. Anal. Math. Model., 32 (2017).
[29] J. Lee, K.-A. Mardal, and R. Winther, *Parameter-robust discretization and preconditioning of Biot's consolidation model*, SIAM J. Sci. Comput., 39 (2017), pp. A1–A24.

[30] J. Lee, E. Piersanti, K.-A. Mardal, and M. Rognes, *A mixed finite element method for nearly incompressible multiple-network poroelasticity*, SIAM J. Sci. Comput., 41 (2019), pp. A722–A747.

[31] S. Lee, M. Wheeler, and T. Wick, *Iterative coupling of flow, geomechanics and adaptive phase-field fracture including level-set crack width approaches*, J. Comput. Appl. Math., 314 (2017), pp. 40–60.

[32] A. Logg, K.-A. Mardal, and G. Wells, eds., *Automated Solution of Differential Equations by the Finite Element Method*, Lect. Notes Comput. Sci. Eng. 84, Springer, New York, 2012.

[33] D. Loghin and A. J. Wathen, *Analysis of preconditioners for saddle-point problems*, SIAM J. Sci. Comput., 25 (2004), pp. 2029–2049.

[34] K.-A. Mardal and R. Winther, *Preconditioning discretizations of systems of partial differential equations*, Numer. Linear Algebra Appl., 18 (2011), pp. 1–40.

[35] A. Mikelić and M. Wheeler, *Convergence of iterative coupling for coupled flow and geomechanics*, Comput. Geosci., 17 (2013), pp. 455–461.

[36] C. Rodrigo, X. Hu, P. Ohm, J. Adler, F. Gaspar, and L. Zikatanov, *New stabilized discretizations for poroelasticity and the Stokes' equations*, Comput. Methods Appl. Mech. Engrg., 341 (2018), pp. 467–484.

[37] D. Schötzau, C. Schwab, and A. Toselli, *Mixed hp-DGFEM for incompressible flows*, SIAM J. Numer. Anal., 40 (2002), pp. 2171–2194.

[38] R. Showalter, *Poroelastic filtration coupled to Stokes flow*, Lecture Notes in Pure and Appl. Math., 242 (2010), pp. 229–241.

[39] E. Storvik, J. Both, K. Kumar, J. Nordbotten, and F. Radu, *On the Optimization of the Fixed-Stress Splitting for Biot's Equations*, https://arxiv.org/abs/1811.06242v2, 2018.

[40] B. Tully and Y. Ventikos, *Cerebral water transport using multiple-network poroelastic theory: Application to normal pressure hydrocephalus*, J. Fluid Mech., 667 (2011), pp. 188–215.

[41] J. Vardakis, D. Chou, B. Tully, C. Hung, T. Lee, P. Tsui, and Y. Ventikos, *Investigating cerebral oedema using poroelasticity*, Med. Eng. Phys., 38 (2016), pp. 48–57.

PARAMETER-ROBUST UZAWA-TYPE ITERATIVE METHODS FOR DOUBLE SADDLE POINT PROBLEMS ARISING IN BIOT'S CONSOLIDATION AND MULTIPLE-NETWORK POROELASTICITY MODELS

Mathematical Models and Methods in Applied Sciences
Vol. 30, No. 13 (2020) 2523–2555

DOI: 10.1142/S0218202520500499

World Scientific
www.worldscientific.com

Parameter-robust Uzawa-type iterative methods for double saddle point problems arising in Biot's consolidation and multiple-network poroelasticity models

Qingguo Hong

Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA
huq11@psu.edu

Johannes Kraus*, Maria Lymbery† and Fadi Philo‡

Faculty of Mathematics, University of Duisburg-Essen, Thea-Leymann-Straße 9, Essen 45127, Germany
**johannes.kraus@uni-due.de*
†*maria.lymbery@uni-due.de*
‡*fadi.philo@uni-due.de*

Received 13 November 2019
Revised 26 July 2020
Accepted 1 September 2020
Published 10 December 2020
Communicated by J. Xu

This work is concerned with the iterative solution of systems of quasi-static multiple-network poroelasticity equations describing flow in elastic porous media that is permeated by single or multiple fluid networks. Here, the focus is on a three-field formulation of the problem in which the displacement field of the elastic matrix and, additionally, one velocity field and one pressure field for each of the $n \geq 1$ fluid networks are the unknown physical quantities. Generalizing Biot's model of consolidation, which is obtained for $n = 1$, the MPET equations for $n \geq 1$ exhibit a double saddle point structure. The proposed approach is based on a framework of augmenting and splitting this three-by-three block system in such a way that the resulting block Gauss–Seidel preconditioner defines a fully decoupled iterative scheme for the flux-, pressure-, and displacement fields. In this manner, one obtains an augmented Lagrangian Uzawa-type method, the analysis of which is the main contribution of this work. The parameter-robust uniform linear convergence of this fixed-point iteration is proved by showing that its rate of contraction is strictly less than one independent of all physical and discretization parameters. The theoretical results are confirmed by a series of numerical tests that compare the new fully decoupled scheme to the very popular partially decoupled fixed-stress split iterative method, which decouples only flow — the flux and pressure fields remain coupled in this case — from the mechanics problem. We further test the performance of the

*Corresponding author.

block-triangular preconditioner defining the new scheme when used to accelerate the generalized minimal residual method (GMRES) algorithm.

Keywords: Biot's consolidation; multiple-network poroelasticity; double saddle point problems; parameter-robust Uzawa method; field-of-values-equivalent preconditioners.

AMS Subject Classification 2020: 65M12, 65M60, 65F10, 65N22, 35Q92

1. Introduction

In this paper, we propose and analyze stationary iterative methods for solving the equations of multiple-network poroelastic theory which describe flow in deformable porous media. The latter is modeled as an elastic solid matrix comprising $n \geq 1$ superimposed fluid networks with possibly vastly varying characteristic length scales and hydraulic conductivities, see e.g. Ref. 53 and the references therein.

Dual-porosity/dual-permeability models have been proposed and studied in a geomechanical context,[7,8] providing a generalization of Biot's consolidation model which is obtained for $n = 1$, see Refs. 12 and 13. Over the last decade, the MPET equations have gradually gained attention as a tool for modeling flow across scales and networks in soft tissue. Biological multicompartmental poroelasticity models can be used to embed more specific medical models, e.g. to describe water transport in the cerebral environment and explore the pathogenesis of acute and chronic hydrocephalus,[51] or to study effects of obstructing cerebrospinal fluid (CSF) transport and to demonstrate the impact of aqueductal stenosis and fourth ventricle outlet obstruction (FVOO),[52,54] or to find medical indications of oedema formation.[20]

Recently, the MPET model has also been used in order to gain a better understanding of the processes involved with the mechanisms behind Alzheimer's disease (AD), the most common form of dementia, cf. Ref. 24. Most prominently, the so-called amyloid hypothesis states that the accumulation of neurotoxic amyloid-β (Aβ) into parenchymal senile plaques or within the walls of arteries is a basic cause of this disease. In Ref. 23, a partial validation of a four-network poroelastic model for metabolic waste clearance is presented in a qualitative way, i.e. by showing a qualitative agreement of the cerebral blood flow (CBF) data obtained from arterial spin labeling (ASL) images and the corresponding model output for different regions of the brain. Although the authors of these papers conclude that there is a need for more experimental and clinical data to optimize the boundary conditions and parameters used in numerical modeling, they also stress the potential of MPET modeling as a testing bed for hypotheses and new theories in neuroscience research.

Regarding the numerical solution of the MPET equations mainly two different approaches have been investigated in the last couple of years. The first one has been proposed in Ref. 37 and uses a mixed finite element formulation based on introducing an additional total pressure variable. Energy estimates for the continuous solutions and *a priori* error estimates for a family of compatible semidiscretizations demonstrate that this formulation is robust for nearly incompressible materi-

als, small storage coefficients, and small or vanishing transfer coefficients between networks.

The second approach is based on a generalization of the classical three-field formulation of Biot's model and explicitly accommodates Darcy's law for each fluid network. This formulation enforces the exact conservation of mass at the price of including additionally n vector fields for the Darcy velocities (fluxes). A parameter-robust stability analysis of this flux-based MPET model has been presented in Ref. 27 along with fully parameter-robust norm-equivalent preconditioners. Following Refs. 26 and 31, the authors propose in Ref. 27 a family of strongly conservative locking-free discretizations for the MPET model and establish the related optimal error estimates for the stationary problems arising from implicit time discretization by the backward Euler method. These results also cover the case of vanishing storage coefficients.

Various works can be found on discretizations and efficient iterative solvers and preconditioning techniques for the quasi-static Biot model addressing two-field,[1,14] three-field,[26,30,36,45] and four-field formulations.[6,35]

Two of the most popular and likely most efficient iterative schemes for solving the equations of poroelasticity are the so-called undrained split and fixed-stress split iterative methods, which, contrary to the drained split and the fixed-strain split methods, are unconditionally stable, see Ref. 32. The first convergence analysis of the former methods has been presented in Ref. 43 for the quasi-static Biot system. Subsequent refined results focus mostly on variants of the fixed-stress method addressing multirate fixed-stress split iterative schemes,[2] fully discrete iterative coupling of flow and geomechanics,[3] heterogeneous media and linearized Biot's equations,[16] two-grid fixed-stress schemes for heterogeneous media,[22] or space-time finite element approximations of the quasi-static Biot system.[9] A strategy for optimizing the stabilization parameter in the fixed-stress split iterative method for the Biot problem in two-field formulation has been presented in Ref. 50.

The fixed-stress method has also been recently successfully used in combination with Anderson acceleration for the solution of nonlinear poromechanics problems.[17] Moreover, monolithic and splitting based solution schemes have been considered and analyzed for solving quasi-static thermo-poroelasticity problems with nonlinear convective transport, see Ref 19. The latter work focuses on the analysis of fully and partially decoupled schemes for heat, mechanics and flow applied to the linearized problem obtained via the so-called L-scheme. All previously mentioned works, in presence of flux and pressure unknowns, solve the flow equations implicitly, i.e. as a coupled subsystem, a strategy which we will not pursue in this paper.

A desirable property of preconditioners, in addition to their uniformity with respect to discretization parameters, is their robustness regarding potentially large variations of the physical parameters. This task can be studied in the framework of operator preconditioning on the level of the continuous model.[42] Targeting Biot's consolidation model the parameter-robustness of norm-equivalent preconditioners has been established in Ref. 36 for the total pressure-based formulation and in

Ref. 26 for the classical three-field formulation based on displacement, Darcy velocity and fluid pressure fields. Both approaches have been generalized to the MPET model.[27,37]

One potential advantage of the approach presented in Ref. 27 is exact mass conservation. A disadvantage, however, is that the presence of n fluxes and n associated pressures makes the system in general more difficult and also more time-consuming to solve. The fixed-stress split iterative method has recently been generalized to be applicable not only to the Biot ($n = 1$) but also to the more general MPET ($n \geq 1$) systems in Ref. 28 which presents a fully parameter-robust convergence analysis and determines a close to optimal acceleration parameter.

However, in the conservative approach obtained from generalizing the classical three-field formulation of Biot's model, the block of n unknown fluxes (with d components each) couples to a block of n pressure unknowns creating a subsystem with $n(d+1)$ scalar quantities of interest as compared to the $(n(d+1)+d)$ unknown scalar functions in the whole system. Hence, considering the above-mentioned four-network model ($n = 4$) in three space dimensions ($d = 3$), for example, this results in a flux–pressure subsystem with approximately 16/19 of the size of the whole system. This explains why a further decoupling of the flux from the pressure block of unknowns in an iterative method is of particular interest in this approach.

The goal of this paper is to propose and analyze a class of fully decoupled iterative schemes, which contrary to the fixed-stress split iterative method also decouple the *flux–pressure* subsystem. In this respect, it can be seen as a continuation of the analysis presented in Ref. 28.

As already mentioned, the target problem is a three-by-three block system with a double saddle point. The abstract canonical form of the operator (matrix) of the related operator equation can be represented in the form

$$\begin{bmatrix} A_1 & 0 & B_1^T \\ 0 & A_2 & B_2^T \\ B_1 & B_2 & -C \end{bmatrix} \tag{1.1}$$

with A_1 and A_2 being symmetric positive definite (SPD) operators and C a symmetric positive semidefinite (SPSD) operator. The operator (1.1) defines a double saddle point problem and can be rearranged in such a way that it has the form

$$\begin{bmatrix} A_1 & B_1^T & 0 \\ B_1 & -C & B_2 \\ 0 & B_2^T & A_2 \end{bmatrix} \tag{1.2}$$

and thus fits the definition of a multiple saddle point operator as given in Ref. 49 where block-diagonal Schur complement preconditioners for multiple saddle point problems of block tridiagonal form are analyzed. We will use a combined augmentation and splitting technique to construct in a block Gauss–Seidel framework fully decoupled augmented Lagrangian Uzawa-type methods for linear systems with an

operator (matrix) of the canonical form (1.1). Although our methodical approach to construct preconditioners is similar to the one taken in recent works, see Refs. 10, 11, and 56, there are also major differences. First, the double saddle point problems considered in Refs. 10 and 11 are generated by operators of the canonical form

$$\begin{bmatrix} A_1 & B_1^T & B_2^T \\ B_1 & 0 & 0 \\ B_2 & 0 & -C \end{bmatrix} \tag{1.3}$$

with A_1 being SPD and C being SPSD. It can easily be seen that the operators (1.1) and (1.3) are of a different type in the sense that they cannot be transferred one into the other by permutations of rows and columns. The second main difference is that the analysis in Refs. 10 and 11 uses arguments from classical linear algebra whereas our convergence proofs use techniques from functional analysis aiming at quantitative bounds that might be useful when applying the proposed iterative methods at the level of finite element approximations of the continuous problems.

The remainder of the paper is organized as follows: In Sec. 2, we first formulate the MPET problem, introduce the notation and transform the problem into a coupled system with a double saddle point operator of the form (1.2). Based on this notation we then recall the fixed-stress split iterative method in a block Gauss–Seidel framework. It follows the construction of a new class of fully decoupled iterative Uzawa-type methods, which requires an additional augmentation step. This section ends with summarizing some preliminary and auxiliary results that are used in the convergence analysis of the new class of methods presented in Sec. 3. The numerical tests in Sec. 5 serve the assessment of the performance of the iterative methods and preconditioners developed in this paper comparing them also with the fixed-stress split iterative method analyzed in Ref. 28.

2. Iterative Coupling Methods for the MPET Problem

2.1. *The MPET system — formulation and notation*

Consider the quasi-static MPET equations in a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$:

$$\boldsymbol{v}_i + K_i \nabla p_i = \mathbf{0} \quad \text{in } \Omega \times (0,T), \tag{2.1a}$$

$$-\alpha_i \operatorname{div} \dot{\boldsymbol{u}} - \operatorname{div} \boldsymbol{v}_i - c_{p_i} \dot{p}_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} \beta_{ij}(p_i - p_j) = g_i \quad \text{in } \Omega \times (0,T), \tag{2.1b}$$

$$-\operatorname{div} \boldsymbol{\sigma} + \sum_{i=1}^{n} \alpha_i \nabla p_i = \boldsymbol{f} \quad \text{in } \Omega \times (0,T). \tag{2.1c}$$

where (2.1a) and (2.1b) are for $i = 1, \ldots, n$. The unknown physical quantities in this system are the displacement field $\boldsymbol{u}$, the seepage velocities, or fluxes, $\boldsymbol{v}_i$, and

the scalar pressure fields p_i. The effective stress and strain tensors are given by

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\epsilon}(\boldsymbol{u}) + \lambda \mathrm{div}(\boldsymbol{u})\boldsymbol{I} \quad \text{and} \quad \boldsymbol{\epsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T), \tag{2.2}$$

respectively, with the Lamé parameters λ and μ defined via the modulus of elasticity E and the Poisson ratio $\nu \in [0, 1/2)$ as follows:

$$\lambda := \frac{\nu E}{(1+\nu)(1-2\nu)}, \quad \mu := \frac{E}{2(1+\nu)}.$$

In (2.1), α_i denote the Biot–Willis coefficients, c_{p_i} the constrained specific storage coefficients, and K_i the hydraulic conductivities, which in this paper for convenience only, are scalars defining the tensor coefficients $\boldsymbol{K}_i = K_i\boldsymbol{I}$ where $\boldsymbol{I}$ is the identity (matrix) operator. Considering the right-hand sides in (2.1c) and (2.1b), $\boldsymbol{f}$ denotes the body force density whereas g_i represent the fluid extractions or injections, see e.g. Ref. 48 and the references therein. The parameters $\beta_{ij} = \beta_{ji}$, $i \neq j$ couple the network pressures and are called network transfer coefficients.

By substituting the expression for the stress tensor from (2.2) in (2.1c) the MPET system takes the form:

$$\boldsymbol{v}_i + K_i \nabla p_i = \boldsymbol{0}, \quad i = 1, \ldots, n, \tag{2.3a}$$

$$-\mathrm{div}\, \boldsymbol{v}_i - c_{p_i}\dot{p}_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} \beta_{ij}(p_i - p_j) - \alpha_i \mathrm{div}\dot{\boldsymbol{u}} = g_i, \quad i = 1, \ldots, n, \tag{2.3b}$$

$$\sum_{i=1}^{n} \alpha_i \nabla p_i - 2\mu \mathrm{div}\, \boldsymbol{\epsilon}(\boldsymbol{u}) - \lambda \nabla \mathrm{div}\, \boldsymbol{u} = \boldsymbol{f}. \tag{2.3c}$$

After imposing proper boundary and initial conditions, see Ref. 27, and using the backward Euler method for time discretization, one has to solve a static problem of the form

$$K_i^{-1}\boldsymbol{v}_i^k + \nabla p_i^k = \boldsymbol{0}, \quad i = 1, \ldots, n, \tag{2.4a}$$

$$-\alpha_i \mathrm{div}\, \boldsymbol{u}^k - \tau \mathrm{div}\, \boldsymbol{v}_i^k - c_{p_i} p_i^k - \tau \sum_{\substack{j=1 \\ j\neq i}}^{n} \beta_{ij}(p_i^k - p_j^k) = g_i^k, \quad i = 1, \ldots, n, \tag{2.4b}$$

$$-2\mu \mathrm{div}\, \boldsymbol{\epsilon}(\boldsymbol{u}^k) - \lambda \nabla \mathrm{div}\, \boldsymbol{u}^k + \sum_{i=1}^{n} \alpha_i \nabla p_i^k = \boldsymbol{f}^k, \tag{2.4c}$$

in each time step, i.e. at every time moment $t_k = t_{k-1} + \tau$, $k = 1, 2, \ldots$. Here, $\boldsymbol{u}^k$, $\boldsymbol{v}_i^k$, p_i^k are approximations of $\boldsymbol{u}$, $\boldsymbol{v}_i$, p_i at $t = t_k$ and $\boldsymbol{f}^k = \boldsymbol{f}(x, t_k)$, $g_i^k = -\tau g_i(x, t_k) - \alpha_i \mathrm{div}(\boldsymbol{u}^{k-1}) - c_{p_i} p_i^{k-1}$ for $i = 1, \ldots, n$. After dividing (2.4) by 2μ,

denoting

$$\frac{\lambda}{2\mu} \to \lambda, \quad \frac{\alpha_i}{2\mu} \to \alpha_i, \quad \frac{\boldsymbol{f}^k}{2\mu} \to \boldsymbol{f}^k, \quad \frac{\tau}{2\mu} \to \tau,$$

$$\frac{c_{p_i}}{2\mu} \to c_{p_i}, \quad \frac{g_i^k}{2\mu} \to g_i^k, \quad i = 1, \ldots, n,$$

and further introducing the new variables

$$\boldsymbol{v}_i := \frac{\tau}{\alpha_i}\boldsymbol{v}_i^k, \quad p_i := \alpha_i p_i^k, \quad \boldsymbol{u} := \boldsymbol{u}^k, \quad \boldsymbol{f} := \boldsymbol{f}^k,$$

$$g_i := \frac{g_i^k}{\alpha_i}, \quad i = 1, \ldots, n,$$

system (2.4) can be presented in the form

$$\tau^{-1}K_i^{-1}\alpha_i^2\boldsymbol{v}_i + \nabla p_i = \boldsymbol{0}, \tag{2.5a}$$

$$-\operatorname{div}\boldsymbol{u} - \operatorname{div}\boldsymbol{v}_i - \frac{c_{p_i}}{\alpha_i^2}p_i + \sum_{\substack{j=1\\ j\neq i}}^{n}\left(-\frac{\tau\beta_{ij}}{\alpha_i^2}p_i + \frac{\tau\beta_{ij}}{\alpha_i\alpha_j}p_j\right) = g_i, \tag{2.5b}$$

$$-\operatorname{div}\boldsymbol{\epsilon}(\boldsymbol{u}) - \lambda\nabla\operatorname{div}\boldsymbol{u} + \sum_{i=1}^{n}\nabla p_i = \boldsymbol{f}, \tag{2.5c}$$

where (2.5a) and (2.5b) are for $i = 1, \ldots, n$, and we have also multiplied (2.4a) by α_i and (2.4b) by α_i^{-1}.

In what follows, we will also make use of the notation $\boldsymbol{v}^T := (\boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_n^T)$, $\boldsymbol{z}^T := (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)$, $\boldsymbol{p}^T := (p_1, \ldots, p_n)$, $\boldsymbol{q}^T := (q_1, \ldots, q_n)$ where $\boldsymbol{v}, \boldsymbol{z} \in \boldsymbol{V} = \boldsymbol{V}_1 \times \cdots \times \boldsymbol{V}_n$, $\boldsymbol{p}, \boldsymbol{q} \in \boldsymbol{P} = P_1 \times \cdots \times P_n$ and $\boldsymbol{U} = \{\boldsymbol{u} \in H^1(\Omega)^d : \boldsymbol{u} = \boldsymbol{0} \text{ on } \Gamma_{\boldsymbol{u},D}\}$, $\boldsymbol{V}_i = \{\boldsymbol{v}_i \in H(\operatorname{div}, \Omega) : \boldsymbol{v}_i \cdot \boldsymbol{n} = 0 \text{ on } \Gamma_{p_i,N}\}$, $P_i = L^2(\Omega)$, and $P_i = L_0^2(\Omega)$ if $\Gamma_{\boldsymbol{u},D} = \Gamma = \partial\Omega$. Using the parameter substitutions

$$R_i^{-1} := \tau^{-1}K_i^{-1}\alpha_i^2, \quad \alpha_{p_i} := \frac{c_{p_i}}{\alpha_i^2}, \quad \beta_{ii} := \sum_{\substack{j=1\\ j\neq i}}^{n}\beta_{ij},$$

$$\alpha_{ij} := \frac{\tau\beta_{ij}}{\alpha_i\alpha_j}, \quad \tilde{\alpha}_{ii} := -\alpha_{p_i} - \alpha_{ii}$$

for $i, j = 1, \ldots, n$, we further rewrite system (2.5) as

$$\mathcal{A}\begin{pmatrix}\boldsymbol{v}\\ \boldsymbol{p}\\ \boldsymbol{u}\end{pmatrix} = \begin{bmatrix}A_v & B_v^T & 0\\ B_v & -C & B_u\\ 0 & B_u^T & A_u\end{bmatrix}\begin{pmatrix}\boldsymbol{v}\\ \boldsymbol{p}\\ \boldsymbol{u}\end{pmatrix} = \begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{g}\\ \boldsymbol{f}\end{pmatrix}, \tag{2.6}$$

where

$$A_v := \begin{bmatrix} R_1^{-1}I & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & R_n^{-1}I \end{bmatrix}, \quad B_v := \begin{bmatrix} -\mathrm{div} & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & -\mathrm{div} \end{bmatrix},$$

$$-C := \begin{bmatrix} \tilde{\alpha}_{11}I & \alpha_{12}I & \dots & \alpha_{1n}I \\ \alpha_{21}I & \ddots & & \alpha_{2n}I \\ \vdots & & \ddots & \vdots \\ \alpha_{n1}I & \alpha_{n2}I & \dots & \tilde{\alpha}_{nn}I \end{bmatrix},$$

$$B_u := \begin{bmatrix} -\mathrm{div}, \dots, -\mathrm{div} \end{bmatrix}^T, \quad A_u := -\mathrm{div}\boldsymbol{\epsilon} - \lambda\nabla\mathrm{div}$$

and I is the identity operator. For the scaled parameters, we make the rather nonrestrictive assumptions

$$\lambda \geq 0, \quad R_1^{-1}, \dots, R_n^{-1} > 0, \quad \alpha_{p_1}, \dots, \alpha_{p_n} \geq 0, \quad \alpha_{ij} \geq 0, \quad i, j = 1, \dots, n. \quad (2.7)$$

From now on, we will use the same symbols for denoting operators and their corresponding coefficient matrices. Additionally, let us introduce

$$\Lambda_1 := \begin{bmatrix} \alpha_{11} & -\alpha_{12} & \dots & -\alpha_{1n} \\ -\alpha_{21} & \alpha_{22} & \dots & -\alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_{n1} & -\alpha_{n2} & \dots & \alpha_{nn} \end{bmatrix}, \quad \Lambda_2 := \begin{bmatrix} \alpha_{p_1} & 0 & \dots & 0 \\ 0 & \alpha_{p_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \alpha_{p_n} \end{bmatrix},$$

i.e. $C = (\Lambda_1 + \Lambda_2) \otimes I$. Further, denote $R^{-1} := \max\{R_i^{-1} : i = 1, \dots, n\}, \lambda_0 := \max\{1, \lambda\}$,

$$\Lambda_3 := \begin{bmatrix} R & 0 & \dots & 0 \\ 0 & R & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & R \end{bmatrix}, \quad \Lambda_4 := \begin{bmatrix} \frac{1}{\lambda_0} & \dots & \dots & \frac{1}{\lambda_0} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{1}{\lambda_0} & \dots & \dots & \frac{1}{\lambda_0} \end{bmatrix},$$

$$\Lambda := \Lambda_1 + \Lambda_2 + \Lambda_3 + \Lambda_4$$

and also, for any block vector $\boldsymbol{z}$ and vector $\boldsymbol{u}$

$$\mathrm{Div}\,\boldsymbol{z} := (\mathrm{div}\,\boldsymbol{z}_1, \dots, \mathrm{div}\,\boldsymbol{z}_n)^T, \quad \underline{\mathrm{Div}}\boldsymbol{u} := (\mathrm{div}\,\boldsymbol{u}, \dots, \mathrm{div}\,\boldsymbol{u})^T.$$

2.2. *The fixed-stress split iterative method revisited*

For any operator $\Lambda_L : \boldsymbol{P} \to \boldsymbol{P}^*$, $\mathcal{A}$ can be decomposed as follows:

$$\mathcal{A} = \begin{bmatrix} A_v & B_v^T & 0 \\ B_v & -C - \Lambda_L & 0 \\ 0 & B_u^T & A_u \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Lambda_L & B_u \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.8}$$

Applying the block Gauss–Seidel method to the above system, we obtain

$$\begin{bmatrix} A_v & B_v^T & 0 \\ B_v & -C - \Lambda_L & 0 \\ 0 & B_u^T & A_u \end{bmatrix} \begin{pmatrix} \boldsymbol{v}^{k+1} \\ \boldsymbol{p}^{k+1} \\ \boldsymbol{u}^{k+1} \end{pmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Lambda_L & B_u \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{v}^{k} \\ \boldsymbol{p}^{k} \\ \boldsymbol{u}^{k} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{g} \\ \boldsymbol{f} \end{pmatrix} \tag{2.9}$$

or, equivalently,

$$\begin{bmatrix} A_v & B_v^T & 0 \\ B_v & -C - \Lambda_L & 0 \\ 0 & B_u^T & A_u \end{bmatrix} \begin{pmatrix} \boldsymbol{v}^{k+1} \\ \boldsymbol{p}^{k+1} \\ \boldsymbol{u}^{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{g} \\ \boldsymbol{f} \end{pmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Lambda_L & B_u \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{v}^{k} \\ \boldsymbol{p}^{k} \\ \boldsymbol{u}^{k} \end{pmatrix}, \tag{2.10}$$

which is (a block variant of) the fixed-stress method. In Ref. 28, a parameter-robust convergence analysis of this method has been presented for the choice

$$\Lambda_L = L \begin{bmatrix} I & \dots & I \\ \vdots & \ddots & \vdots \\ I & \dots & I \end{bmatrix} \quad \text{where } L \geq \frac{1}{\lambda + c_K^2}. \tag{2.11}$$

I is the identity operator and c_K the constant in the estimate

$$\|\boldsymbol{\epsilon}(\boldsymbol{w})\| \geq c_K \|\operatorname{div} \boldsymbol{w}\| \quad \text{for all } \boldsymbol{w} \in \boldsymbol{U}. \tag{2.12}$$

Here, $\|\cdot\|$ denotes the L^2 norm, on the left-hand side of (2.12) of a tensor-valued and on the right-hand side of a scalar-valued function. Note that (2.12) holds true for example for $c_K = 1/\sqrt{d}$ where d is the space dimension.

2.3. *Uzawa-type methods in block Gauss–Seidel framework*

Now for any positive definite operator $M : \boldsymbol{P}^* \to \boldsymbol{P}$, we consider the equivalent augmented MPET system

$$\hat{\mathcal{A}} \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{p} \\ \boldsymbol{u} \end{pmatrix} = \begin{bmatrix} A_v + B_v^T M B_v & B_v^T - B_v^T M C & B_v^T M B_u \\ -B_v & C & -B_u \\ 0 & B_u^T & A_u \end{bmatrix} \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{p} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} B_v^T M \boldsymbol{g} \\ -\boldsymbol{g} \\ \boldsymbol{f} \end{pmatrix}. \tag{2.13}$$

Further, for any positive definite operator $S : \boldsymbol{P} \to \boldsymbol{P}^*$, we decompose $\hat{\mathcal{A}}$ in the form

$$\hat{\mathcal{A}} = \hat{\mathcal{A}}_L + \hat{\mathcal{A}}_U := \begin{bmatrix} A_v + B_v^T M B_v & 0 & 0 \\ -B_v & S & 0 \\ 0 & B_u^T & A_u \end{bmatrix} + \begin{bmatrix} 0 & B_v^T - B_v^T M C & B_v^T M B_u \\ 0 & -S + C & -B_u \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.14}$$

Next, applying the block Gauss–Seidel method to the above system yields

$$\hat{\mathcal{A}}_L \begin{pmatrix} \boldsymbol{v}^{k+1} \\ \boldsymbol{p}^{k+1} \\ \boldsymbol{u}^{k+1} \end{pmatrix} = \begin{pmatrix} B_v^T M \boldsymbol{g} \\ -\boldsymbol{g} \\ \boldsymbol{f} \end{pmatrix} - \hat{\mathcal{A}}_U \begin{pmatrix} \boldsymbol{v}^k \\ \boldsymbol{p}^k \\ \boldsymbol{u}^k \end{pmatrix}. \tag{2.15}$$

System (2.15) can be expressed in terms of bilinear forms as follows:

Algorithm 1. Fully decoupled iterative scheme for weak flux–pressure–displacement formulation of MPET problem.

Step a: Given $\boldsymbol{p}^k$ and $\boldsymbol{u}^k$, we first solve for $\boldsymbol{v}^{k+1}$, such that for all $\boldsymbol{z} \in \boldsymbol{V}$ there holds

$$\begin{aligned}(A_v \boldsymbol{v}^{k+1}, \boldsymbol{z}) + (M \operatorname{Div} \boldsymbol{v}^{k+1}, \operatorname{Div} \boldsymbol{z}) = &-(M\boldsymbol{g}, \operatorname{Div} \boldsymbol{z}) + (\boldsymbol{p}^k, \operatorname{Div} \boldsymbol{z}) \\ &- (M(\Lambda_1 + \Lambda_2)\boldsymbol{p}^k, \operatorname{Div} \boldsymbol{z}) \\ &- (M \underline{\operatorname{Div}}\, \boldsymbol{u}^k, \operatorname{Div} \boldsymbol{z}).\end{aligned}$$

Step b: Given $\boldsymbol{u}^k$ and $\boldsymbol{v}^{k+1}$, we solve for $\boldsymbol{p}^{k+1}$, such that for all $\boldsymbol{q} \in \boldsymbol{P}$ there holds

$$\begin{aligned}(S\boldsymbol{p}^{k+1}, \boldsymbol{q}) = &-(\boldsymbol{g}, \boldsymbol{q}) + (S\boldsymbol{p}^k, \boldsymbol{q}) - ((\Lambda_1 + \Lambda_2)\boldsymbol{p}^k, \boldsymbol{q})) - (\underline{\operatorname{Div}}\, \boldsymbol{u}^k, \boldsymbol{q}) \\ &- (\operatorname{Div} \boldsymbol{v}^{k+1}, \boldsymbol{q}).\end{aligned}$$

Step c: Given $\boldsymbol{p}^{k+1}$ and $\boldsymbol{v}^{k+1}$, we solve for $\boldsymbol{u}^{k+1}$, such that for all $\boldsymbol{w} \in \boldsymbol{U}$ there holds

$$(\boldsymbol{\epsilon}(\boldsymbol{u}^{k+1}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div} \boldsymbol{u}^{k+1}, \operatorname{div} \boldsymbol{w}) = (\boldsymbol{f}, \boldsymbol{w}) + (\boldsymbol{p}^{k+1}, \underline{\operatorname{Div}}\, \boldsymbol{w}).$$

2.4. *Preliminary results*

We first present a result from linear algebra which will be useful in the proof of Theorem 3.2 in Sec. 3.

Lemma 2.1. *For any $a > 0$ and $b > 0$, denote $\boldsymbol{e} = (1, \ldots, 1)^T \in \mathbb{R}^n$ and let $(aI_{n\times n} + b\boldsymbol{e}\boldsymbol{e}^T)^{-1} = (b_{ij})_{n\times n}$. Then we have that*

$$0 < \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} = \frac{n}{(a+nb)}. \tag{2.16}$$

Proof. The proof is based on the Sherman–Morrison–Woodbury formula and follows the arguments of the proof of Lemma 1 in Ref. 27. □

Next, let us recall some well-known results.[15,18]

Lemma 2.2. *There exists a constant $\beta_s > 0$ such that*

$$\inf_{(q_1,\ldots,q_n)\in P_1\times\cdots\times P_n} \sup_{\boldsymbol{u}\in\boldsymbol{U}} \frac{(\operatorname{div}\boldsymbol{u}, \sum_{i=1}^n q_i)}{\|\boldsymbol{u}\|_1 \, \|\sum_{i=1}^n q_i\|} \geq \beta_s. \tag{2.17}$$

Lemma 2.3. *There exists a constant $\beta_d > 0$ such that*

$$\inf_{q\in P_i} \sup_{\boldsymbol{v}\in\boldsymbol{V}_i} \frac{(\operatorname{div}\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_{\operatorname{div}}\|q\|} \geq \beta_d, \quad i = 1, \ldots, n. \tag{2.18}$$

Here, $\|\cdot\|_1$ and $\|\boldsymbol{v}\|_{\operatorname{div}}$ denote the standard H^1 and $H(\operatorname{div})$ norms of vector-valued functions, respectively, i.e. $\|\boldsymbol{u}\|_1^2 := \int_\Omega \nabla\boldsymbol{u} : \nabla\boldsymbol{u} + \boldsymbol{u}\cdot\boldsymbol{u}\, dx$ and $\|\boldsymbol{v}\|_{\operatorname{div}}^2 := \int_\Omega \operatorname{div}\boldsymbol{v}\operatorname{div}\boldsymbol{v} + \boldsymbol{v}\cdot\boldsymbol{v}\, dx$.

Our task will be to study the errors

$$\boldsymbol{e}_u^k = \boldsymbol{u}^k - \boldsymbol{u} \in \boldsymbol{U}, \tag{2.19a}$$

$$\boldsymbol{e}_{v_i}^k = \boldsymbol{v}_i^k - \boldsymbol{v}_i \in \boldsymbol{V}_i, \quad i = 1, \ldots, n, \tag{2.19b}$$

$$e_{p_i}^k = p_i^k - p_i \in P_i, \quad i = 1, \ldots, n, \tag{2.19c}$$

of the kth iterates $\boldsymbol{u}^k$, $\boldsymbol{v}_i^k$, p_i^k, $i = 1, \ldots, n$, generated by Algorithm 1. For that reason, we consider the following error equations:

$$\begin{aligned}(A_v\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{z}) - (\boldsymbol{e}_{\boldsymbol{p}}^k, \operatorname{Div}\boldsymbol{z}) + (M\underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \operatorname{Div}\boldsymbol{z}) + (M\operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \operatorname{Div}\boldsymbol{z})\\ + (M(\Lambda_1+\Lambda_2)\boldsymbol{e}_{\boldsymbol{p}}^k, \operatorname{Div}\boldsymbol{z}) = 0,\end{aligned} \tag{2.20a}$$

$$\begin{aligned}(S\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \boldsymbol{q}) - (S\boldsymbol{e}_{\boldsymbol{p}}^k, \boldsymbol{q}) + (\underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \boldsymbol{q}) + (\operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{q})\\ +((\Lambda_1+\Lambda_2)\boldsymbol{e}_{\boldsymbol{p}}^k, \boldsymbol{q}) = 0,\end{aligned} \tag{2.20b}$$

$$(\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1}), \boldsymbol{\epsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{e}_{\boldsymbol{u}}^{k+1}, \operatorname{div}\boldsymbol{w}) - (\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \underline{\operatorname{Div}}\,\boldsymbol{w}) = 0, \tag{2.20c}$$

where the error block-vectors $\boldsymbol{e}_v^k$ and $\boldsymbol{e}_p^k$ are given by $(\boldsymbol{e}_{\boldsymbol{v}}^k)^T = ((\boldsymbol{e}_{v_1}^k)^T, \ldots, (\boldsymbol{e}_{v_n}^k))^T$, $(\boldsymbol{e}_{\boldsymbol{p}}^k)^T = (e_{p_1}^k, \ldots, e_{p_n}^k)$.

To complete the design of Algorithm 1, we need to specify M and S. By Lemma 2.3, we have that for all $\boldsymbol{e}_{p_i}^{k+1} \in P_i$ there exists $\boldsymbol{\psi}_i \in \boldsymbol{V}_i$ such that $\operatorname{div} \boldsymbol{\psi}_i = \boldsymbol{e}_{p_i}^{k+1}$ and $\|\boldsymbol{\psi}_i\|_{\mathrm{div}} \leq \beta_d^{-1} \|\boldsymbol{e}_{p_i}^{k+1}\|$ for all $i = 1, \ldots, n$, i.e. $\operatorname{Div} \boldsymbol{\psi} = \boldsymbol{e}_{\boldsymbol{p}}^{k+1}$ and $\|\boldsymbol{\psi}\|_{\mathrm{div}} \leq \beta_v^{-1} \|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|$. Setting $\boldsymbol{q} = S^{-1} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}$ in (2.20b) and $\boldsymbol{z} = \boldsymbol{\psi}$ in (2.20a), from $\operatorname{Div} \boldsymbol{\psi} = \boldsymbol{e}_{\boldsymbol{p}}^{k+1}$ it follows that

$$
\begin{aligned}
&(A_v \boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{\psi}) - (\boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) + (M \underline{\operatorname{Div}}\, \boldsymbol{e}_{\boldsymbol{u}}^{k+1}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) + (M \operatorname{Div} \boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) \\
&\quad + (M(\Lambda_1 + \Lambda_2) \boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) = 0, \qquad (2.21a) \\
&(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) - (\boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) + (S^{-1} \underline{\operatorname{Div}}\, \boldsymbol{e}_{\boldsymbol{u}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) + (S^{-1} \operatorname{Div} \boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) \\
&\quad + (S^{-1}(\Lambda_1 + \Lambda_2) \boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) = 0. \qquad (2.21b)
\end{aligned}
$$

Subtracting (2.21a) from (2.21b) yields

$$
\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 = (A_v \boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{\psi}) - ((S^{-1} - M)(\underline{\operatorname{Div}}\, \boldsymbol{e}_{\boldsymbol{u}}^{k} + \operatorname{Div} \boldsymbol{e}_{\boldsymbol{v}}^{k+1} + (\Lambda_1 + \Lambda_2) \boldsymbol{e}_{\boldsymbol{p}}^{k}), \boldsymbol{e}_{\boldsymbol{p}}^{k+1}),
$$

implying

$$
\begin{aligned}
\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 &\leq \|A_v^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{v}}^{k+1}\| \|A_v^{\frac{1}{2}} \boldsymbol{\psi}\| \\
&\quad + \|(S^{-1} - M)(\underline{\operatorname{Div}}\, \boldsymbol{e}_{\boldsymbol{u}}^{k+1} + \operatorname{Div} \boldsymbol{e}_{\boldsymbol{v}}^{k+1} + (\Lambda_1 + \Lambda_2) \boldsymbol{e}_{\boldsymbol{p}}^{k})\| \|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\| \\
&\leq \sqrt{R^{-1}} \|A_v^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{v}}^{k+1}\| \|\boldsymbol{\psi}\| \\
&\quad + \|(S^{-1} - M)(\underline{\operatorname{Div}}\, \boldsymbol{e}_{\boldsymbol{u}}^{k} + \operatorname{Div} \boldsymbol{e}_{\boldsymbol{v}}^{k+1} + (\Lambda_1 + \Lambda_2) \boldsymbol{e}_{\boldsymbol{p}}^{k})\| \|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\| \\
&\leq \beta_d^{-1} \sqrt{R^{-1}} \|A_v^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{v}}^{k+1}\| \|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\| + \|(S^{-1} - M)(\underline{\operatorname{Div}} \boldsymbol{e}_{\boldsymbol{u}}^{k} + \operatorname{Div} \boldsymbol{e}_{\boldsymbol{v}}^{k+1} \\
&\quad + (\Lambda_1 + \Lambda_2) \boldsymbol{e}_{\boldsymbol{p}}^{k})\| \|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|.
\end{aligned}
$$

We conclude that

$$
\begin{aligned}
\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\| &\leq \beta_d^{-1} \sqrt{R^{-1}} \|A_v^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{v}}^{k+1}\| \\
&\quad + \|(S^{-1} - M)(\underline{\operatorname{Div}} \boldsymbol{e}_{\boldsymbol{u}}^{k} + \operatorname{Div} \boldsymbol{e}_{\boldsymbol{v}}^{k+1} + (\Lambda_1 + \Lambda_2) \boldsymbol{e}_{\boldsymbol{p}}^{k})\|. \qquad (2.22)
\end{aligned}
$$

Estimate (2.22) suggests choosing $S = M^{-1}$ in order to minimize the upper bound for $\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|$. This results in the following statement.

Lemma 2.4. *Consider Algorithm* 1 *and let* $S = M^{-1}$, *then we have*

$$
\|A_v^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{v}}^{k+1}\|^2 \geq R \beta_d^2 \|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 = \beta_d^2 \|\Lambda_3^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2. \qquad (2.23)
$$

The relationship $S = M^{-1}$ reduces our design task to the determination of either S or M. In the remainder of this paper, we analyze and numerically test Algorithm 1 for the specific choice

$$
S := \Lambda_1 + \Lambda_2 + L_1 \Lambda_3 + L_2 \Lambda_4, \qquad (2.24)
$$

where L_1 and L_2 are scalar parameters which are later to be determined.

3. Convergence Theory of Uzawa-Type Algorithms for MPET

This section is devoted to the convergence analysis of Algorithm 1. Our aim is to establish a uniform bound on the convergence rate, i.e. a bound independent of any model and discretization parameters.

We start with deriving some useful auxiliary results presented in the following two lemmas. These afterwards assist us in establishing a parameter-robust upper bound on the pressure error in a weighted norm.

Lemma 3.1. *Considering Algorithm* 1 *with* S *as defined in* (2.24), *the errors* $\boldsymbol{e}_{\boldsymbol{u}}^k$, $\boldsymbol{e}_{\boldsymbol{v}}^k$, *and* $\boldsymbol{e}_{\boldsymbol{p}}^k$ *defined in* (2.19) *satisfy the following estimate*:

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \frac{\lambda}{2}\|\operatorname{div}\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2 + \|A_v^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 \\
&\quad + \frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 \le \frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 + \frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 \\
&\quad + \left(\frac{\lambda_0}{2(c_K^2+\lambda)} - \frac{L_2}{2} - \frac{L_1 R\lambda_0}{2n}\right)\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2. \qquad (3.1)
\end{aligned}$$

Proof. By setting $\boldsymbol{q} = M\operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}$ in (2.20b) and $\boldsymbol{z} = \boldsymbol{e}_{\boldsymbol{v}}^{k+1}$ in (2.20a) we obtain

$$\begin{aligned}
&(A_v\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{v}}^{k+1}) - (\boldsymbol{e}_{\boldsymbol{p}}^{k}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) + (M\underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \\
&\quad + (M\operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) + (M(\Lambda_1+\Lambda_2)\boldsymbol{e}_{\boldsymbol{p}}^{k}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) = 0, \\
&(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) = (\boldsymbol{e}_{\boldsymbol{p}}^{k}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) - (M\underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \\
&\quad - (M\operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) - (M(\Lambda_1+\Lambda_2)\boldsymbol{e}_{\boldsymbol{p}}^{k}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1})
\end{aligned}$$

from where it immediately follows that

$$(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) = (A_v\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{v}}^{k+1}). \qquad (3.2)$$

Choosing $\boldsymbol{q} = \boldsymbol{e}_{\boldsymbol{p}}^{k+1}$ in (2.20b) and $\boldsymbol{w} = \boldsymbol{e}_{\boldsymbol{u}}^{k+1}$ in (2.20c) yields

$$(\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1}), \boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})) + \lambda(\operatorname{div}\boldsymbol{e}_{\boldsymbol{u}}^{k+1}, \operatorname{div}\boldsymbol{e}_{\boldsymbol{u}}^{k+1}) - (\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}) = 0, \qquad (3.3a)$$

$$\begin{aligned}
&(S\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) = (S\boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) - (\underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) - (\operatorname{Div}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) \\
&\quad - ((\Lambda_1+\Lambda_2)\boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}). \qquad (3.3b)
\end{aligned}$$

Next, summing (3.3a) and (3.3b) and applying (3.2) it follows that

$$\begin{aligned}
&\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda\|\operatorname{div}\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2 + \|S^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 - ((L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) \\
&\quad = (\underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \underline{\operatorname{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) - \|A_v^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}\|^2. \qquad (3.4)
\end{aligned}$$

In order to simplify (3.4) we first rewrite $\|S^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 - ((L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1})$, that is,

$$\begin{aligned}
&\|S^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 - ((L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1}) \\
&\quad = \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \frac{L_1}{2}(\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 - \|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 + \|\Lambda_3^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1} - \boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2) \\
&\qquad + \frac{L_2}{2}\big(\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 - \|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 + \|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1} - \boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2\big) \\
&\quad \geq \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 - \frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 \\
&\qquad - \frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 + \left(\frac{L_2}{2} + \frac{L_1 R\lambda_0}{2n}\right)\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1} - \boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2. \qquad (3.5)
\end{aligned}$$

Second, we estimate $(\underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^{k}, \boldsymbol{e}_{\boldsymbol{p}}^{k+1})$. By setting $\boldsymbol{w} = \boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k}$ in (2.20c) we obtain

$$\begin{aligned}
(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}, \underline{\mathrm{Div}}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})) &= (\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k}), \boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})) + \lambda(\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k}), \mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}) \\
&\leq \frac{1}{2}(\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2 + \lambda\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2) \\
&\quad + \frac{1}{2}(\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda\|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2). \qquad (3.6)
\end{aligned}$$

In order to estimate the right-hand side of (3.6), we subtract the kth error from the $(k+1)$th error and choose $\boldsymbol{w} = \boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k}$ in (2.20c) and herewith obtaining

$$\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2 + \lambda\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2 = \left(\sum_{i=1}^{n}(\boldsymbol{e}_{p_i}^{k+1} - \boldsymbol{e}_{p_i}^{k}), \mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\right).$$

Applying Cauchy's inequality further yields

$$\begin{aligned}
&\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2 + \lambda\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2 \\
&\quad = \left(\sum_{i=1}^{n}(\boldsymbol{e}_{p_i}^{k+1} - \boldsymbol{e}_{p_i}^{k}), \mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\right) \\
&\quad \leq \left\|\sum_{i=1}^{n}(\boldsymbol{e}_{p_i}^{k+1} - \boldsymbol{e}_{p_i}^{k})\right\| \|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\| \\
&\quad = \sqrt{\lambda_0}\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1} - \boldsymbol{e}_{\boldsymbol{p}}^{k})\|\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|. \qquad (3.7)
\end{aligned}$$

Noting that

$$(c_K^2 + \lambda)\|\mathrm{div}\,\boldsymbol{w}\|^2 \leq \|\boldsymbol{\epsilon}(\boldsymbol{w})\|^2 + \lambda\|\mathrm{div}\,\boldsymbol{w}\|^2, \qquad (3.8)$$

which follows from (2.12), we directly obtain

$$(c_K^2 + \lambda)\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2 \leq \sqrt{\lambda_0}\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1} - \boldsymbol{e}_{\boldsymbol{p}}^{k})\|\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1} - \boldsymbol{e}_{\boldsymbol{u}}^{k})\|,$$

from (3.7). The latter estimate implies

$$\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1}-\boldsymbol{e}_{\boldsymbol{u}}^{k})\| \leq \frac{\sqrt{\lambda_0}}{c_K^2+\lambda}\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|.$$

By using the above inequality in (3.7), it follows that

$$\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1}-\boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2+\lambda\|\mathrm{div}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1}-\boldsymbol{e}_{\boldsymbol{u}}^{k})\|^2 \leq \tfrac{\lambda_0}{c_K^2+\lambda}\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2. \tag{3.9}$$

Now, combining (3.6) and (3.9) yields

$$\begin{aligned}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1},\underline{\mathrm{Div}}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1}-\boldsymbol{e}_{\boldsymbol{u}}^{k})) \leq{}& \frac{\lambda_0}{2(c_K^2+\lambda)}\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2\\ &+\frac{1}{2}(\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2+\lambda\|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2).\end{aligned} \tag{3.10}$$

Finally, inserting (3.5) and (3.10) in (3.4) we have that

$$\begin{aligned}&\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2+\lambda\|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2+\|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\\ &\quad+\frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\\ &\leq \frac{\lambda_0}{2(c_K^2+\lambda)}\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2+\frac{1}{2}(\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2+\lambda\|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2)\\ &\quad-\|A_v^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{v}}^{k+1}\|^2+\frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2+\frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2\\ &\quad-\left(\frac{L_2}{2}+\frac{L_1R\lambda_0}{2n}\right)\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2\end{aligned}$$

which shows (3.1). □

The next lemma provides a preliminary estimate for the pressure errors.

Lemma 3.2. *Consider Algorithm* 1 *with* S *as in* (2.24)*. Then the errors* $\boldsymbol{e}_{\boldsymbol{p}}^{k}$ *defined in* (2.19) *satisfy*

$$\begin{aligned}&\frac{\lambda_0}{2(\beta_s^{-2}+\lambda)}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\\ &\quad+\frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\\ &\leq \frac{L_1}{2}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2+\frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2\\ &\quad+\left(\frac{\lambda_0}{2(c_K^2+\lambda)}-\frac{L_2}{2}-\frac{L_1R\lambda_0}{2n}\right)\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2.\end{aligned}$$

Proof. From Lemma 2.2, it follows that for all $\sum_{i=1}^n \boldsymbol{e}_{p_i}^{k+1} \in P_i$ there exists $\boldsymbol{w}_0 \in \boldsymbol{U}$ such that $\operatorname{div} \boldsymbol{w}_0 = \frac{1}{\sqrt{\lambda_0}} \sum_{i=1}^n \boldsymbol{e}_{p_i}^{k+1}$ and $\|\boldsymbol{w}_0\|_1 \leq \beta_s^{-1} \frac{1}{\sqrt{\lambda_0}} \| \sum_{i=1}^n \boldsymbol{e}_{p_i}^{k+1} \| = \beta_s^{-1} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|$. Also,

$$\underline{\operatorname{Div}}\, \boldsymbol{w}_0 = \begin{pmatrix} \frac{1}{\sqrt{\lambda_0}} \sum_{i=1}^n \boldsymbol{e}_{p_i}^{k+1} \\ \vdots \\ \frac{1}{\sqrt{\lambda_0}} \sum_{i=1}^n \boldsymbol{e}_{p_i}^{k+1} \end{pmatrix} = \sqrt{\lambda_0} \Lambda_4 \boldsymbol{e}_{\boldsymbol{p}}^{k+1}.$$

Setting $\boldsymbol{w} = \boldsymbol{w}_0$ in (2.20c), it follows that

$$\begin{aligned} \sqrt{\lambda_0} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 &= (\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1}), \boldsymbol{\epsilon}(\boldsymbol{w}_0)) + \lambda (\operatorname{div} \boldsymbol{e}_{\boldsymbol{u}}^{k+1}, \operatorname{div} \boldsymbol{w}_0) \\ &\leq (\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda \|\operatorname{div} \boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2)^{\frac{1}{2}} (\|\boldsymbol{\epsilon}(\boldsymbol{w}_0)\|^2 + \lambda \|\operatorname{div} \boldsymbol{w}_0\|^2)^{\frac{1}{2}} \\ &\leq (\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda \|\operatorname{div} \boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2)^{\frac{1}{2}} (\beta_s^{-2} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \lambda \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2)^{\frac{1}{2}} \\ &= (\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda \|\operatorname{div} \boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2)^{\frac{1}{2}} (\beta_s^{-2} + \lambda)^{\frac{1}{2}} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\| \end{aligned}$$

and, therefore,

$$\frac{\lambda_0}{\beta_s^{-2} + \lambda} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 \leq \|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda \|\operatorname{div} \boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2. \tag{3.11}$$

Using (3.11) and (2.23) in (3.1), we have

$$\begin{aligned} &\frac{\lambda_0}{2(\beta_s^{-2} + \lambda)} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \beta_d^2 \|\Lambda_3^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \|(\Lambda_1 + \Lambda_2)^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \frac{L_1}{2} \|\Lambda_3^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 \\ &\quad + \frac{L_2}{2} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 \leq \frac{L_1}{2} \|\Lambda_3^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 + \frac{L_2}{2} \|\Lambda_4^{\frac{1}{2}} \boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 \\ &\quad + \left(\frac{\lambda_0}{2(c_K^2 + \lambda)} - \frac{L_2}{2} - \frac{L_1 R \lambda_0}{2n} \right) \|\Lambda_4^{\frac{1}{2}} (\boldsymbol{e}_{\boldsymbol{p}}^{k+1} - \boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2, \end{aligned} \tag{3.12}$$

which completes the proof. □

The following two theorems present the main convergence results for Algorithm 1.

Theorem 3.1. *Consider Algorithm* 1. *For any* $\theta > 0$ *and* $L_2 \geq \frac{\lambda_0}{(c_K^2 + \lambda)(1 + \frac{\theta \beta_d^2 R \lambda_0}{n})}$, $L_1 = \theta \beta_d^2 L_2$, *the errors* $\boldsymbol{e}_{\boldsymbol{p}}^k$ *defined in* (2.19) *satisfy the estimate*:

$$\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|_{\boldsymbol{P}_\theta}^2 \leq \operatorname{rate}^2(\lambda, R, \theta) \|\boldsymbol{e}_{\boldsymbol{p}}^{k}\|_{\boldsymbol{P}_\theta}^2 \tag{3.13}$$

with

$$\operatorname{rate}^2(\lambda, R, \theta) \leq \frac{1}{C+1}, \quad C := \min\left\{ \frac{\lambda_0}{\beta_s^{-2} + \lambda}, 2\theta^{-1} \right\} L_2^{-1}$$

and

$$\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|_{\boldsymbol{P}_\theta}^2 := \|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2. \tag{3.14}$$

(1) *For* $\theta = \theta_0 := \beta_d^{-2}$ *and* $L_2 = \frac{\lambda_0}{(c_K^2+\lambda)(1+\frac{R\lambda_0}{n})}$, *we obtain the convergence factor under the norm* $\|\cdot\|_{\boldsymbol{P}_{\theta_0}}$ *estimated by*

$$\mathrm{rate}^2(\lambda,R) \le \frac{1}{\frac{c_0(c_K^2+\lambda)\left(1+\frac{\lambda_0 R}{n}\right)}{\lambda_0}+1} \le \max\left\{\frac{1}{c_0+1}, \frac{1}{c_0c_K^2+1}, \frac{1}{2}\right\},$$

where $c_0 = \min\{\frac{\lambda_0}{\beta_s^{-2}+\lambda}, 2\beta_d^2\}$. *Here, for any* $\boldsymbol{x}\in\boldsymbol{P}$

$$\|\boldsymbol{x}\|_{\boldsymbol{P}_{\theta_0}}^2 := \|\Lambda_3^{\frac{1}{2}}\boldsymbol{x}\|^2 + \|\Lambda_4^{\frac{1}{2}}\boldsymbol{x}\|^2 + \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{x}\|^2.$$

(2) *For the best choice* $\theta = \theta_* := \frac{2(\beta_s^{-2}+\lambda)}{\lambda_0}$ *and* $L_2 = \frac{\lambda_0}{(c_K^2+\lambda)(1+\frac{2\beta_d^2(\beta_s^{-2}+\lambda)R}{n})}$, *the errors* $\boldsymbol{e}_{\boldsymbol{p}}^k$ *satisfy the estimate*

$$\begin{aligned}\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|_{\boldsymbol{P}_{\theta_*}}^2 &\le \mathrm{rate}^2(\lambda,R) \le \frac{1}{\frac{(c_K^2+\lambda)\left(1+\frac{2\beta_d^2(\beta_s^{-2}+\lambda)R}{n}\right)}{(\beta_s^{-2}+\lambda)}+1}\\ &\le \max\left\{\frac{\beta_s^{-2}}{c_K^2+\beta_s^{-2}}, \frac{1}{2}\right\},\end{aligned} \tag{3.15}$$

where

$$\begin{aligned}\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|_{\boldsymbol{P}_{\theta_*}}^2 := \|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 &+ \frac{2(\beta_s^{-2}+\lambda)\beta_v^2}{\lambda_0}\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\\ &+ \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2.\end{aligned} \tag{3.16}$$

Proof. In view of the estimate presented in Lemma 3.2, we want to find L_2 and L_1 subject to the condition

$$\frac{\lambda_0}{2(c_K^2+\lambda)} - \frac{L_2}{2} - \frac{L_1R\lambda_0}{2n} \le 0. \tag{3.17}$$

For any $\theta > 0$, we rewrite (3.12) as

$$\begin{aligned}&\frac{\lambda_0}{2(\beta_s^{-2}+\lambda)}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \theta^{-1}\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\\ &\quad + \frac{L_1}{2\theta\beta_d^2}\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 + \frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 \le \frac{L_1}{2\theta\beta_d^2}\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2 + \frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2\\ &\quad + \left(\frac{\lambda_0}{2(c_K^2+\lambda)} - \frac{L_2}{2} - \frac{L_1R\lambda_0}{2n}\right)\|\Lambda_4^{\frac{1}{2}}(\boldsymbol{e}_{\boldsymbol{p}}^{k+1}-\boldsymbol{e}_{\boldsymbol{p}}^{k})\|^2,\end{aligned} \tag{3.18}$$

namely,

$$\left(\frac{\lambda_0}{2(\beta_s^{-2}+\lambda)}+\frac{L_2}{2}\right)\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\left(\theta^{-1}+\frac{L_1}{2\theta\beta_d^2}\right)\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2$$
$$+\|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\leq\frac{L_1}{2\theta\beta_d^2}\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2+\frac{L_2}{2}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2. \tag{3.19}$$

Then, for $L_2\leq 1$ we obtain

$$\begin{aligned}&\min\left\{\frac{\lambda_0}{2(\beta_s^{-2}+\lambda)}+\frac{L_2}{2},\theta^{-1}+\frac{L_1}{2\theta\beta_d^2}\right\}\\&\quad\times\big(\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\big)\\&\leq\max\left\{\frac{L_1}{2\theta\beta_d^2},\frac{L_2}{2}\right\}\big(\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2+\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2\big).\end{aligned} \tag{3.20}$$

Now, choose $L_1=\theta\beta_d^2L_2$. Then, condition (3.17) becomes

$$\frac{\lambda_0}{2(c_K^2+\lambda)}-\frac{L_2}{2}-\frac{\theta\beta_d^2L_2R\lambda_0}{2n}\leq 0\quad\text{or}\quad L_2\geq\frac{\frac{\lambda_0}{(c_K^2+\lambda)}}{1+\frac{\theta\beta_d^2R\lambda_0}{n}} \tag{3.21}$$

and we can simplify (3.20) as follows:

$$\begin{aligned}&\min\left\{\frac{\lambda_0}{2(\beta_s^{-2}+\lambda)}+\frac{L_2}{2},\theta^{-1}+\frac{L_2}{2}\right\}\\&\quad\times\big(\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2+\|(\Lambda_1+\Lambda_2)^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2\big)\\&\leq\frac{L_2}{2}\big(\theta\beta_d^2\|\Lambda_3^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2+\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k}\|^2\big),\end{aligned}$$

which shows (3.13). Statements (1) and (2) are direct consequences of (3.13) for the particular choices of θ in the corresponding norms. □

Note that estimate (3.15) not only indicates that the convergence rate of the Uzawa-type iterative method has a uniform, with respect to the parameters, upper-bound being strictly less than 1, but also that it is bounded by a number much smaller than 1 if λ is large. Moreover, the presented analysis of the Uzawa-type scheme results in a new, parameter-optimized block-triangular preconditioner that can be used to accelerate the convergence of the GMRES method if the latter is applied to the augmented system (2.13). The parameter-robust uniform convergence estimates for the new Uzawa-type method imply the field-of-values equivalence of this preconditioner for the augmented system.

Theorem 3.2. *Consider Algorithm* 1 *with* S *as introduced in* (2.24)*. Then the errors* $\boldsymbol{e}_{\boldsymbol{u}}^k$ *and* $\boldsymbol{e}_{\boldsymbol{v}}^k$ *defined in* (2.19) *satisfy the estimates*:

$$\|\boldsymbol{e}_{\boldsymbol{u}}^k\|_{\boldsymbol{U}}\leq C_u[\mathrm{rate}(\lambda,R)]^k,\quad\|\boldsymbol{e}_{\boldsymbol{v}}^k\|_{\boldsymbol{V}_{\theta_*}}\leq C_v[\mathrm{rate}(\lambda,R)]^k, \tag{3.22}$$

where

$$\|\boldsymbol{e}_{\boldsymbol{v}}^k\|_{\boldsymbol{V}_{\theta_*}}^2 := (A_v\boldsymbol{e}_{\boldsymbol{v}}^k, \boldsymbol{e}_{\boldsymbol{v}}^k) + (S^{-1}\mathrm{Div}\boldsymbol{e}_{\boldsymbol{v}}^k, \mathrm{Div}\boldsymbol{e}_{\boldsymbol{v}}^k), \quad \|\boldsymbol{u}\|_{\boldsymbol{U}}^2 := \|\boldsymbol{\epsilon}(\boldsymbol{u})\|^2 + \lambda\|\mathrm{div}\boldsymbol{u}\|^2 \quad (3.23)$$

and the constants C_u and C_v are independent of the model parameters and the time step size.

Proof. First, we estimate $\|\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|_{\boldsymbol{U}}^2$. By setting $\boldsymbol{w} = \boldsymbol{e}_{\boldsymbol{u}}^{k+1}$ in (2.20c), applying Cauchy's inequality and using (3.8) we obtain

$$\begin{aligned}\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda\|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2 &= \left(\sum_{i=1}^n \boldsymbol{e}_{p_i}^{k+1}, \mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\right) \le \left\|\sum_{i=1}^n \boldsymbol{e}_{p_i}^{k+1}\right\| \cdot \|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\| \\ &= \sqrt{\lambda_0}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|\|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\| \\ &\le \sqrt{\lambda_0}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|\sqrt{\frac{1}{c_K^2+\lambda}}(\|\boldsymbol{\epsilon}(\boldsymbol{e}_{\boldsymbol{u}}^{k+1})\|^2 + \lambda\|\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|^2),\end{aligned}$$

or, equivalently,

$$\|\boldsymbol{e}_{\boldsymbol{u}}^{k+1}\|_{\boldsymbol{U}}^2 \le \frac{\lambda_0}{c_K^2+\lambda}\|\Lambda_4^{\frac{1}{2}}\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|^2 \le \frac{\lambda_0}{c_K^2+\lambda}\|\boldsymbol{e}_{\boldsymbol{p}}^{k+1}\|_{\boldsymbol{P}_\theta}^2. \quad (3.24)$$

In order to estimate $\|\boldsymbol{e}_{\boldsymbol{v}}^{k+1}\|_{\boldsymbol{V}_{\theta_*}}^2$ we set $\boldsymbol{z} = \boldsymbol{e}_{\boldsymbol{v}}^{k+1}$ in (2.20a) and apply the Cauchy inequality to derive

$$\begin{aligned}&(A_v\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{v}}^{k+1}) + (S^{-1}\mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \\ &\quad= (\boldsymbol{e}_{\boldsymbol{p}}^k, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) - (S^{-1}\underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \\ &\qquad- (S^{-1}(\Lambda_1+\Lambda_2)\boldsymbol{e}_{\boldsymbol{p}}^k, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \\ &\quad= (S^{-1}(L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) - (S^{-1}\underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \\ &\quad\le (S^{-1}(L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k, (L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k) + \frac{1}{4}(S^{-1}\mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \\ &\qquad+ (S^{-1}\underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k) + \frac{1}{4}(S^{-1}\mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}). \quad (3.25)\end{aligned}$$

From the definition of S, see (2.24), that of $\|\cdot\|_{\boldsymbol{P}_{\theta_*}}$, see (3.16), and noting that $L_1 = \theta\beta_d^2 L_2$, see Theorem 3.1, we have

$$\begin{aligned}&(S^{-1}(L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k, (L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k) \\ &\qquad\le ((L_1\Lambda_3 + L_2\Lambda_4)^{-1}(L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k, (L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k) \\ &\qquad= ((L_1\Lambda_3 + L_2\Lambda_4)\boldsymbol{e}_{\boldsymbol{p}}^k, \boldsymbol{e}_{\boldsymbol{p}}^k) \le L_2\|\boldsymbol{e}_{\boldsymbol{p}}^k\|_{\boldsymbol{P}_{\theta_*}}^2. \quad (3.26)\end{aligned}$$

Then (3.25) can be rewritten in the form

$$(A_v\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{v}}^{k+1}) + \frac{1}{2}(S^{-1}\mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}) \le L_2\|\boldsymbol{e}_{\boldsymbol{p}}^k\|_{\boldsymbol{P}_{\theta_*}}^2 + \|S^{-\frac{1}{2}}\underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k\|^2.$$

Again, from the definition of S, and noting that $L_1\Lambda_3 + L_2\Lambda_4 = (L_1 R I_{n\times n} + \frac{L_2}{\lambda_0}\boldsymbol{e}\boldsymbol{e}^T)$, by choosing $a = L_1 R$ and $b = \frac{L_2}{\lambda_0}$ in Lemma 2.1, it follows that

$$\begin{aligned}\|S^{-\frac{1}{2}}\underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k\|^2 &\le ((L_1\Lambda_3 + L_2\Lambda_4)^{-1}\underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \underline{\mathrm{Div}}\,\boldsymbol{e}_{\boldsymbol{u}}^k)\\ &= \left(\left(\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}\right)\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k\right)\\ &= \frac{n\lambda_0}{L_1 R\lambda_0 + nL_2}(\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k) \le (c_K^2 + \lambda)(\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k).\end{aligned}$$

Therefore, from (3.24), we have

$$\begin{aligned}\|\boldsymbol{e}_{\boldsymbol{v}}^k\|_{\boldsymbol{V}_{\theta_*}}^2 &= (A_v\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \boldsymbol{e}_{\boldsymbol{v}}^{k+1}) + \frac{1}{2}(S^{-1}\mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1}, \mathrm{Div}\,\boldsymbol{e}_{\boldsymbol{v}}^{k+1})\\ &\le L_2\|\boldsymbol{e}_{\boldsymbol{p}}^k\|_{\boldsymbol{P}_{\theta_*}}^2 + (c_K^2 + \lambda)(\mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k, \mathrm{div}\,\boldsymbol{e}_{\boldsymbol{u}}^k) \le L_2\|\boldsymbol{e}_{\boldsymbol{p}}^k\|_{\boldsymbol{P}_{\theta_*}}^2 + \|\boldsymbol{e}_{\boldsymbol{u}}^k\|_U^2\\ &\le L_2\|\boldsymbol{e}_{\boldsymbol{p}}^k\|_{\boldsymbol{P}_{\theta_*}}^2 + \frac{\lambda_0}{c_K^2 + \lambda}\|\boldsymbol{e}_{\boldsymbol{p}}^k\|_{\boldsymbol{P}_{\theta_*}}^2\\ &= \left(L_2 + \frac{\lambda_0}{c_K^2 + \lambda}\right)\|\boldsymbol{e}_{\boldsymbol{p}}^k\|_{\boldsymbol{P}_{\theta_*}}^2,\end{aligned}$$

which completes the proof. □

Remark 3.1. Note that for the particular choice of S and M in this section, the block-triangular matrix on the left-hand side of (2.15) provides a field-of-values-equivalent preconditioner with equivalence constants independent of any model and discretization parameters.

4. The Discrete MPET Problem

Mass conservative discretizations for the MPET model are considered in this section, cf. Refs. 26 and 27. The analysis here can also be utilized for other stable discretizations of the three-field formulation of the MPET model.[30,46]

4.1. *Notation*

Let $\mathcal{T}_h$ be a shape-regular triangulation of the domain Ω into triangles/tetrahedrons where the subscript h denotes the mesh-size. Furthermore, let $\mathcal{E}_h^I$ and $\mathcal{E}_h^B$ define the set of all interior edges/faces and the set of all boundary edges/faces of $\mathcal{T}_h$, respectively, with their union being written as $\mathcal{E}_h$.

For $s \ge 1$ we introduce the broken Sobolev spaces

$$H^s(\mathcal{T}_h) = \{\phi \in L^2(\Omega), \text{ such that } \phi|_T \in H^s(T) \text{ for all } T \in \mathcal{T}_h\}.$$

Define T_1 and T_2 to be two elements from the triangulation which share an edge or face e and $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ to be the corresponding unit normal vectors to e which

point to the exterior of T_1 and T_2. For $q \in H^1(\mathcal{T}_h)$, $\boldsymbol{v} \in H^1(\mathcal{T}_h)^d$ and $\boldsymbol{\tau} \in H^1(\mathcal{T}_h)^{d\times d}$ and any $e \in \mathcal{E}_h^I$, the jumps $[\cdot]$ and averages $\{\cdot\}$ are defined by

$$[q] = q|_{\partial T_1 \cap e} - q|_{\partial T_2 \cap e}, \quad [\boldsymbol{v}] = \boldsymbol{v}|_{\partial T_1 \cap e} - \boldsymbol{v}|_{\partial T_2 \cap e},$$
$$\{\boldsymbol{v}\} = \frac{1}{2}(\boldsymbol{v}|_{\partial T_1 \cap e} \cdot \boldsymbol{n}_1 - \boldsymbol{v}|_{\partial T_2 \cap e} \cdot \boldsymbol{n}_2), \quad \{\boldsymbol{\tau}\} = \frac{1}{2}(\boldsymbol{\tau}|_{\partial T_1 \cap e}\boldsymbol{n}_1 - \boldsymbol{\tau}|_{\partial T_2 \cap e}\boldsymbol{n}_2),$$

whereas in the case of $e \in \mathcal{E}_h^B$, $[q] = q|_e$, $[\boldsymbol{v}] = \boldsymbol{v}|_e$, $\{\boldsymbol{v}\} = \boldsymbol{v}|_e \cdot \boldsymbol{n}$, $\{\boldsymbol{\tau}\} = \boldsymbol{\tau}|_e \boldsymbol{n}$.

4.2. *Mixed finite element spaces and discrete formulation*

So as to discretize the flow equations, a mixed finite element method has been used to approximate fluxes and pressures. The displacement field of the mechanics problem is approximated using a discontinuous Galerkin method. The following finite element spaces are employed:

$$\boldsymbol{U}_h = \{\boldsymbol{u} \in H(\mathrm{div};\Omega) : \boldsymbol{u}|_T \in \boldsymbol{U}(T), T \in \mathcal{T}_h;\ \boldsymbol{u}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\},$$
$$\boldsymbol{V}_{i,h} = \{\boldsymbol{v} \in H(\mathrm{div};\Omega) : \boldsymbol{v}|_T \in \boldsymbol{V}_i(T),\ T \in \mathcal{T}_h;\ \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\}, \quad i = 1,\dots,n,$$
$$P_{i,h} = \left\{q \in L^2(\Omega) : q|_T \in Q_i(T),\ T \in \mathcal{T}_h; \int_\Omega q dx = 0\right\}, \quad i = 1,\dots,n,$$

where $\boldsymbol{V}_i(T)/Q_i(T) = \mathrm{RT}_{l-1}(T)/\mathrm{P}_{l-1}(T)$, $\boldsymbol{U}(T) = \mathrm{BDM}_l(T)$ or $\boldsymbol{U}(T) = \mathrm{BDFM}_l(T)$ for $l \geq 1$. One should note that $\mathrm{div}\,\boldsymbol{U}(T) = \mathrm{div}\,\boldsymbol{V}_i(T) = Q_i(T)$ for each of these choices.

Also, it has been shown in Refs. 26 and 27 that for all $\boldsymbol{u} \in \boldsymbol{U}_h$, $[\boldsymbol{u}_n] = 0$, from which follows that $[\boldsymbol{u}] = [\boldsymbol{u}_t]$. Here, $\boldsymbol{u}_n$ and $\boldsymbol{u}_t$ are the normal and tangential component of $\boldsymbol{u}$, respectively.

Defining $\boldsymbol{v}_h^T = (\boldsymbol{v}_{1,h}^T,\dots,\boldsymbol{v}_{n,h}^T)$, $\boldsymbol{p}_h^T = (p_{1,h},\dots,p_{n,h})$, $\boldsymbol{z}_h^T = (\boldsymbol{z}_{1,h}^T,\dots,\boldsymbol{z}_{n,h}^T)$, $\boldsymbol{q}_h^T = (q_{1,h},\dots,q_{n,h})$, and further $\boldsymbol{V}_h = \boldsymbol{V}_{1,h} \times \cdots \times \boldsymbol{V}_{n,h}$, $\boldsymbol{P}_h = P_{1,h} \times \cdots \times P_{n,h}$, $\boldsymbol{X}_h = \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h$, we consider the following discrete variational problem: Find $(\boldsymbol{u}_h;\boldsymbol{v}_h;\boldsymbol{p}_h) \in \boldsymbol{X}_h$, such that for any $(\boldsymbol{w}_h;\boldsymbol{z}_h;\boldsymbol{q}_h) \in \boldsymbol{X}_h$ and $i = 1,\dots,n$

$$(R_i^{-1}\boldsymbol{v}_{i,h}, \boldsymbol{z}_{i,h}) - (p_{i,h}, \mathrm{div}\,\boldsymbol{z}_{i,h}) = 0, \tag{4.1a}$$
$$-(\mathrm{div}\,\boldsymbol{u}_h, q_{i,h}) - (\mathrm{div}\,\boldsymbol{v}_{i,h}, q_{i.h}) + \tilde{\alpha}_{ii}(p_{i,h}, q_{i,h}) + \sum_{\substack{j=1\\ j\neq i}}^{n} \alpha_{ij}(p_{j,h}, q_{i,h}) = (g_i, q_{i,h}), \tag{4.1b}$$
$$a_h(\boldsymbol{u}_h, \boldsymbol{w}_h) + \lambda(\mathrm{div}\,\boldsymbol{u}_h, \mathrm{div}\,\boldsymbol{w}_h) - \sum_{i=1}^{n}(p_{i,h}, \mathrm{div}\,\boldsymbol{w}_h) = (\boldsymbol{f}, \boldsymbol{w}_h), \tag{4.1c}$$

where

$$a_h(\boldsymbol{u},\boldsymbol{w}) = \sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{e_p}(\boldsymbol{u}) : \boldsymbol{e_p}(\boldsymbol{w})dx - \sum_{e\in\mathcal{E}_h}\int_e \{\boldsymbol{e_p}(\boldsymbol{u})\}\cdot[\boldsymbol{w}_t]ds$$
$$- \sum_{e\in\mathcal{E}_h}\int_e \{\boldsymbol{e_p}(\boldsymbol{w})\}\cdot[\boldsymbol{u}_t]ds + \sum_{e\in\mathcal{E}_h}\int_e \eta h_e^{-1}[\boldsymbol{u}_t]\cdot[\boldsymbol{w}_t]ds, \qquad (4.2)$$

$\tilde{\alpha}_{ii} = -\alpha_{p_i} - \alpha_{ii}$, and η is a stabilization parameter which is independent of λ, R_i^{-1}, α_{p_i}, α_{ij}, $i,j\in\{1,\ldots,n\}$, the network scale n, and the mesh-size h.

In the derivation of the discrete variational problem (4.1), homogeneous Dirichlet boundary conditions for $\boldsymbol{u}$ and homogeneous Neumann boundary conditions for p_i, $i=1,2,\ldots,n$, have been assumed for each case over the entire domain boundary. The DG discretizations for more general (rescaled) boundary conditions and the stability analysis of the related discrete variational problems can be found in Refs. 26 and 27. The iterative scheme for flux–pressure–displacement formulation of the discrete MPET problem, analogous to Algorithm 1, is as follows:

Algorithm 2. Fully decoupled iterative scheme for flux–pressure–displacement formulation of discrete MPET problem.

Step a: Given $\boldsymbol{p}_h^k$ and $\boldsymbol{u}_h^k$, we first solve for $\boldsymbol{v}_h^{k+1}$, such that for all $\boldsymbol{z}_h\in\boldsymbol{V}_h$ there holds

$$(A_v\boldsymbol{v}_h^{k+1},\boldsymbol{z}_h) + (M\mathrm{Div}\,\boldsymbol{v}_h^{k+1},\mathrm{Div}\,\boldsymbol{z}_h)$$
$$= -(M\boldsymbol{g},\mathrm{Div}\,\boldsymbol{z}_h) + ((I - M(\Lambda_1+\Lambda_2))\boldsymbol{p}_h^k,\mathrm{Div}\,\boldsymbol{z}_h) - (M\underline{\mathrm{Div}}\,\boldsymbol{u}_h^k,\mathrm{Div}\,\boldsymbol{z}_h).$$

Step b: Given $\boldsymbol{u}_h^k$ and $\boldsymbol{v}_h^{k+1}$, we solve for $\boldsymbol{p}_h^{k+1}$, such that for all $\boldsymbol{q}_h\in\boldsymbol{P}_h$ there holds

$$(S\boldsymbol{p}_h^{k+1},\boldsymbol{q}_h)$$
$$= -(\boldsymbol{g},\boldsymbol{q}_h) + (S\boldsymbol{p}_h^k,\boldsymbol{q}_h) - ((\Lambda_1+\Lambda_2)\boldsymbol{p}_h^k,\boldsymbol{q}_h) - (\underline{\mathrm{Div}}\boldsymbol{u}_h^k,\boldsymbol{q}_h) - (\mathrm{Div}\,\boldsymbol{v}_h^{k+1},\boldsymbol{q}_h).$$

Step c: Given $\boldsymbol{p}_h^{k+1}$ and $\boldsymbol{v}_h^{k+1}$, we solve for $\boldsymbol{u}_h^{k+1}$, such that for all $\boldsymbol{w}_h\in\boldsymbol{U}_h$ there holds

$$a_h(\boldsymbol{u}_h^{k+1},\boldsymbol{w}_h) = (\boldsymbol{f},\boldsymbol{w}_h) + (\boldsymbol{p}_h^{k+1},\underline{\mathrm{Div}}\,\boldsymbol{w}_h).$$

In *Step a*, a coupled $H(\mathrm{div})$ problem is solved. As mentioned in Remark 6 of Ref. 27, we can apply orthogonal transformations to the flux and pressure subsystems which decouple the fluxes from each other and also the pressures from each other. For fluxes, this procedure results in n decoupled $H(\mathrm{div})$ problems for the operators $I+\bar{\mu}_i\nabla\mathrm{div}$, $i=1,2,\ldots,n$, where $\bar{\mu}_i$ are the eigenvalues of an $n\times n$ coefficient matrix, n denoting the number of networks, i.e. $n\in\{1,2,4,8\}$ in the

examples presented in Sec. 5; correspondingly, the decoupling of the pressure subsystem yields n, essentially, well-conditioned independent L^2 problems.

There are several works addressing the solution of nearly singular $H(\mathrm{div})$ problems and we may resort to Hiptmair–Xu preconditioners[25] and the robust subspace correction methods presented in Refs. 39, 40, 57 and 58. There also exist multigrid methods that serve this purpose.[5,55] In case of highly varying permeability (conductivity) coefficient, the auxiliary space multigrid preconditioners based on additive Schur complement approximation proposed in Ref. 34 provide a parameter-robust alternative.

In *Step c*, to obtain A_u^{-1} for the elasticity subproblem, one can use the multigrid method proposed in Ref. 29 for the discontinuous Galerkin discretization and the multigrid methods proposed in Refs. 38 and 47 for conforming elements, which are all robust with respect to the Lamé parameter λ. Following the methodology of the convergence analysis presented for the continuous MPET problem, statements analogous to those presented in Theorems 3.1 and 3.2 can also be proven for Algorithm 2.

5. Numerical Results

In the following, we consider four numerical test settings to demonstrate the effectiveness and accuracy of the proposed Uzawa-type iterative schemes for the MPET model.

First, numerical results are presented for the single network problem, i.e. the Biot model, in Fig. 1. These validate the theoretical convergence estimates of the linear stationary iterative method based on Algorithm 1 which has been additionally assessed against the preconditioned GMRES algorithm. In the second and third tests, the performance of Algorithm 1 is compared with the preconditioned GMRES algorithm and the fixed-stress algorithm as proposed in Ref. 28, cf. (2.10), for the two-network and four-network MPET problems. Finally, a scaling test demonstrating the behavior of the preconditioned GMRES and the Uzawa-type algorithms for different numbers of networks is performed.

The block Gauss–Seidel preconditioner that we have used to accelerate the GMRES method equals the lower block-triangular matrix in the left-hand side of (2.15) where $M = S^{-1}$ and S is given in (2.24).

All the numerical results in this section have been conducted on the FEniCS computing platform.[4,41] In all test cases the set-up is as follows:

- The domain $\Omega \in \mathbb{R}^2$ is the unit square which is partitioned into $2N^2$ congruent right-angled triangles.
- The discretization setting follows Sec. 4, see also Refs. 27 and 28, i.e. we use discontinuous piecewise constant elements, lowest-order Raviart–Thomas elements and Brezzi–Douglas–Marini elements to approximate the pressures, fluxes, and displacement fields, respectively.

- For all experiments conducted using Algorithm 1, whose implementation corresponds to Algorithm 2, we set

$$L_2 = \frac{\lambda_0}{(c_k^2+\lambda)(1+\frac{2\beta_d^2(\beta_s^{-2}+\lambda)R}{n})}, \quad L_1 = \frac{2(\beta_s^{-2}+\lambda)\beta_d^2}{\lambda_0}L_2$$

  and $\beta_s^2 = \beta_d^2 = 0.18$, see Ref. 21.
- The stopping criterium of the iterative process is the reduction of the initial preconditioned residual by a factor 10^8 where a random vector has been used in the initialization.

5.1. *The Biot's consolidation model*

Consider system (2.1) for $n = 1$, i.e. a system for which only one pressure and one flux exists, where for $(x, y) \in \Omega$

$$g = R_1\left(\frac{\partial^2\phi_2}{\partial x^2} + \frac{\partial^2\phi_2}{\partial y^2}\right) - \alpha_{p_1}(\phi_2 - 1),$$

$$\phi_1 = (x-1)^2(y-1)^2x^2y^2, \quad \phi_2 = 900(x-1)^2(y-1)^2x^2y^2,$$

and

$$\boldsymbol{f} = \frac{1}{2}\begin{pmatrix} -\dfrac{\partial^3\phi_1}{\partial x^2\partial y} - \dfrac{\partial^3\phi_1}{\partial y^3} + 2\dfrac{\partial\phi_2}{\partial x} \\ \dfrac{\partial^3\phi_1}{\partial x\partial y^2} + \dfrac{\partial^3\phi_1}{\partial x^3} + 2\dfrac{\partial\phi_2}{\partial y} \end{pmatrix}.$$

Experiments over a wide-range of input parameters α_p, λ, R_1^{-1} have been run with Algorithm 1 and the preconditioned GMRES algorithm and are shown in Fig. 1. In all test cases, the number of Uzawa-type iterations required to achieve the prescribed solution accuracy is bounded by a constant independent of all model and discretization parameters. Clearly, the GMRES preconditioned algorithm demonstrates better convergence behavior for small λ.

5.2. *The Biot–Barenblatt model*

In the next test, system (2.1) is considered for $n = 2$ where the problem setting is as per the cantilever bracket benchmark problem in Ref. 44. We denote the bottom, right, top, and left parts of $\Gamma = \partial\Omega$ by Γ_1, Γ_2, Γ_3, and Γ_4 and, also, we impose $\boldsymbol{u} = 0$ on Γ_4, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I})\boldsymbol{n} = (0,0)^T$ on $\Gamma_1 \cup \Gamma_2$, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I})\boldsymbol{n} = (0,-1)^T$ on Γ_3, $p_1 = 2$ on Γ and $p_2 = 20$ on Γ. Further, we set $\boldsymbol{f} = \boldsymbol{0}$, $g_1 = 0$, and $g_2 = 0$. Table 1 shows the reference values of the model parameters as given in Ref. 33.

Figures 2–4 present a comparison between the preconditioned GMRES algorithm, the fixed-stress split algorithm as presented in Ref. 28 with a tuning parameter $L = 1/(1+\lambda)$ and Algorithm 1. As can be seen, from Figs. 2 and 4 for λ being

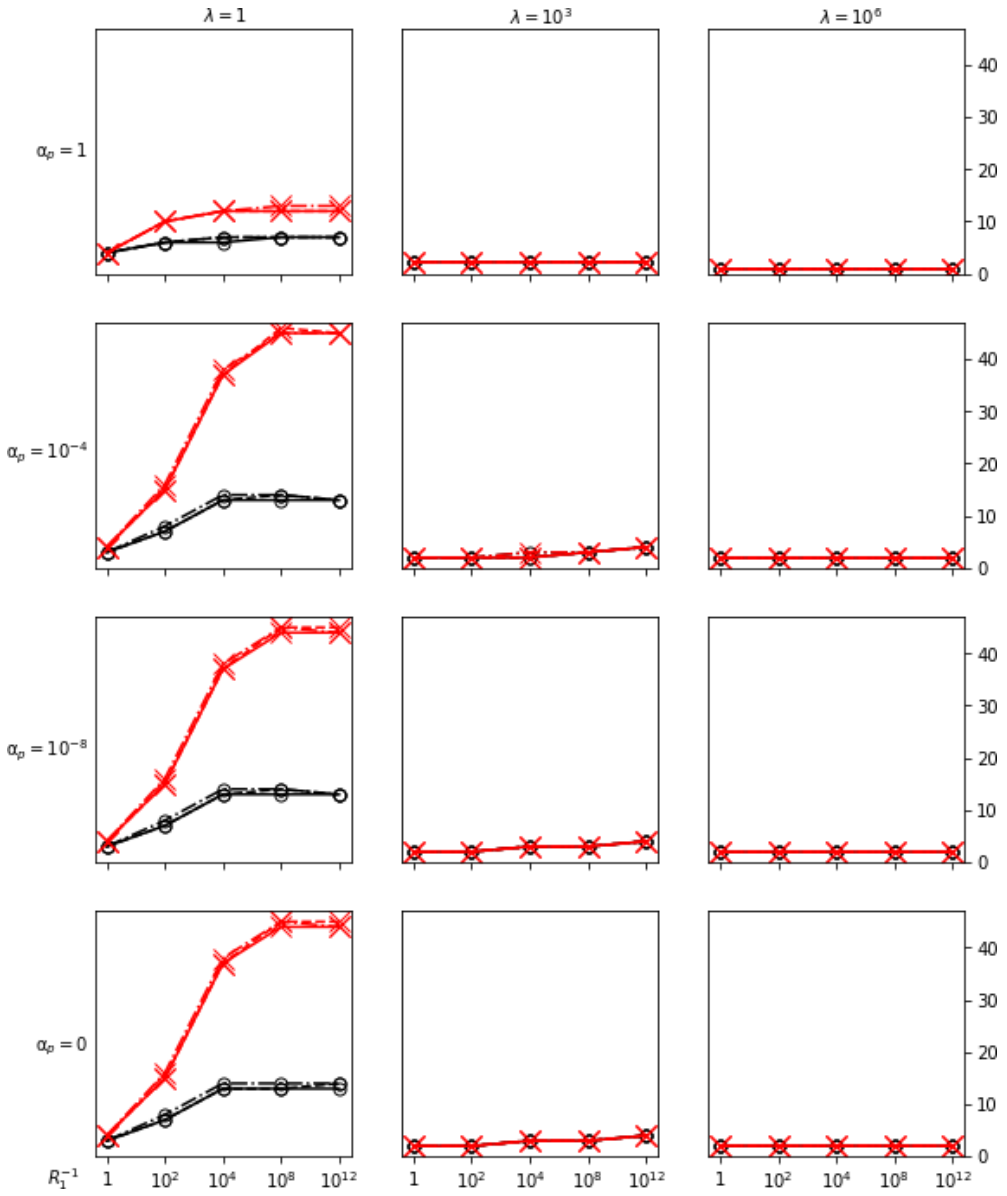


Fig. 1. (Color online) Number of preconditioned GMRES (small black circles) and augmented Uzawa-type (red crosses) iterations for preconditioned residual reduction by a factor 10^8 when solving the Biot problem. These tests have been performed for $h = 1/32$ (dash-dotted line), $h = 1/64$ (dashed line), and $h = 1/128$ (full line).

sufficiently large, the Uzawa-type method shows similar convergence behavior to the preconditioned GMRES and fixed-stress methods.

Furthermore, all the numerical results included in Figs. 2–4 demonstrate the robust performance of the Uzawa-type algorithm with respect to mesh refinements and variation of the hydraulic conductivities K_1 and K_2, as well as λ.

5.3. *The four-network model*

Now, we consider system (2.1) for $n = 4$. This test setting is analogous to the previous example, i.e. $\partial\Omega = \bar{\Gamma}_1 \cup \bar{\Gamma}_2 \cup \bar{\Gamma}_3 \cup \bar{\Gamma}_4$ with Γ_1, Γ_2, Γ_3, Γ_4 denoting the

Table 1. Reference values of model parameters for the Barenblatt model.

Parameter	Value	Unit
$\widehat{\lambda}$	$4.2 * 10^{6}$	$\mathrm{N\,m^{-2}}$
μ	$2.4 * 10^{6}$	$\mathrm{N\,m^{-2}}$
c_{p_1}	$5.4 * 10^{-8}$	$\mathrm{N^{-1}\,m^{2}}$
c_{p_2}	$1.4 * 10^{-8}$	$\mathrm{N^{-1}\,m^{2}}$
α_1	0.95	
α_2	0.12	
β	$5.0 * 10^{-10}$	$\mathrm{N^{-1}\,m^{2}\,s^{-1}}$
	$1.0 * 10^{-8}$	$\mathrm{N^{-1}\,m^{2}\,s^{-1}}$
K_1	$6.18 * 10^{-12}$	$\mathrm{N^{-1}\,m^{4}\,s^{-1}}$
K_2	$2.72 * 10^{-11}$	$\mathrm{N^{-1}\,m^{4}\,s^{-1}}$

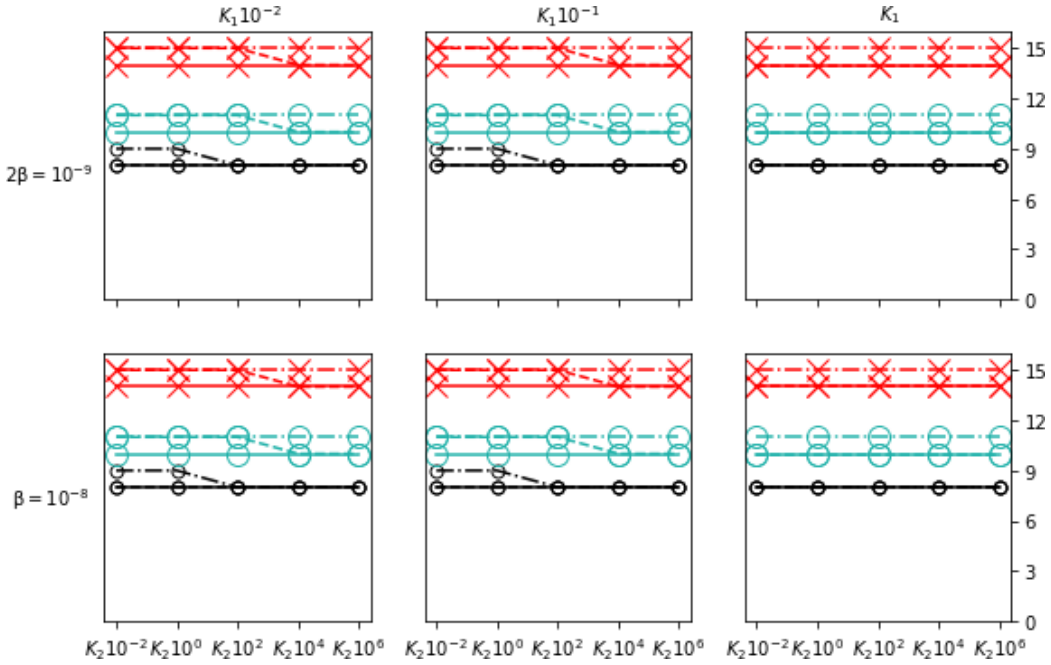


Fig. 2. (Color online) Number of preconditioned GMRES (small black circles), fixed-stress split (big green circles), and augmented Uzawa-type (red crosses) iterations for preconditioned residual reduction by a factor 10^8 when solving the Barenblatt problem, $\lambda = \widehat{\lambda}$. These tests have been performed for $h = 1/16$ (dash-dotted line), $h = 1/32$ (dashed line), and $h = 1/64$ (full line).

bottom, right, top, and left boundaries, respectively, $\boldsymbol{u} = 0$ on Γ_4, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I} - p_3\boldsymbol{I} - p_4\boldsymbol{I})\boldsymbol{n} = (0,0)^T$ on $\Gamma_1 \cup \Gamma_2$, $(\boldsymbol{\sigma} - p_1\boldsymbol{I} - p_2\boldsymbol{I} - p_3\boldsymbol{I} - p_4\boldsymbol{I})\boldsymbol{n} = (0,-1)^T$ on Γ_3, $p_1 = 2$ on Γ, $p_2 = 20$ on Γ, $p_3 = 30$ on Γ, $p_4 = 40$ on Γ. All right-hand sides have been chosen to be zero. The reference values of the parameters are taken from Ref. 52 and presented in Table 2.

The main aim of the numerical experiments discussed in this section is, again, the comparison between the three algorithms, namely the preconditioned GMRES algorithm, the fixed-stress split algorithm with $L = 1/(1+\lambda)$ and the fully decoupling Algorithm 1.

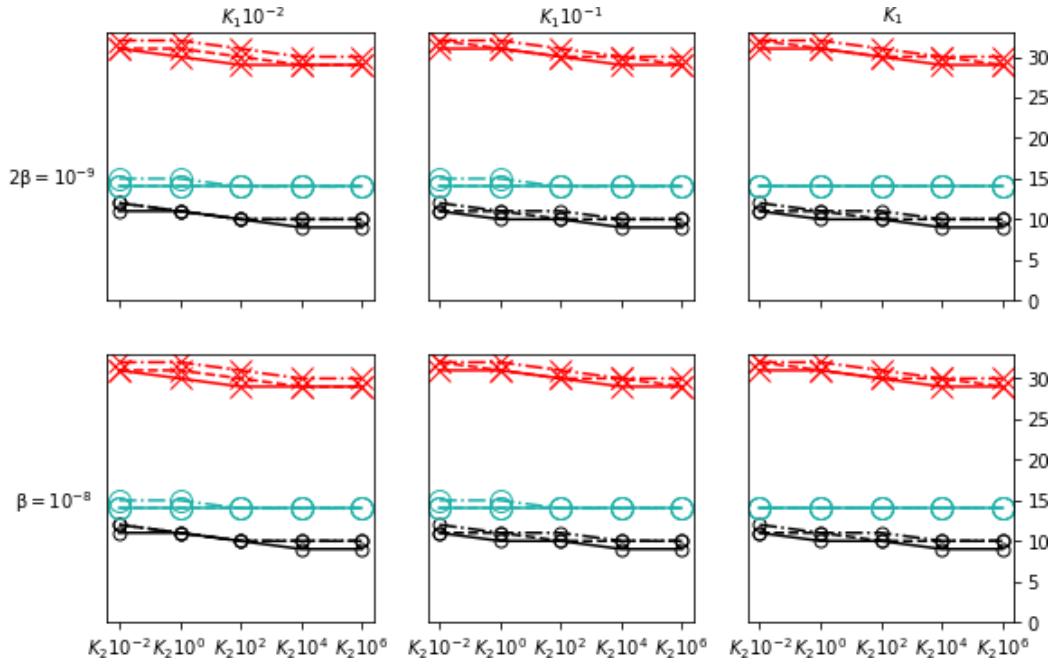


Fig. 3. (Color online) Number of preconditioned GMRES (small black circles), fixed-stress split (big green circles), and augmented Uzawa-type (red crosses) iterations for preconditioned residual reduction by a factor 10^8 when solving the Barenblatt problem, $\lambda := 0.01 * \widehat{\lambda}$. These tests have been performed for $h = 1/16$ (dash-dotted line), $h = 1/32$ (dashed line), and $h = 1/64$ (full line).

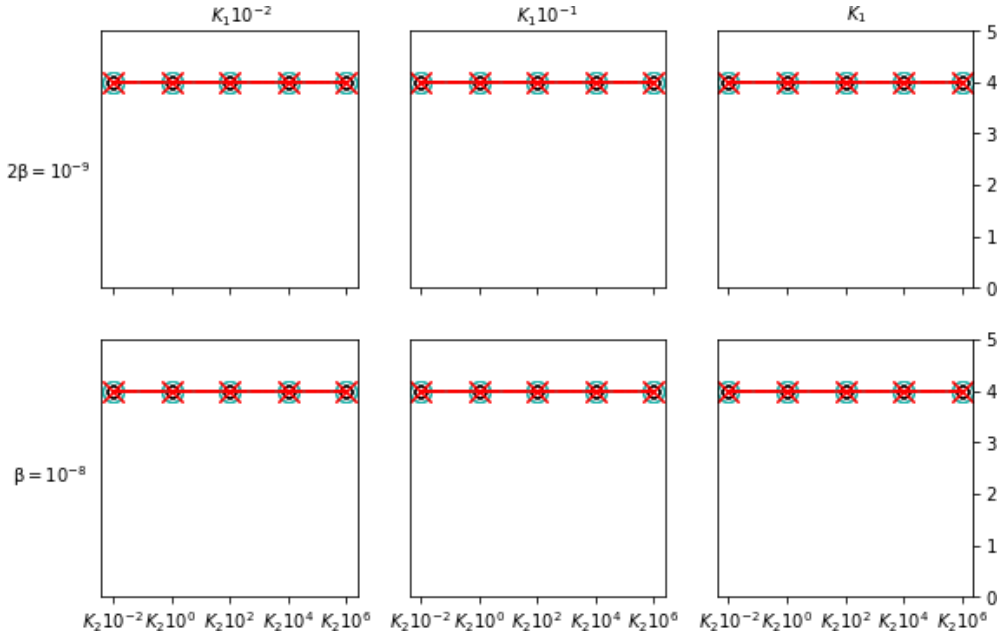


Fig. 4. (Color online) Number of preconditioned GMRES (small black circles), fixed-stress split (big green circles), and augmented Uzawa-type (red crosses) iterations for preconditioned residual reduction by a factor 10^8 when solving the Barenblatt problem, $\lambda := 100 * \widehat{\lambda}$. These tests have been performed for $h = 1/16$ (dash-dotted line), $h = 1/32$ (dashed line), and $h = 1/64$ (full line).

Figure 5 shows that Algorithm 1 exhibits a convergence behavior similar to that of the preconditioned GMRES method and the fixed-stress split iterative scheme over a wide-range of parameters as tabulated. Moreover, the presented numerical results demonstrate the robustness of the newly proposed algorithm with respect

Table 2. Reference values of model parameters for the four-network MPET model.

Parameter	Value	Unit
λ	505	$\mathrm{N\,m^{-2}}$
μ	216	$\mathrm{N\,m^{-2}}$
$c_{p_1} = c_{p_2} = c_{p_3} = c_{p_4}$	$4.5 * 10^{-10}$	$\mathrm{N^{-1}\,m^2}$
$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$	0.99	
$\beta_{12} = \beta_{24}$	$1.5 * 10^{-19}$	$\mathrm{N^{-1}\,m^2\,s^{-1}}$
β_{23}	$2.0 * 10^{-19}$	$\mathrm{N^{-1}\,m^2\,s^{-1}}$
β_{34}	$1.0 * 10^{-13}$	$\mathrm{N^{-1}\,m^2\,s^{-1}}$
$K_1 = K_2 = K_4 = K$	$3.75 * 10^{-6}$	$\mathrm{N^{-1}\,m^4\,s^{-1}}$
K_3	$1.57 * 10^{-9}$	$\mathrm{N^{-1}\,m^4\,s^{-1}}$

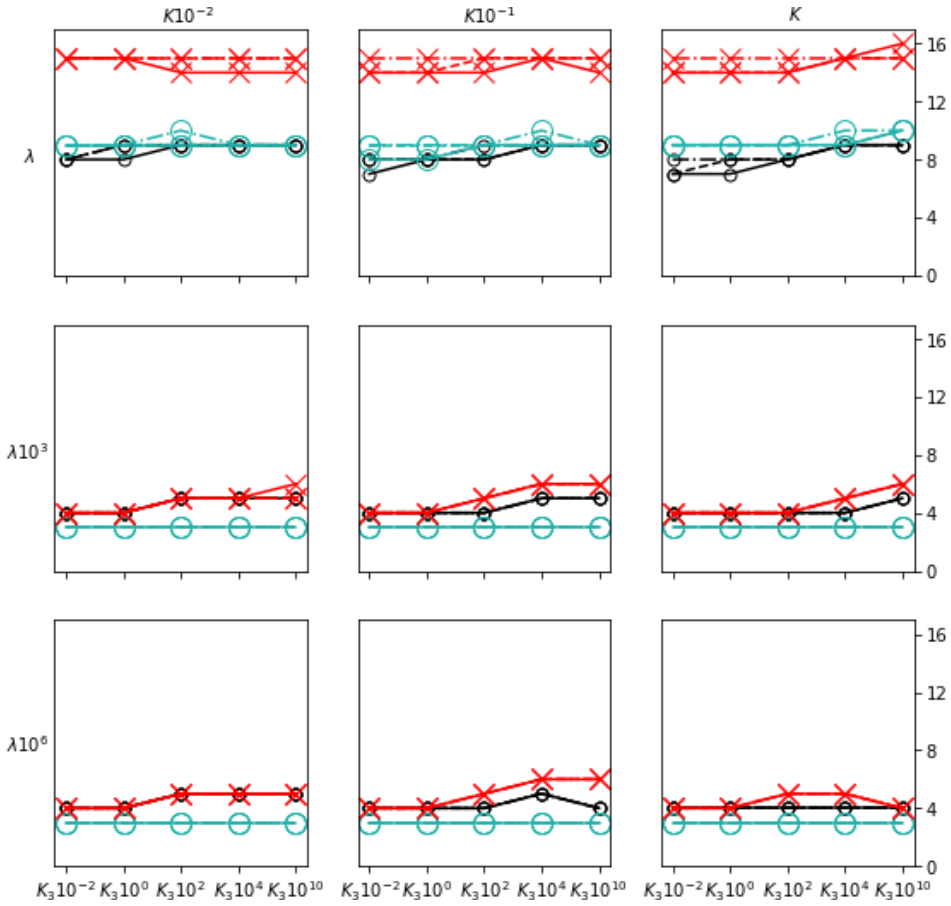


Fig. 5. (Color online) Number of preconditioned GMRES (small black circles), fixed-stress split (big green circles), and augmented Uzawa-type (red crosses) iterations for preconditioned residual reduction by a factor 10^8 when solving the four-network MPET problem. These tests have been performed for $h = 1/16$ (dash-dotted line), $h = 1/32$ (dashed line), and $h = 1/64$ (full line).

to large variations of the coefficients K_3, $K = K_1 = K_2 = K_4$ and λ and the mesh parameter h.

In order to further compare the performance of the preconditioned GMRES, fixed-stress split and augmented Uzawa-type algorithms we present one final table, Table 3, with elapsed times measured in seconds. These numerical tests have been

Table 3. Computational times in seconds for the preconditioned GMRES (t_G), fixed-stress split (t_F), and augmented Uzawa-type (t_U) algorithms to reach preconditioned residual reduction by a factor 10^8 in the norm induced by the preconditioner when solving the four-network MPET problem on a mesh with $h = 1/64$.

		$K_3 \cdot 10^{-2}$			K_3			$K_3 \cdot 10^{2}$			$K_3 \cdot 10^{4}$			$K_3 \cdot 10^{10}$		
		t_G	t_F	t_U	t_G	n_F	t_U	t_G	t_F	t_U	t_G	t_F	t_U	t_G	t_F	t_U
λ	$K \cdot 10^{-2}$	15.54	8.98	7.26	15.39	9.12	7.21	15.51	9.09	7.16	15.90	9.17	7.21	15.83	9.33	7.24
	K	15.32	9.20	7.68	15.60	9.13	7.66	15.40	8.91	7.13	15.75	9.12	7.41	16.09	9.26	7.68
	$K \cdot 10^{2}$	15.25	9.17	7.53	15.47	9.27	7.19	15.24	9.08	7.52	15.44	9.11	7.28	15.64	9.17	7.37
$\lambda \cdot 10^{3}$	$K \cdot 10^{-2}$	14.87	7.80	5.45	15.00	7.74	5.68	14.95	7.56	5.93	15.16	8.05	5.48	15.31	8.64	6.10
	K	14.71	7.78	5.38	14.81	7.91	5.42	14.74	8.10	5.75	15.05	8.03	6.68	15.23	8.07	6.05
	$K \cdot 10^{2}$	14.92	8.91	6.78	14.97	8.92	6.69	14.90	8.80	6.96	14.83	8.64	5.21	14.87	9.27	5.28
$\lambda \cdot 10^{6}$	$K \cdot 10^{-2}$	14.98	8.95	6.14	15.02	9.06	7.07	14.81	7.67	7.12	14.75	7.65	5.90	14.96	7.53	5.81
	K	14.91	8.89	5.40	15.08	8.96	5.61	15.12	9.03	7.01	15.06	9.12	6.33	15.19	9.32	6.20
	$K \cdot 10^{2}$	14.72	9.27	4.92	14.88	8.99	5.00	15.09	8.96	5.26	15.52	9.19	5.40	15.12	9.24	5.54

conducted on a Dell Precision 5540 notebook with an Intel Core i7-9 9850H processor and 64 GB RAM. As the results indicate, the Uzawa-type method is the computationally most efficient among the three, here, clearly seen in terms of total solution time when direct methods are used to solve the respective subproblems. A similar behavior can also be expected when iterative solvers of lower complexity replace the direct ones.

5.4. *Scaling test*

Finally, we present a scaling test demonstrating the convergence behavior of the preconditioned GMRES and augmented Uzawa-type algorithms with respect to the number of fluid networks n. These methods have been tested for $n = 1, 2, 4, 8$.

In order to perform a reasonable comparison, we have assumed that all network transfer coefficients are equal to 0 irrelevant to the number of networks, $\lambda = 10^3$, $R_i^{-1} = 10^4$, $\alpha_{p_i} = 10^{-4}$, $i = 1, \ldots, n$. The test setting is similar to those of the previously considered Biot–Barenblatt and four-network models, i.e. $\partial\Omega = \bar{\Gamma}_1 \cup \bar{\Gamma}_2 \cup \bar{\Gamma}_3 \cup \bar{\Gamma}_4$ with Γ_1, Γ_2, Γ_3, Γ_4 being the bottom, right, top, and left boundaries, respectively, $\boldsymbol{u} = 0$ on Γ_4, $(\boldsymbol{\sigma} - \sum_{i=1}^{n} p_i \boldsymbol{I})\boldsymbol{n} = (0,0)^T$ on $\Gamma_1 \cup \Gamma_2$, $(\boldsymbol{\sigma} - \sum_{i=1}^{n} p_i \boldsymbol{I})\boldsymbol{n} = (0,-1)^T$ on Γ_3, and $p_i = 10$, $i = 1, \ldots, n$ on Γ. As previously, all the right-hand sides have been chosen to be zero.

We have conducted the numerical tests on a mesh with a mesh-size $h = 1/32$. In all test cases, the number of required iterations to reach a preconditioned residual reduction by a factor 10^8 equals 4. This clearly indicates the robustness of the proposed algorithms with respect to the number of networks as suggested by our theoretical findings.

6. Concluding Remarks

The main contribution of this paper is the development of a new augmented Lagrangian Uzawa algorithm for three-by-three double saddle point block systems arising in Biot's and multiple-network poroelasticity models. The proposed method

fully decouples the fluid velocity, fluid pressure and solid displacement fields, contrary to the fixed-stress iterative scheme, which decouples only the flow from the mechanics problem. In this manner the subsystems that need to be solved in every iteration become considerably smaller, especially in the models where multiple fluid networks are present.

The presented convergence analysis proves the parameter-robust linear convergence of the new algorithm and additionally offers explicit formulas for a proper choice of required stabilization parameters. All numerical tests confirm the robustness and efficiency of the new fully decoupled iterative scheme and also its superiority in terms of computational work over existing methods.

References

1. J. Adler, F. Gaspar, X. Hu, C. Rodrigo and L. Zikatanov, Robust block preconditioners for Biot's model, in *Domain Decomposition Methods in Science and Engineering XXIV. DD 2017*, Lecture Notes in Computational Science and Engineering, Vol. 125 (Springer, 2019), pp. 3–16.
2. T. Almani, K. Kumar, A. Dogru, G. Singh and M. Wheeler, Convergence analysis of multirate fixed-stress split iterative schemes for coupling flow with geomechanics, *Comput. Methods Appl. Mech. Engrg.* **311** (2016) 180–207.
3. T. Almani, K. Kumar and M. Wheeler, Convergence and error analysis of fully discrete iterative coupling schemes for coupling flow with geomechanics, *Comput. Geosci.* **21** (2017) 1157–1172.
4. M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. Rognes and G. Wells, The FEniCS project version 1.5, *Arch. Numer. Softw.* **3** (2015) 9–23.
5. D. Arnold, R. Falk and R. Winther, Preconditioning in $H(\mathrm{div})$ and applications, *Math. Comp.* **66** (1997) 957–984.
6. T. Bærland, J. Lee, K.-A. Mardal and R. Winther, Weakly imposed symmetry and robust preconditioners for Biot's consolidation model, *Comput. Methods Appl. Math.* **17** (2017) 377–396.
7. M. Bai, D. Elsworth and J.-C. Roegiers, Multiporosity/multipermeability approach to the simulation of naturally fractured reservoirs, *Water Resour. Res.* **29** (1993) 1621–1633.
8. G. Barenblatt, G. Zheltov and I. Kochina, Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata], *J. Appl. Math. Mech.* **24** (1960) 1286–1303.
9. M. Bause, F. Radu and U. Köcher, Space-time finite element approximation of the Biot poroelasticity system with iterative coupling, *Comput. Methods Appl. Mech. Engrg.* **320** (2017) 745–768.
10. M. Benzi and F. Beik, Iterative methods for double saddle point systems, *SIAM J. Matrix Anal. Appl.* **39** (2018) 902–921.
11. M. Benzi and F. Beik, Uzawa-type and augmented Lagrangian methods for double saddle point systems, in *Structured Matrices in Numerical Linear Algebra*, eds. P. D. A. Bini, P. F. D. Benedetto, P. E. Tyrtyshnikov and P. M. V. Barel (Springer International Publishing, 2019), Chap. 11.
12. M. Biot, General theory of three-dimensional consolidation, *J. Appl. Phys.* **12** (1941) 155–164.

13. M. Biot, Theory of elasticity and consolidation for a porous anisotropic solid, *J. Appl. Phys.* **26** (1955) 182–185.
14. D. Boffi, M. Botti and D. Di Pietro, A nonconforming high-order method for the Biot problem on general meshes, *SIAM J. Sci. Comput.* **38** (2016) A1508–A1537.
15. D. Boffi, F. Brezzi and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics, Vol. 44 (Springer, 2013).
16. J. W. Both, M. Borregales, J. Nordbotten, K. Kumar and F. Radu, Robust fixed stress splitting for Biot's equations in heterogeneous media, *Appl. Math. Lett.* **68** (2017) 101–108.
17. J. Both, K. Kumar, J. Nordbotten and F. Radu, Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media, *Comput. Math. Appl.* **77** (2019) 1479–1502.
18. F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, *Rev. Fr. Automat. Inform. Rech. Opér. Sér. Rouge* **8** (1974) 129–151.
19. M. K. Brun, E. Ahmed, I. Berre, J. M. Nordbotten and F. A. Radu, Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport, *Comput. Math. Appl.* **80** (2020) 1964–1984.
20. D. Chou, J. Vardakis, L. Guo, B. Tully and Y. Ventikos, A fully dynamic multi-compartmental poroelastic system: Application to aqueductal stenosis, *J. Biomech.* **49** (2016) 2306–2312.
21. M. Costabel and M. Dauge, On the inequalities of Babuška-Aziz, Friedrichs and Horgan-Payne, *Arch. Ration. Mech. Anal.* **217** (2015) 873–898.
22. S. Dana and M. Wheeler, Convergence analysis of two-grid fixed stress iterative scheme for coupled flow and deformation in heterogeneous poroelastic media, *Comput. Methods Appl. Mech. Engrg.* **341** (2018) 788–806.
23. L. Guo, Z. Li, J. Lyu, Y. Mei, J. Vardakis, D. Chen, C. Han, X. Lou and Y. Ventikos, On the validation of a multiple-network poroelastic model using arterial spin labeling MRI data, *Front. Comput. Neurosci.* **13**.
24. L. Guo, J. Vardakis, T. Lassila, M. Mitolo, N. Ravikumar, D. Chou, M. Lange, A. Sarrami-Foroushani, B. Tully, Z. Taylor, S. Varma, A. Venneri, A. Frangi and Y. Ventikos, Subject-specific multi-poroelastic model for exploring the risk factors associated with the early stages of Alzheimer's disease, *Interface Focus* **8** (2018) 20170019.
25. R. Hiptmair and J. Xu, Nodal auxiliary space preconditioning in $H(\mathrm{curl})$ and $H(\mathrm{div})$ spaces, *SIAM J. Numer. Anal.* **45** (2007) 2483–2509 (electronic).
26. Q. Hong and J. Kraus, Parameter-robust stability of classical three-field formulation of Biot's consolidation model, *Electron. Trans. Numer. Anal.* **48** (2018) 202–226.
27. Q. Hong, J. Kraus, M. Lymbery and F. Philo, Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models, *Numer. Linear Algebra Appl.* **26** (2019) e2242, arXiv:1806.00353v2.
28. Q. Hong, J. Kraus, M. Lymbery and M. F. Wheeler, Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems, *Multiscale Model. Simul.* **18** (2020) 916–941.
29. Q. Hong, J. Kraus, J. Xu and L. Zikatanov, A robust multigrid method for discontinuous Galerkin discretizations of Stokes and linear elasticity equations, *Numer. Math.* **132** (2016) 23–49.
30. X. Hu, C. Rodrigo, F. Gaspar and L. Zikatanov, A nonconforming finite element method for the Biot's consolidation model in poroelasticity, *J. Comput. Appl. Math.* **310** (2017) 143–154.

31. G. Kanschat and B. Riviere, A finite element method with strong mass conservation for Biot's linear consolidation model, *J. Sci. Comput.* **77** (2018) 1762–1779.
32. J. Kim, H. Tchelepi and R. Juanes, Stability, accuracy and efficiency of sequential methods for coupled flow and geomechanics, *SPE J.* **16**.
33. A. Kolesov and P. Vabishchevich, Splitting schemes with respect to physical processes for double-porosity poroelasticity problems, *Russ. J. Numer. Anal. Math. Model.* **32** (2017) 99–113.
34. J. Kraus, R. Lazarov, M. Lymbery, S. Margenov and L. Zikatanov, Preconditioning heterogeneous $H(\mathrm{div})$ problems by additive Schur complement approximation and applications, *SIAM J. Sci. Comput.* **38** (2016) A875–A898.
35. J. Lee, Robust error analysis of coupled mixed methods for Biot's consolidation model, *J. Sci. Comput.* **69** (2016) 610–632.
36. J. Lee, K.-A. Mardal and R. Winther, Parameter-robust discretization and preconditioning of Biot's consolidation model, *SIAM J. Sci. Comput.* **39** (2017) A1–A24.
37. J. Lee, E. Piersanti, K.-A. Mardal and M. Rognes, A mixed finite element method for nearly incompressible multiple-network poroelasticity, *SIAM J. Sci. Comput.* **41** (2019) A722–A747.
38. Y. Lee, J. Wu and J. Chen, Robust multigrid method for the planar linear elasticity problems, *Numer. Math.* **113** (2009) 473–496.
39. Y. J. Lee, J. Wu, L. Xu and J. Zikatanov, Robust subspace correction methods for nearly singular systems, *Math. Models Methods Appl. Sci.* **17** (2007) 1937–1963.
40. Y. Lee, J. Wu, J. Xu and L. Zikatanov, A sharp convergence estimate for the method of subspace corrections for singular systems of equations, *Math. Comp.* **77** (2008) 831.
41. A. Logg, K.-A. Mardal, G. Wells *et al.*, *Automated Solution of Differential Equations by the Finite Element Method* (Springer, 2012).
42. K.-A. Mardal and R. Winther, Preconditioning discretizations of systems of partial differential equations, *Numer. Linear Algebra Appl.* **18** (2011) 1–40.
43. A. Mikelić and M. Wheeler, Convergence of iterative coupling for coupled flow and geomechanics, *Comput. Geosci.* **17**.
44. N. A. for Finite element methods & standards (Great Britain), *The Standard NAFEMS Benchmarks* (NAFEMS, 1990).
45. R. Oyarzúa and R. Ruiz-Baier, Locking-free finite element methods for poroelasticity, *SIAM J. Numer. Anal.* **54** (2016) 2951–2973.
46. C. Rodrigo, X. Hu, P. Ohm, J. Adler, F. Gaspar and L. Zikatanov, New stabilized discretizations for poroelasticity and the Stokes' equations, *Comput. Methods Appl. Mech. Eng.* **341** (2018) 467–484.
47. J. Schöberl, Multigrid methods for a parameter dependent problem in primal variables, *Numer. Math.* **84** (1999) 97–119.
48. R. Showalter, Poroelastic filtration coupled to Stokes flow, *Lecture Notes in Pure and Appl. Math.* **242** (2010) 229–241.
49. J. Sogn and W. Zulehner, Schur complement preconditioners for multiple saddle point problems of block tridiagonal form with application to optimization problems, *IMA J. Numer. Anal.* **39** (2019) 1328–1359.
50. E. Storvik, J. Both, K. Kumar, J. Nordbotten and F. Radu, On the optimization of the fixed-stress splitting for Biot's equations, *Int. J. Numer. Methods Engrg.* **120** (2019) 179–194.
51. B. Tully and Y. Ventikos, Cerebral water transport using multiple-network poroelastic theory: Application to normal pressure hydrocephalus, *J. Fluid Mech.* **667** (2011) 188–215.

52. J. Vardakis, D. Chou, B. Tully, C. Hung, T. Lee, P. Tsui and Y. Ventikos, Investigating cerebral oedema using poroelasticity, *Med. Eng. Phys.* **38** (2016) 48–57.
53. J. C. Vardakis, L. Guo, T. W. Peach, T. Lassila, M. Mitolo, D. Chou *et al.*, Fluid-structure interaction for highly complex, statistically defined, biological media: Homogenisation and a 3d multi-compartmental poroelastic model for brain biomechanics, *J. Fluids Struct.* **91**.
54. J. Vardakis, B. Tully and Y. Ventikos, Exploring the efficacy of endoscopic ventriculostomy for hydrocephalus treatment via a multicompartmental poroelastic model of CSF transport: A computational perspective, *PLoS ONE* **8** (2013) e84577.
55. P. S. Vassilevski and R. D. Lazarov, Preconditioning mixed finite element saddle-point elliptic problems, *Numer. Linear Algebra Appl.* **3** (1996) 1–20.
56. J. White, N. Castelletto and H. Tchelepi, Block-partitioned solvers for coupled poromechanics: A unified framework, *Comput. Methods Appl. Mech. Engrg.* **303** (2016) 55–74.
57. J. Xu, Iterative methods by space decomposition and subspace correction, *SIAM Rev.* **34** (1992) 581–613.
58. J. Xu and L. Zikatanov, The method of alternating projections and the method of subspace corrections in Hilbert space, *J. Amer. Math. Soc.* **15** (2002) 573–597.

UNIFORMLY WELL-POSED HYBRIDIZED DISCONTINUOUS GALERKIN HYBRID MIXED DISCRETIZATIONS FOR BIOT'S CONSOLIDATION MODEL

Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 384 (2021) 113991
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Uniformly well-posed hybridized discontinuous Galerkin/hybrid mixed discretizations for Biot's consolidation model

Johannes Kraus[a], Philip L. Lederer[b,*], Maria Lymbery[a], Joachim Schöberl[b]

[a] *Faculty of Mathematics, University Duisburg–Essen, 45127 Essen, Germany*
[b] *Institute for Analysis and Scientific Computing, TU Wien, 1040 Wien, Austria*

Received 15 December 2020; received in revised form 12 May 2021; accepted 1 June 2021
Available online xxxx

Abstract

We consider the quasi-static Biot's consolidation model in a three-field formulation with the three unknown physical quantities of interest being the displacement $\boldsymbol{u}$ of the solid matrix, the seepage velocity $\boldsymbol{v}$ of the fluid and the pore pressure p. As conservation of fluid mass is a leading physical principle in poromechanics, we preserve this property using an $\boldsymbol{H}(\mathrm{div})$-conforming ansatz for $\boldsymbol{u}$ and $\boldsymbol{v}$ together with an appropriate pressure space. This results in Stokes and Darcy stability and exact, that is, pointwise mass conservation of the discrete model.

The proposed discretization technique combines a hybridized discontinuous Galerkin method for the elasticity subproblem with a mixed method for the flow subproblem, also handled by hybridization. The latter allows for a static condensation step to eliminate the seepage velocity from the system while preserving mass conservation. The system to be solved finally only contains degrees of freedom related to $\boldsymbol{u}$ and p resulting from the hybridization process and thus provides, especially for higher-order approximations, a very cost-efficient family of physics-oriented space discretizations for poroelasticity problems.

We present the construction of the discrete model, theoretical results related to its uniform well-posedness along with optimal error estimates and parameter-robust preconditioners as a key tool for developing uniformly convergent iterative solvers. Finally, the cost-efficiency of the proposed approach is illustrated in a series of numerical tests for three-dimensional test cases.

Keywords: Biot's consolidation model; Strongly mass conserving high-order discretizations; Parameter-robust LBB stability; Norm-equivalent preconditioners; Hybrid discontinuous Galerkin methods; Hybrid mixed methods

1. Introduction

Poroelastic models describing the mechanical behavior of fluid saturated porous media find a wide range of applications in many different fields of science, medicine and engineering. The theory of poroelasticity was initially conceived by Maurice Anthony Biot who proposed a soil consolidation model to calculate the settlement of structures placed on fluid-saturated porous soils, see [1,2].

Recently, interest in Biot's consolidation equations has been revived due to their newly discovered applications in medicine, see e.g. [3] and [4], where they have been studied in the context of human cancellous bone samples

* Corresponding author.
E-mail addresses: johannes.kraus@uni-due.de (J. Kraus), philip.lederer@tuwien.ac.at (P.L. Lederer), maria.lymbery@uni-due.de (M. Lymbery), joachim.schoeberl@tuwien.ac.at (J. Schöberl).

https://doi.org/10.1016/j.cma.2021.113991
0045-7825/

and risk factors associated with the early stages of Alzheimer's disease, respectively. Their numerical solution has consequently been a subject of active research. One major challenge is that the parameters involved in Biot's model can vary over many orders of magnitude and, therefore, it is vital that not only the variational formulation of the problem is stable but also that the iterative solution method is uniformly convergent over the whole range of admissible model parameter values.

A rigorous stability and convergence analysis for finite element (FE) approximations of the two-field formulation of Biot's equations where the velocity field has been eliminated from the unknowns has first been presented in [5,6]. The derived a priori error estimates are valid for both semidiscrete and fully discrete formulations, where the backward Euler method is used for time-discretization and inf–sup stable finite elements are used for space discretization.

Other recent developments in discretizing Biot-type models are related to the stabilization of conforming methods [7], stable finite volume methods [8], discretizations for total-pressure-based formulations [9,10], including conservative discontinuous finite volume and mixed schemes [11], enriched Galerkin methods [12,13], space–time finite element approximations [14], and methods for two-phase flow and non-linear extensions of the Biot problem [13,15], to mention only but a few. Finally, and, nevertheless, important in the context of the present research, are the extensions of abovementioned discretization techniques to multicompartmental (multiple network) poroelasticity problems presented in [16,17].

The subject of the study in this paper is the standard three-field formulation of Biot's model in which the unknown fields are the displacement, seepage velocity and fluid pressure. Discretizations based on three-field-formulation have originally been proposed in [18,19] where continuous-in-time and discrete-in-time error estimates have been proved. This approach has also been extended to discontinuous Galerkin approximations of the displacement field in [20] and other nonconforming approximations, e.g., using modified rotated bilinear elements [21], or Crouzeix–Raviart elements for the displacements in [22]. More recently, in [23], a family of strongly mass conserving discretizations based on the $\boldsymbol{H}(\mathrm{div})$-conforming discontinuous Galerkin (DG) discretization of the displacement field has been suggested and its parameter-robust stability and near best approximation properties proven. Time-dependent error estimates for the same family of discretizations have been proved in [24]. Note that these approaches are based on the inf–sup stability of the corresponding Stokes discretization scheme which were originally stated in [25–27] and the Brinkman problem [28,29].

Hybridization techniques have been applied to discretizations of Biot's model in the recent works [30] and [31]. Whereas in [30] the authors introduced a hybridized $\boldsymbol{H}(\mathrm{div})$-conforming DG method for the two-field formulation, the work [31] starts from a lowest-order conforming stabilized discretization of the three-field formulation and uses hybridization for the flow subsystem as it was first presented in [32].

The aim of the present work is the construction, analysis and numerical testing of a new family of higher-order mass conserving hybridized/hybrid mixed FE discretizations for the three-field formulation of Biot's model. The main focus lies on a well-posedness analysis in properly scaled norms resulting in estimates with constants that are independent of any problem parameters. As a consequence, we obtain norm-equivalent preconditioners and optimal near best approximation estimates. Recently, norm-equivalent preconditioners for Biot's consolidation model have been also derived for a total-pressure based formulation in [9], a two-field formulation in [33], and for a classical flux-based three-field formulation in [23]. Non-linear extensions of Biot's consolidation model in which, e.g., the volumetric stress and fluid density are given by certain non-linear functions [34], large deformations are combined with deformation-dependent permeability [35] or with non-linear fluid compressibility [36], have been considered. The application of the family of higher-order mass conserving hybridized/hybrid mixed FE discretizations presented in this paper to such models is possible and work in progress, e.g., based on the approach in [37].

The most popular splitting methods (e.g., fixed stress or undrained split) can be described on a continuous level and the problems that have to be solved in each iteration of these iterative coupling schemes can then be discretized using the proposed techniques. For example, the combination of the presented discretizations with the methods described and analyzed in [17,38] is uncomplicated and straightforward to implement.

The paper is structured as follows. In Section 2 the governing equations are stated and the three-field formulation of Biot's model is discussed. Its semi-discretization in time by the implicit Euler method along with a proper rescaling of the parameters results in a static boundary value problem and is presented in Section 3. The latter then is discretized in space by a new family of hybridized discontinuous Galerkin/hybrid mixed methods while addressing the advantages of this approach. The main theoretical results follow in Section 4 where the uniform

boundedness and the parameter-robust inf–sup stability of the underlying bilinear form are proven to be independent of all model and discretization parameters. Furthermore, the corresponding parameter-robust preconditioners and error estimates are provided. In Section 5 the theoretical results of this paper are complemented by a series of numerical tests assessing the approximation quality and cost efficiency of these preconditioners for the proposed family of higher-order hybridized discontinuous Galerkin/hybrid mixed discretizations.

2. Problem formulation

2.1. Governing equations

We consider a porous medium, which is linearly elastic, homogeneous, isotropic and saturated by an incompressible Newtonian fluid. Then Biot's consolidation model, see [1,39], for a bounded Lipschitz domain $\Omega \in \mathbb{R}^d$, $d \in \{2, 3\}$,

$$-\mathrm{div}(2\tilde{\mu}\epsilon(\boldsymbol{u})) - \tilde{\lambda}\nabla \mathrm{div}\boldsymbol{u} + \alpha\nabla p = \tilde{\boldsymbol{f}}, \quad \text{in } \Omega \times (0, T), \tag{1a}$$

$$\frac{\partial}{\partial t}(S_0 p + \alpha \mathrm{div}\boldsymbol{u}) - \mathrm{div}(K\nabla p) = \tilde{g}, \quad \text{in } \Omega \times (0, T), \tag{1b}$$

relates the deformation $\boldsymbol{u}$ and the fluid pressure p for a given body force density $\tilde{\boldsymbol{f}}$ and mass source or sink $\tilde{g}$. For convenience, we assume a scalar conductivity coefficient K. We use bold symbols to denote vector- or tensor-valued quantities, e.g., $\boldsymbol{\epsilon}(\boldsymbol{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^T)$ denoting the symmetric gradient. Further, $\tilde{\lambda}$ and $\tilde{\mu}$ are the Lamé parameters, α is the Biot–Willis parameter and S_0 the constrained specific storage coefficient.

The three-field [18,20] formulation is based on the primary variables $(\boldsymbol{u}, \boldsymbol{w}, p)$ and reads

$$-\mathrm{div}\boldsymbol{\sigma} = \tilde{\boldsymbol{f}}, \quad \text{in } \Omega \times (0, T), \tag{2a}$$

$$K^{-1}\boldsymbol{w} + \nabla p = \boldsymbol{0}, \quad \text{in } \Omega \times (0, T), \tag{2b}$$

$$\frac{\partial}{\partial t}(S_0 p + \alpha\, \mathrm{div}\, \boldsymbol{u}) + \mathrm{div}\, \boldsymbol{w} = \tilde{g}, \quad \text{in } \Omega \times (0, T), \tag{2c}$$

where $\boldsymbol{w}$ denotes the seepage velocity, $\tilde{\boldsymbol{\sigma}} := 2\tilde{\mu}\boldsymbol{\epsilon}(\boldsymbol{u}) + \tilde{\lambda}\mathrm{div}(\boldsymbol{u})\boldsymbol{I}$ is the total stress and $\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}} - \alpha p\boldsymbol{I}$ the effective stress. If not mentioned otherwise, we assume homogeneous Dirichlet boundary conditions for the displacement $\boldsymbol{u}$ and homogeneous Neumann conditions for the pressure p. In this context, let $\boldsymbol{H}_0^1(\Omega)$, $\boldsymbol{H}_0(\mathrm{div}, \Omega)$ denote the standard vector-valued Sobolev spaces where the subscript 0 refers to homogeneous essential boundary conditions. Further, let $L_0^2(\Omega)$ denote the space of square integrable functions with zero mean value. Following the standard procedure, one derives the weak formulation: Find $(\boldsymbol{u}, \boldsymbol{w}, p) \in \boldsymbol{H}_0^1(\Omega) \times \boldsymbol{H}_0(\mathrm{div}, \Omega) \times L_0^2(\Omega)$ such that

$$\tilde{a}(\boldsymbol{u}, \boldsymbol{v}) - (\alpha\, p, \mathrm{div}\, \boldsymbol{v}) = (\tilde{\boldsymbol{f}}, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega), \tag{3a}$$

$$(K^{-1}\boldsymbol{w}, \boldsymbol{z}) - (p, \mathrm{div}\, \boldsymbol{z}) = 0, \qquad \forall \boldsymbol{z} \in \boldsymbol{H}_0(\mathrm{div}, \Omega), \tag{3b}$$

$$-(\alpha\, \mathrm{div}\, \partial_t \boldsymbol{u}, q) - (\mathrm{div}\, \boldsymbol{w}, q) - (S_0 \partial_t p, q) = -(\tilde{g}, q), \qquad \forall q \in L_0^2(\Omega), \tag{3c}$$

where

$$\tilde{a}(\boldsymbol{u}, \boldsymbol{v}) := 2\tilde{\mu} \int_\Omega \varepsilon(\boldsymbol{u}) : \varepsilon(\boldsymbol{v}) \mathrm{d}x + \tilde{\lambda} \int_\Omega \mathrm{div}\, \boldsymbol{u}\, \mathrm{div}\, \boldsymbol{v} \mathrm{d}x. \tag{4}$$

Finally, system (3) is completed with suitable initial conditions $\boldsymbol{u}(\cdot, 0) = \boldsymbol{u}_0(\cdot)$ and $p(\cdot, 0) = p_0(\cdot)$.

3. Hybridized discontinuous Galerkin/hybrid mixed discretizations of the Biot problem

3.1. Strongly mass conserving discretization of the Biot problem

The starting point for this subsection is a family of strongly mass conserving discretizations of the three-field formulation of the quasi-static Biot model based on a discontinuous Galerkin (DG) formulation of the mechanics subproblem, as proposed in [23]. After time discretization by the implicit Euler scheme, the method for the arising static problem in each time step can be expressed as follows:

Find the time-step functions $(\boldsymbol{u}^k, \boldsymbol{w}^k, p^k) := (\boldsymbol{u}(\boldsymbol{x}, t_k), \boldsymbol{w}(\boldsymbol{x}, t_k), p(\boldsymbol{x}, t_k)) \in \boldsymbol{H}^1_0(\Omega) \times \boldsymbol{H}_0(\mathrm{div}, \Omega) \times L^2_0(\Omega)$ which solve the following system of equations

$$\tilde{a}(\boldsymbol{u}^k, \boldsymbol{v}) - (\alpha\, p^k, \mathrm{div}\, \boldsymbol{v}) = (\tilde{\boldsymbol{f}}^k, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}^1_0(\Omega), \tag{5a}$$

$$(K^{-1}\boldsymbol{w}^k, \boldsymbol{z}) - (p^k, \mathrm{div}\, \boldsymbol{z}) = 0, \qquad \forall \boldsymbol{z} \in \boldsymbol{H}_0(\mathrm{div}, \Omega), \tag{5b}$$

$$-(\alpha\, \mathrm{div}(\boldsymbol{u}^k - \boldsymbol{u}^{k-1}), q) - \tau(\mathrm{div}\, \boldsymbol{w}^k, q) - (S_0(p^k - p^{k-1}), q) = -\tau(\tilde{g}^k, q), \quad \forall q \in L^2_0(\Omega), \tag{5c}$$

where τ is the time-step parameter and $\tilde{\boldsymbol{f}}^k = \tilde{\boldsymbol{f}}(\boldsymbol{x}, t_k)$, $\tilde{g}^k = \tilde{g}(\boldsymbol{x}, t_k)$.

For the space discretization, consider a shape-regular triangulation $\mathcal{T}_h$ whose set of facets are denoted by $\mathcal{F}_h$. We introduce the following finite element spaces

$$\begin{aligned}
\boldsymbol{U}_h &:= \{\boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{div}, \Omega) : \boldsymbol{v}|_T \in \boldsymbol{U}(T),\ T \in \mathcal{T}_h\},\\
\boldsymbol{W}_h &:= \{\boldsymbol{z} \in \boldsymbol{H}_0(\mathrm{div}, \Omega) : \boldsymbol{z}|_T \in \boldsymbol{W}(T),\ T \in \mathcal{T}_h\},\\
P_h &:= \{q \in L^2_0(\Omega) : q|_T \in P(T),\ T \in \mathcal{T}_h\}.
\end{aligned}$$

The local spaces $\boldsymbol{U}(T)$, $\boldsymbol{W}(T)$, $P(T)$ are either $\mathrm{BDM}_\ell(T)$, $\mathrm{RT}_{\ell-1}(T)$, $\mathrm{P}_{\ell-1}(T)$ or by $\mathrm{BDFM}_\ell(T)$, $\mathrm{RT}_{\ell-1}(T)$, $\mathrm{P}_{\ell-1}(T)$ where $\mathrm{BD(F)M}_\ell(T)$, $\mathrm{RT}_{\ell-1}(T)$, and $\mathrm{P}_{\ell-1}(T)$ denote the local Brezzi–Douglas–(Fortin–)Marini space of order ℓ, the Raviart–Thomas space of order $\ell - 1$, and full polynomials of degree $\ell - 1$, respectively. A definition of these local spaces can be found, for example, in [40]. Since the exact solution belongs to $\boldsymbol{H}^1_0(\Omega)$, the discrete space $\boldsymbol{U}_h$ leads to a nonconforming method where tangential continuity is incorporated via a DG-formulation of the mechanics subproblem. The benefit of the proposed combination of FE spaces is that they provide exact fluid mass conservation, i.e., (2c) is fulfilled pointwise.

We present the definitions of some trace operators next. Let $F = \partial T_1 \cap \partial T_2$ be a common facet of two adjacent elements $T_1, T_2 \in \mathcal{T}_h$, and let $\boldsymbol{n}_1, \boldsymbol{n}_2$ be the corresponding outward pointing unit normal vectors. For any interior facet $F \not\subset \partial\Omega$ and element-wise smooth and scalar-valued function q, vector-valued function $\boldsymbol{v}$ and tensor-valued function $\boldsymbol{\tau}$, their averages and jumps on the facet F are defined by

$$\{\boldsymbol{v}\} = \frac{1}{2}(\boldsymbol{v}_1 \cdot \boldsymbol{n}_1 - \boldsymbol{v}_2 \cdot \boldsymbol{n}_2), \quad \{\boldsymbol{\tau}\} = \frac{1}{2}(\boldsymbol{\tau}_1\boldsymbol{n}_1 - \boldsymbol{\tau}_2\boldsymbol{n}_2), \quad [q] = q_1 - q_2, \quad [\boldsymbol{v}] = \boldsymbol{v}_1 - \boldsymbol{v}_2,$$

where the subscript i, $i = 1, 2$, with the functions q, $\boldsymbol{v}$ and $\boldsymbol{\tau}$ refers to their evaluation on $T_i \cap F$. For any boundary facet $F \subset \partial\Omega$, these quantities are given as

$$\{\boldsymbol{v}\} = \boldsymbol{v}|_F \cdot \boldsymbol{n}, \quad \{\boldsymbol{\tau}\} = \boldsymbol{\tau}|_F\boldsymbol{n}, \quad [q] = q|_F, \quad [\boldsymbol{v}] = \boldsymbol{v}|_F.$$

With these definitions at hand, the formulation of the method is as follows: Find $(\boldsymbol{u}_h, \boldsymbol{w}_h, p_h) \in \boldsymbol{U}_h \times \boldsymbol{W}_h \times P_h$, such that

$$a_h(\boldsymbol{u}_h, \boldsymbol{v}_h) - (p_h, \mathrm{div}\boldsymbol{v}_h) = (\boldsymbol{f}, \boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \boldsymbol{U}_h, \tag{6a}$$

$$(R^{-1}\boldsymbol{w}_h, \boldsymbol{z}_h) - (p_h, \mathrm{div}\boldsymbol{z}_h) = 0, \quad \forall \boldsymbol{z}_h \in \boldsymbol{W}_h, \tag{6b}$$

$$-(\mathrm{div}\boldsymbol{u}_h, q_h) - (\mathrm{div}\boldsymbol{w}_h, q_h) - (Sp_h, q_h) = (g, q_h), \quad \forall q_h \in P_h. \tag{6c}$$

This system has been derived by dividing system (5) by $2\tilde{\mu}$ and, additionally, Eq. (5b) by the time step size τ and furthermore by applying the substitutions $\boldsymbol{u}_h = \alpha\boldsymbol{u}^k_h$, $\boldsymbol{w}_h = \tau\boldsymbol{w}^k_h$, $p_h = \alpha^2 p^k_h/2\tilde{\mu}$. The right-hand sides in (6) are $\boldsymbol{f} = \alpha\tilde{\boldsymbol{f}}(\boldsymbol{x}, t_k)/2\tilde{\mu}$ and $g = (\tau\tilde{g}(\boldsymbol{x}, t_k) - \alpha\mathrm{div}(\boldsymbol{u}_h(\boldsymbol{x}, t_{k-1})) - S_0 p(\boldsymbol{x}, t_{k-1}))/2\tilde{\mu}$,

$$a_h(\boldsymbol{u}_h, \boldsymbol{v}_h) := a^{\mathrm{DG}}_h(\boldsymbol{u}_h, \boldsymbol{v}_h) + \lambda \int_\Omega \mathrm{div}\boldsymbol{u}_h\, \mathrm{div}\boldsymbol{v}_h \mathrm{d}\, x$$

and

$$\lambda := \frac{\tilde{\lambda}}{2\tilde{\mu}}, \quad R := \frac{2\tilde{\mu}\tau}{\alpha^2}K > 0, \quad S := \frac{2\tilde{\mu}S_0}{\alpha^2}. \tag{7}$$

Note that the discrete bilinear form $a_h(\cdot, \cdot)$ is obtained from scaling the bilinear form in (4) by $1/2\tilde{\mu}$. We denote the tangential component of any vector field on a facet by its symbol with a subscript t. Then the symmetric interior

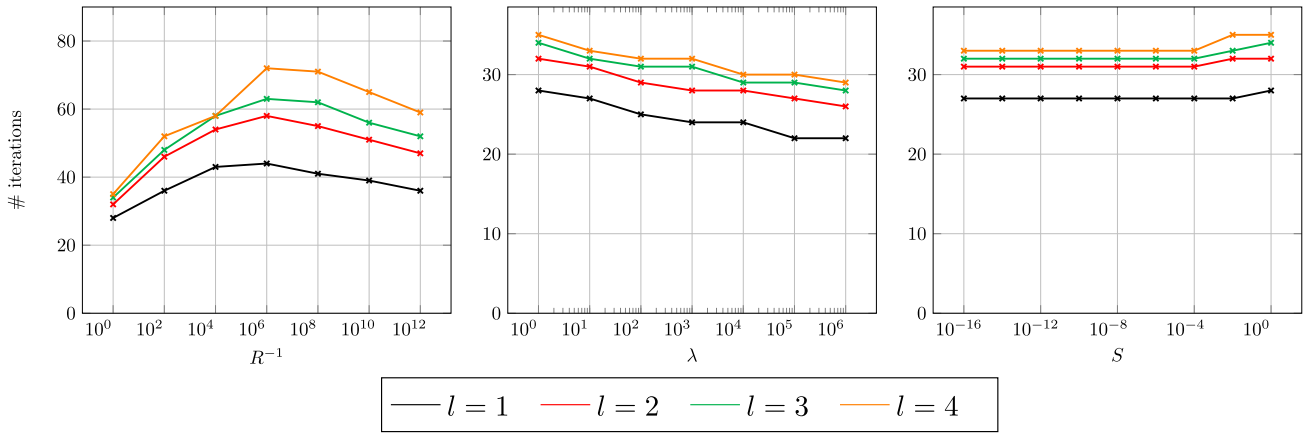


Fig. 1. Robustness of the preconditioner defined in (63).

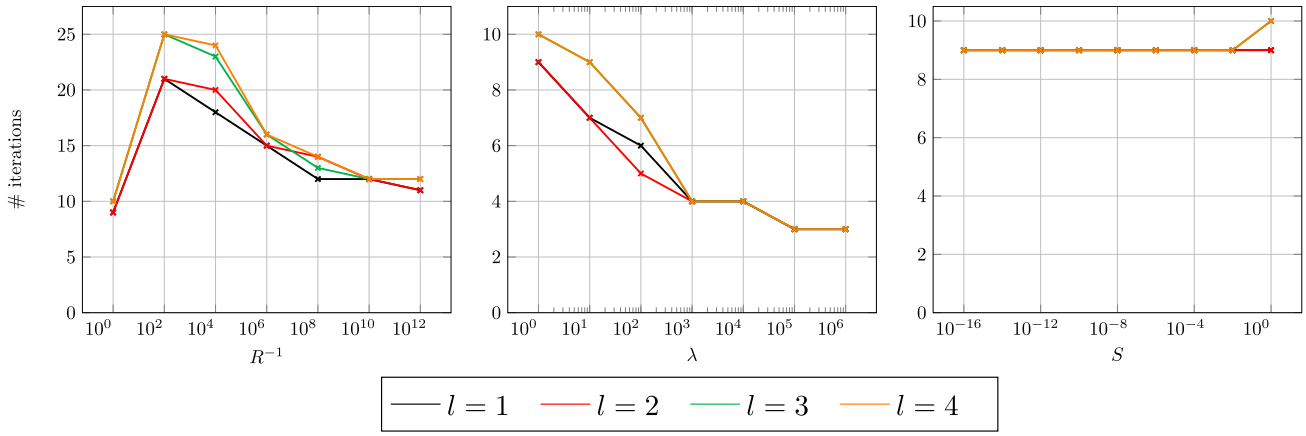


Fig. 2. Robustness of the preconditioner defined in (64).

penalty Galerkin (SIPG) bilinear form $a_h^{\mathrm{DG}}(\cdot,\cdot)$ is defined as

$$a_h^{\mathrm{DG}}(\boldsymbol{u},\boldsymbol{v}) := \sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{\epsilon}(\boldsymbol{u}) : \boldsymbol{\epsilon}(\boldsymbol{v})\mathrm{d}\,x - \sum_{F\in\mathcal{F}_h}\int_F \{\boldsymbol{\epsilon}(\boldsymbol{u})\}\cdot[\boldsymbol{v}_t]\mathrm{d}\,s$$
$$- \sum_{F\in\mathcal{F}_h}\int_F \{\boldsymbol{\epsilon}(\boldsymbol{v})\}\cdot[\boldsymbol{u}_t]\mathrm{d}\,s + \sum_{F\in\mathcal{F}_h}\eta\ell^2 h_F^{-1}\int_F [\boldsymbol{u}_t]\cdot[\boldsymbol{v}_t]\mathrm{d}\,s \tag{8}$$

with a sufficiently large stabilization parameter η independent of all model parameters, i.e., λ, R, S, and discretization parameters h (mesh size), and τ (time step), and l (polynomial order). Note that in this paper the constants in all parameter-robust estimates are independent of all model parameters and the discretization parameters h and τ. While the numerical results, see Fig. 1 and 2, show only a mild dependence on the polynomial order, we want to emphasize that we do not claim uniform robustness in l.

3.2. Hybridized DG method

When dealing with Stokes-type problems, $\boldsymbol{H}(\mathrm{div})$-conforming discretizations possess several advantages over H^1-conforming discretizations. This is mainly due to the fact that they allow for a suitable approximation of the incompressibility constraint which results in favorable properties such as pointwise divergence-free solutions and pressure robustness, see, e.g., [25,26]. However, the incorporation of (tangential) continuity in standard DG schemes leads to a significantly increased number of (globally) coupled degrees of freedom (dof). To overcome this, in *hybridized DG methods*, one decouples element unknowns by introducing additional unknowns on the facets through which (tangential) continuity is imposed weakly, see, e.g., [41,42].

In the context of an $\boldsymbol{H}(\mathrm{div})$-conforming hybridized DG discretization, one introduces an additional space

$$\widehat{\boldsymbol{U}}_h := \{\hat{\boldsymbol{u}} \in \boldsymbol{L}^2(\mathcal{F}_h) : \hat{\boldsymbol{u}}|_F \in \boldsymbol{P}_\ell(F) \text{ and } \hat{\boldsymbol{u}}|_F \cdot \boldsymbol{n} = 0,\ \ F \in \mathcal{F}_h;\ \ \hat{\boldsymbol{u}} = \boldsymbol{0} \text{ on } \partial\Omega\}$$

for the approximation of the tangential trace of the displacement field $\boldsymbol{u}$. Here, $\boldsymbol{L}^2(\mathcal{F}_h)$ denotes the space of vector-valued square integrable functions on the skeleton $\mathcal{F}_h$ and $\boldsymbol{P}_\ell(F)$ the vector-valued polynomial space of order ℓ on each facet $F \in \mathcal{F}_h$. We replace the bilinear form $a_h^{\mathrm{DG}}(\cdot,\cdot)$ defined in (8) by $a_h^{\mathrm{HDG}}(\cdot,\cdot)$ given by

$$\begin{aligned} a_h^{\mathrm{HDG}}((\boldsymbol{u},\hat{\boldsymbol{u}}),(\boldsymbol{v},\hat{\boldsymbol{v}})) := \sum_{T\in\mathcal{T}_h} \Bigg[&\int_T \boldsymbol{\epsilon}(\boldsymbol{u}) : \boldsymbol{\epsilon}(\boldsymbol{v})\,\mathrm{d}x + \int_{\partial T} \boldsymbol{\epsilon}(\boldsymbol{u})\boldsymbol{n}\cdot(\hat{\boldsymbol{v}}-\boldsymbol{v})_t\,\mathrm{d}s \\ &+ \int_{\partial T} \boldsymbol{\epsilon}(\boldsymbol{v})\boldsymbol{n}\cdot(\hat{\boldsymbol{u}}-\boldsymbol{u})_t\,\mathrm{d}s + \eta\ell^2 h^{-1}\int_{\partial T}(\hat{\boldsymbol{u}}-\boldsymbol{u})_t\cdot(\hat{\boldsymbol{v}}-\boldsymbol{v})_t\mathrm{d}s\Bigg], \end{aligned} \tag{9}$$

where $(\boldsymbol{u},\hat{\boldsymbol{u}}),(\boldsymbol{v},\hat{\boldsymbol{v}}) \in \overline{\boldsymbol{U}}_h := \boldsymbol{U}_h \times \widehat{\boldsymbol{U}}_h$. Our approach for exactly divergence-free hybridized discontinuous Galerkin methods will be based on [42] as well as its improvements presented in [43,44]. The resulting method for the Biot problem now reads as: Find $((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),\boldsymbol{w}_h,p_h) \in \overline{\boldsymbol{U}}_h \times \boldsymbol{W}_h \times P_h$, such that

$$a_h((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)) - (p_h,\mathrm{div}\boldsymbol{v}_h) = (\boldsymbol{f},\boldsymbol{v}_h), \quad \forall(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)\in\overline{\boldsymbol{U}}_h, \tag{10a}$$

$$(R^{-1}\boldsymbol{w}_h,\boldsymbol{z}_h) - (p_h,\mathrm{div}\boldsymbol{z}_h) = 0, \quad \forall\boldsymbol{z}_h\in\boldsymbol{W}_h, \tag{10b}$$

$$-(\mathrm{div}\boldsymbol{u}_h,q_h) - (\mathrm{div}\boldsymbol{w}_h,q_h) - (Sp_h,q_h) = (g,q_h), \quad \forall q_h\in P_h, \tag{10c}$$

where

$$a_h((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)) := a_h^{\mathrm{HDG}}((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)) + \lambda\int_\Omega \mathrm{div}\boldsymbol{u}_h\mathrm{div}\boldsymbol{v}_h\,\mathrm{d}x \tag{11}$$

and $a_h^{\mathrm{HDG}}((\cdot,\cdot),(\cdot,\cdot))$ is defined in (9).

3.3. A family of hybridized DG/hybrid mixed methods

In this subsection, we enrich the hybridization idea by additionally introducing a hybrid mixed formulation for the flow subproblem. While the stability analysis presented in [23,45] uses properly scaled $\boldsymbol{H}(\mathrm{div})$ and L^2 norms for the flow subproblem, we hybridize the latter one in the present work. This approach has the advantage that when solving the full saddle-point problem with some preconditioned iterative method, one needs to invert a div–grad type operator instead of a grad–div operator in order to apply the preconditioner which is easier and more cost-efficient in general. Note that the solution of the hybridized system is the same as that of the non-hybridized one.

The additional hybridization step can be expressed as follows. First, one enforces the normal continuity of the velocity by a Lagrange multiplier. To this end, we introduce the following finite element spaces

$$\boldsymbol{W}_h^- := \{\boldsymbol{z}\in\boldsymbol{L}^2(\Omega) : \boldsymbol{z}|_T \in \boldsymbol{W}(T),\ \ T\in\mathcal{T}_h\}, \quad \widehat{P}_h := \prod_{F\in\mathcal{F}_h} \mathrm{P}_{l-1}(F), \quad \overline{P}_h := P_h \times \widehat{P}_h,$$

where $\boldsymbol{W}(T)$ can be chosen in the same way as before. Here, the space $\boldsymbol{W}_h^-$ is simply a discontinuous version of the space $\boldsymbol{W}_h$. Further, note that $\widehat{P}_h$ is chosen as the normal trace space of $\boldsymbol{W}_h$, e.g., in the case $\boldsymbol{W}(T) = RT_0$ (thus $l-1=0$) the normal traces on each facet are constant and so correspondingly we also choose $\widehat{P}_h$ to be defined as facet wise constants. Based on these spaces, we next define for all $\boldsymbol{w}_h \in \boldsymbol{W}_h^-$ and $(p_h,\hat{p}_h)\in\overline{P}_h$ the bilinear form

$$b((p_h,\hat{p}_h),\boldsymbol{w}_h) = \sum_{T\in\mathcal{T}_h}\left(\int_T \mathrm{div}\,\boldsymbol{w}_h\,p_h\mathrm{d}x - \int_{\partial T}\boldsymbol{w}_h\cdot\boldsymbol{n}\,\hat{p}_h\mathrm{d}s\right). \tag{12}$$

This bilinear form can be interpreted as a distributional version of the inner product "$(\mathrm{div}\,\boldsymbol{w}_h, p_h)$" since functions in $\boldsymbol{W}_h^-$ are not normal continuous. Therefore, variational problem (10), when using a hybrid mixed formulation of the flow subproblem, is expressed as: Find $((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),\boldsymbol{w}_h,(p_h,\hat{p}_h)) \in \overline{\boldsymbol{U}}_h \times \boldsymbol{W}_h^- \times \overline{P}_h$, such that

$$a_h((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)) - (p_h,\mathrm{div}\boldsymbol{v}_h) = (\boldsymbol{f},\boldsymbol{v}_h), \quad \forall(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)\in\overline{\boldsymbol{U}}_h, \tag{13a}$$

$$(R^{-1}\boldsymbol{w}_h,\boldsymbol{z}_h) - b((p_h,\hat{p}_h),\boldsymbol{z}_h) = 0, \quad \forall\boldsymbol{z}_h\in\boldsymbol{W}_h^-, \tag{13b}$$

$$-(\mathrm{div}\boldsymbol{u}_h,q_h) - b((q_h,\hat{q}_h),\boldsymbol{w}_h) - (Sp_h,q_h) = (g,q_h), \quad \forall(q_h,\hat{q}_h)\in\overline{P}_h. \tag{13c}$$

Note that if we test this system with the test function $((\mathbf{0},\mathbf{0}),\mathbf{0},(0,\hat{q}_h))$, we obtain

$$b((0,\hat{p}_h),\boldsymbol{w}_h) = -\sum_{T\in\mathcal{T}_h}\int_{\partial T}\boldsymbol{w}_h\cdot\boldsymbol{n}\,\hat{p}_h\mathrm{d}s = \sum_{F\in\mathcal{F}_h}\int_F[\boldsymbol{w}_h\cdot n]\hat{p}_h\mathrm{d}s = 0.$$

Hence, by choosing $\hat{q}_h = [\boldsymbol{w}_h\cdot n]$ on each facet $F\in\mathcal{F}_h$, the above equation demonstrates that the velocity solution of (13) is normal continuous, i.e. $\boldsymbol{w}_h\in\boldsymbol{W}_h$.

In the next section, we extend the parameter-robust stability results from [23,45] to the hybridized three-field formulation given by systems (10) and (13).

4. Parameter-robust stability, preconditioners and optimal error estimates

4.1. Parameter-robust well-posedness

4.1.1. Parameter-dependent norms

First, let us recall the norms previously used in the parameter robust stability analysis presented in [23]. These are, for the infinite dimensional spaces $\boldsymbol{U}$, $\boldsymbol{W}$, P,

$$\|\boldsymbol{v}\|_{\boldsymbol{U}}^2 := \|\boldsymbol{\epsilon}(\boldsymbol{v})\|_0^2 + \lambda\|\mathrm{div}\boldsymbol{v}\|_0^2, \tag{14a}$$
$$\|\boldsymbol{z}\|_{\boldsymbol{W}}^2 := R^{-1}\|\boldsymbol{z}\|_0^2 + \gamma^{-1}\|\mathrm{div}\boldsymbol{z}\|_0^2, \tag{14b}$$
$$\|\boldsymbol{z}\|_{\boldsymbol{W}^-}^2 := R^{-1}\|\boldsymbol{z}\|_0^2, \tag{14c}$$
$$\|q\|_P^2 := \gamma\|q\|_0^2, \tag{14d}$$

where the parameter γ can be defined as $\gamma := \lambda_0^{-1} + R + S \eqsim \max\{\lambda_0^{-1}, R, S\}$, with $\lambda_0 = \max\{1,\lambda\} \eqsim 1+\lambda$, or exactly as in [23] where γ has been defined as $\gamma := \max\{(\min\{\lambda, R^{-1}\})^{-1}, S\}$. Due to the non-conformity of the DG discretization, the norm for the discrete displacement space $\boldsymbol{U}_h$ is based on the standard DG norm

$$\|\boldsymbol{v}_h\|_{\mathrm{DG}}^2 := \sum_{T\in\mathcal{T}_h}\|\nabla\boldsymbol{v}_h\|_{0,T}^2 + \sum_{F\in\mathcal{F}_h}h_F^{-1}\|[(\boldsymbol{v}_h)_t]\|_{0,F}^2 + \sum_{T\in\mathcal{T}_h}h_T^2|\boldsymbol{v}_h|_{2,T}^2 \tag{15}$$

and defined by

$$\|\boldsymbol{v}_h\|_{\boldsymbol{U}_h}^2 := \|\boldsymbol{v}_h\|_{\mathrm{DG}}^2 + \lambda\|\mathrm{div}\boldsymbol{v}_h\|_0^2. \tag{16}$$

Next, we introduce the hybridized discontinuous Galerkin (HDG) norm

$$\|(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)\|_{\mathrm{HDG}}^2 := \sum_{T\in\mathcal{T}_h}\left(\|\nabla\boldsymbol{v}_h\|_{0,T}^2 + h_T^{-1}\|(\hat{\boldsymbol{v}}_h-\boldsymbol{v}_h)_t\|_{0,\partial T}^2 + h_T^2|\boldsymbol{v}_h|_{2,T}^2\right), \tag{17}$$

based on which we can define a discrete norm on the extended displacement space $\overline{\boldsymbol{U}}_h$, i.e.,

$$\|(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)\|_{\overline{\boldsymbol{U}}_h}^2 := \|(\boldsymbol{v}_h,\hat{\boldsymbol{v}}_h)\|_{\mathrm{HDG}}^2 + \lambda\|\mathrm{div}\boldsymbol{v}_h\|_0^2. \tag{18}$$

Moreover, we define the following discrete norm on the extended pressure space $\overline{P}_h$

$$\|(q_h,\hat{q}_h)\|_{\mathrm{HDG}}^2 := \sum_{T\in\mathcal{T}_h}\left(\|\nabla q_h\|_{0,T}^2 + h_T^{-1}\|\hat{q}_h-q_h\|_{0,\partial T}^2 + h_T^2|q_h|_{2,T}^2\right), \tag{19a}$$
$$\|(q_h,\hat{q}_h)\|_{\overline{P}_h}^2 := R\|(q_h,\hat{q}_h)\|_{\mathrm{HDG}}^2 + \gamma\|q_h\|_0^2, \tag{19b}$$

where

$$\gamma = S + \frac{1}{\lambda_0} \simeq \max\left\{S, \frac{1}{\lambda_0}\right\}. \tag{20}$$

Finally, we consider the following two product spaces

$$\overline{\boldsymbol{X}}_h := \overline{\boldsymbol{U}}_h\times\boldsymbol{W}_h\times P_h, \tag{21a}$$
$$\overline{\overline{\boldsymbol{X}}}_h := \overline{\boldsymbol{U}}_h\times\boldsymbol{W}_h^-\times\overline{P}_h \tag{21b}$$

equipped with the norms

$$|||((\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h), z_h, q_h)|||^2_{\overline{X}_h} := \|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|^2_{\overline{U}_h} + \|z_h\|^2_{\boldsymbol{W}} + \|q_h\|^2_P, \tag{22a}$$

$$|||((\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h), z_h, (q_h, \hat{q}_h))|||^2_{\overline{\overline{X}}_h} := \|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|^2_{\overline{U}_h} + \|z_h\|^2_{\boldsymbol{W}^-} + \|(q_h, \hat{q}_h)\|^2_{\overline{P}_h} \tag{22b}$$

in the context of problems (10) and (13), respectively.

4.1.2. Uniform well-posedness of the time-discrete problem

The well-posedness of the three-field formulation (3) on the continuous and discrete levels has been addressed and answered in [46–49] using semi-group theory and Galerkin discretization methods. After time discretization by an implicit or semi-implicit time integration scheme, the continuous three-field formulation results in a variational problem of the form: Find $\boldsymbol{x} \in \boldsymbol{X}$ such that

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{F}(\boldsymbol{y}), \qquad \forall \boldsymbol{y} \in \boldsymbol{X} := \boldsymbol{U} \times \boldsymbol{W} \times P, \tag{23}$$

where $\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}) := a(\boldsymbol{u}, \boldsymbol{v}) - (p, \mathrm{div}\boldsymbol{v}) + (R^{-1}\boldsymbol{w}, \boldsymbol{z}) - (p, \mathrm{div}\boldsymbol{z}) - (\mathrm{div}\boldsymbol{u}, q) - (\mathrm{div}\boldsymbol{w}, q) - (Sp, q)$ and $\mathcal{F}(\cdot) \in \boldsymbol{X}'$ denotes a corresponding linear form, which depends on the time integrator.

As it is well known, the abstract variational problem (23) is well-posed under the following necessary and sufficient conditions, see [50].

Theorem 1. *Assume that $\mathcal{F} \in \boldsymbol{X}'$ and the bilinear form $\mathcal{A}(\cdot, \cdot)$ in (23) satisfies the following conditions:*

- *$\mathcal{A}(\cdot, \cdot)$ is bounded, i.e., there exists a constant $C > 0$ such that*

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{y}) \leq C |||\boldsymbol{x}|||_{\boldsymbol{X}} |||\boldsymbol{y}|||_{\boldsymbol{X}} \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}; \tag{24}$$

- *There exists a constant $\beta > 0$ such that*

$$\inf_{\boldsymbol{x} \in \boldsymbol{X}} \sup_{\boldsymbol{y} \in \boldsymbol{X}} \frac{\mathcal{A}(\boldsymbol{x}, \boldsymbol{y})}{|||\boldsymbol{x}|||_{\boldsymbol{X}} |||\boldsymbol{y}|||_{\boldsymbol{X}}} \geq \beta > 0. \tag{25}$$

Then there exists a unique solution $\boldsymbol{x}^* \in \boldsymbol{X}$ of the variational problem (23). Further, the solution satisfies the stability estimate

$$|||\boldsymbol{x}^*|||_{\boldsymbol{X}} \leq \frac{1}{\beta} \sup_{\boldsymbol{y} \in \boldsymbol{X}} \frac{\mathcal{F}(\boldsymbol{y})}{|||\boldsymbol{y}|||_{\boldsymbol{X}}} =: \frac{1}{\beta} \|\mathcal{F}\|_{\boldsymbol{X}'}.$$

Besides for the establishment of well-posedness on the continuous and discrete levels, boundedness, i.e., property (24), and inf–sup stability, i.e., property (25), is crucial in the error analysis and for the construction of preconditioners and iterative solution methods for the algebraic problems arising from the discretization of (23). Furthermore, aiming at parameter-independent error, or near-best approximation estimates and parameter-robust preconditioners, it is essential that the constants C and β in (24) and (25) are independent of any physical (model) and discretization parameters.

Definition 1. We call problem (23) uniformly well-posed on its parameter space (or, in short, uniformly well-posed) under the norm $|||\cdot|||_{\boldsymbol{X}}$ if the conditions of Theorem 1 are satisfied and the constants C and β in (24) and (25) do not depend on any of the problem parameters.

Remark 1. The parameter space is the space of all problem parameters, i.e., physical parameters of the continuous mathematical model but also discretization parameters when $\mathcal{A}(\cdot, \cdot)$ represents a semi- or fully discrete problem.

Uniform well-posedness of the time-discrete problem resulting from the three-field formulation of Biot's consolidation model has first been proven in [23] using the norm

$$|||(\boldsymbol{v}, z, q)|||^2_{\boldsymbol{X}} := \|\boldsymbol{v}\|^2_{\boldsymbol{U}} + \|z\|^2_{\boldsymbol{W}} + \|q\|^2_P \tag{26}$$

where $\|\cdot\|_{\boldsymbol{U}}$, $\|\cdot\|_{\boldsymbol{W}}$, $\|\cdot\|_P$ are defined in (14). In the remainder of Section 4.1, we extend the uniform well-posedness analysis from [23,45] to the three-field formulations (10) and (13).

4.1.3. Hybridized DG method

Following the approach presented in [23], we will show that problem (10) is uniformly well-posed. Initially, we rewrite (10) in the form: Find $\bar{\boldsymbol{x}}_h := ((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), \boldsymbol{w}_h, p_h) \in \overline{\boldsymbol{U}}_h \times \boldsymbol{W}_h \times P_h =: \overline{\boldsymbol{X}}_h$, such that

$$\overline{\mathcal{A}}_h(\bar{\boldsymbol{x}}_h, \bar{\boldsymbol{y}}_h) = \overline{\mathcal{F}}_h(\bar{\boldsymbol{y}}_h), \qquad \forall \bar{\boldsymbol{y}}_h \in \overline{\boldsymbol{X}}_h, \tag{27}$$

where with $\bar{\boldsymbol{y}}_h := ((\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h), \boldsymbol{z}_h, q_h)$ we have

$$\begin{aligned}\overline{\mathcal{A}}_h(\bar{\boldsymbol{x}}_h, \bar{\boldsymbol{y}}_h) :=& a_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)) - (p_h, \mathrm{div}\boldsymbol{v}_h) + (R^{-1}\boldsymbol{w}_h, \boldsymbol{z}_h) \\ &- (p_h, \mathrm{div}\boldsymbol{z}_h) - (\mathrm{div}\boldsymbol{u}_h, q_h) - (\mathrm{div}\boldsymbol{w}_h, q_h) - (Sp_h, q_h), \qquad (28a)\\ \overline{\mathcal{F}}_h(\bar{\boldsymbol{y}}_h) :=& (\boldsymbol{f}, \boldsymbol{v}_h) + (g, q_h), \qquad (28b)\end{aligned}$$

and $a_h((\cdot,\cdot),(\cdot,\cdot))$ is defined in (11). Next, we recall two auxiliary results crucial for establishing the main result of this subsection.

Lemma 2. *The following discrete inf–sup condition*

$$\inf_{q_h \in P_h} \sup_{(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h) \in \overline{\boldsymbol{U}}_h} \frac{(\mathrm{div}\boldsymbol{v}_h, q_h)}{\|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\mathrm{HDG}} \|q_h\|_0} \geq \bar{\beta}_{S,d} > 0, \tag{29}$$

holds where $\|\cdot\|_{\mathrm{HDG}}$ *is the HDG norm defined in* (17).

Proof. As shown, for example, in [51,52], the following inf–sup condition holds true:

$$\inf_{q_h \in P_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{U}_h} \frac{(\mathrm{div}\boldsymbol{v}_h, q_h)}{\|\boldsymbol{v}_h\|_{\mathrm{DG}} \|q_h\|_0} \geq \beta_{S,d} > 0. \tag{30}$$

Moreover, for all $\boldsymbol{v}_h \in \boldsymbol{U}_h$ there exists $\hat{\boldsymbol{v}}_h \in \widehat{\boldsymbol{U}}_h$ such that $\|\boldsymbol{v}_h\|_{\mathrm{DG}} \geq C\|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\mathrm{HDG}}$ with a constant C depending only on mesh regularity. Combining the latter estimate with (30) yields (29).

The proof of the following theorem also makes use of the boundedness and coercivity of the bilinear form $a_h^{\mathrm{HDG}}((\cdot,\cdot),(\cdot,\cdot))$ on $\overline{\boldsymbol{U}}_h$ defined in (9), i.e.,

$$|a_h^{\mathrm{HDG}}((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h))| \leq C_a \|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\mathrm{HDG}} \|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\mathrm{HDG}} \tag{31}$$

for all $(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h) \in \overline{\boldsymbol{U}}_h$ and

$$a_h^{\mathrm{HDG}}((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)) \geq C_c \|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\mathrm{HDG}}^2 \quad \text{for all } (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h) \in \overline{\boldsymbol{U}}_h, \tag{32}$$

see e.g. [42,44].

Theorem 3. *Problem* (27)–(28) *is uniformly well-posed under the norm* $|||\cdot|||_{\overline{X}_h}$ *defined in* (22a)*, that is,*

$$\overline{\mathcal{A}}(\bar{\boldsymbol{x}}_h, \bar{\boldsymbol{y}}_h) \leq \overline{C} |||\bar{\boldsymbol{x}}_h|||_{\overline{X}_h} |||\bar{\boldsymbol{y}}_h|||_{\overline{X}_h} \qquad \forall \bar{\boldsymbol{x}}_h, \bar{\boldsymbol{y}}_h \in \overline{\boldsymbol{X}}_h, \tag{33}$$

$$\inf_{\bar{\boldsymbol{x}}_h \in \overline{\boldsymbol{X}}_h} \sup_{\bar{\boldsymbol{y}}_h \in \overline{\boldsymbol{X}}_h} \frac{\overline{\mathcal{A}}(\bar{\boldsymbol{x}}_h, \bar{\boldsymbol{y}}_h)}{|||\bar{\boldsymbol{x}}_h|||_{\overline{X}_h} |||\bar{\boldsymbol{y}}_h|||_{\overline{X}_h}} \geq \bar{\beta} > 0. \tag{34}$$

Proof. To show (33) one uses Cauchy–Schwarz inequality, the continuity of the bilinear form $a_h((\cdot,\cdot),(\cdot,\cdot))$ in the norm $\|\cdot\|_{\overline{U}_h}$ on $\overline{\boldsymbol{U}}_h$, i.e.,

$$|a_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h))| \leq \overline{C}_a \|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\overline{U}_h} \|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\overline{U}_h} \quad \forall (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h) \in \overline{\boldsymbol{U}}_h, \tag{35}$$

which follows from (31) as well as the definitions of $a_h((\cdot,\cdot),(\cdot,\cdot))$, $\|\cdot\|_{\overline{U}_h}$ and $|||\cdot|||_{\overline{X}_h}$, see (11), (18) and (22a), respectively.

The proof of (34) follows exactly the lines of the proof of Theorem 4.4 in [23] replacing the DG bilinear form (8) by the HDG bilinear form (9) and the DG norm (15) by the HDG norm (17).

4.1.4. Hybridized DG/hybrid mixed method

Consider the HDG/hybrid mixed method for the three-field formulation as stated in (13). To prove the uniform well-posedness of this fully discrete problem, as we did with (10), we rewrite (13) in the form: Find $\overline{\overline{\boldsymbol{x}}}_h := ((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), \boldsymbol{w}_h, (p_h, \hat{p}_h)) \in \overline{\boldsymbol{U}}_h \times \boldsymbol{W}_h \times \overline{P}_h =: \overline{\overline{\boldsymbol{X}}}_h$ such that

$$\overline{\overline{\mathcal{A}}}_h(\overline{\overline{\boldsymbol{x}}}_h, \overline{\overline{\boldsymbol{y}}}_h) = \overline{\overline{\mathcal{F}}}_h(\overline{\overline{\boldsymbol{y}}}_h), \qquad \forall \overline{\overline{\boldsymbol{y}}}_h \in \overline{\overline{\boldsymbol{X}}}_h, \tag{36}$$

where with $\overline{\overline{\boldsymbol{y}}}_h := ((\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h), \boldsymbol{z}_h, (q_h, \hat{q}_h))$ we have

$$\begin{aligned}\overline{\overline{\mathcal{A}}}_h(\overline{\overline{\boldsymbol{x}}}_h, \overline{\overline{\boldsymbol{y}}}_h) :=& a_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)) - (p_h, \operatorname{div}\boldsymbol{v}_h) + (R^{-1}\boldsymbol{w}_h, \boldsymbol{z}_h) \\ &- b((p_h, \hat{p}_h), \boldsymbol{z}_h) - (\operatorname{div}\boldsymbol{u}_h, q_h) - b((q_h, \hat{q}_h), \boldsymbol{w}_h) - (Sp_h, q_h), \end{aligned} \tag{37a}$$

$$\overline{\overline{\mathcal{F}}}_h(\overline{\overline{\boldsymbol{y}}}_h) := (\boldsymbol{f}, \boldsymbol{v}_h) + (g, q_h), \tag{37b}$$

and $a_h((\cdot,\cdot),(\cdot,\cdot))$ and $b((\cdot,\cdot),\cdot)$ are defined in (11) and (12), respectively. Before proving the main theorem, we need another auxiliary result given by the following lemma.

Lemma 4. *There holds the following discrete inf–sup condition*

$$\inf_{(q_h,\hat{q}_h)\in\overline{P}_h} \sup_{\boldsymbol{z}_h\in\boldsymbol{V}_h} \frac{b((q_h,\hat{q}_h),\boldsymbol{z}_h)}{\|\boldsymbol{z}_h\|_0 \|(q_h,\hat{q}_h)\|_{\mathrm{HDG}}} \geq \overline{\beta}_{D,d} > 0 \tag{38}$$

where $\|(\cdot,\cdot)\|_{\mathrm{HDG}}$ *is defined in* (19a).

Proof. A direct proof of (38) can be readily constructed, similarly as for the inf–sup condition in [53,54], using the definition of the degrees of freedom for the Raviart–Thomas space, see [40], and standard scaling arguments.

Such inf–sup conditions with mesh-dependent norms are widely used in structural mechanics, see, e.g., [55].

Theorem 5. *Problem* (36)–(37) *is uniformly well-posed under the norm* $|||\cdot|||_{\overline{\overline{\boldsymbol{X}}}_h}$ *defined in* (22b).

Proof. We start with proving the boundedness of the bilinear form $\overline{\overline{\mathcal{A}}}_h(\cdot,\cdot)$, i.e.,

$$\overline{\overline{\mathcal{A}}}_h(\overline{\overline{\boldsymbol{x}}}_h, \overline{\overline{\boldsymbol{y}}}_h) \leq \overline{\overline{C}} |||\overline{\overline{\boldsymbol{x}}}_h|||_{\overline{\overline{\boldsymbol{X}}}_h} |||\overline{\overline{\boldsymbol{y}}}_h|||_{\overline{\overline{\boldsymbol{X}}}_h} \qquad \forall \overline{\overline{\boldsymbol{x}}}_h, \overline{\overline{\boldsymbol{y}}}_h \in \overline{\overline{\boldsymbol{X}}}_h. \tag{39}$$

First we note that

$$\begin{aligned} b((p_h,\hat{p}_h),\boldsymbol{w}_h) &= \sum_{T\in\mathcal{T}_h} \int_T \operatorname{div}\boldsymbol{w}_h p_h \,\mathrm{d}x - \int_{\partial T} \boldsymbol{w}_h\cdot\boldsymbol{n}\,\hat{p}_h \,\mathrm{d}s \\ &= \sum_{T\in\mathcal{T}_h} \int_T -\boldsymbol{w}_h\cdot\nabla p_h \,\mathrm{d}x - \int_{\partial T} \boldsymbol{w}_h\cdot\boldsymbol{n}(\hat{p}_h - p_h)\,\mathrm{d}s \\ &= \sum_{T\in\mathcal{T}_h} \int_T -\boldsymbol{w}_h\cdot\nabla p_h \,\mathrm{d}x - \int_{\partial T} h\boldsymbol{w}_h\cdot\boldsymbol{n}\frac{1}{h}(\hat{p}_h - p_h)\,\mathrm{d}s \\ &\leq \sqrt{\|\boldsymbol{w}_h\|_0^2 + \sum_{T\in\mathcal{T}_h} h\|\boldsymbol{w}_h\cdot n\|_{\partial T}} \|(p_h,\hat{p}_h)\|_{\mathrm{HDG}} \\ &\leq C_b \|\boldsymbol{w}_h\|_0 \|(p_h,\hat{p}_h)\|_{\mathrm{HDG}}, \end{aligned} \tag{40}$$

where we have used standard scaling arguments in the last step of (40), i.e. the constant C_b depends only on the mesh regularity.

Further, using the continuity of the bilinear form $a_h((\cdot,\cdot),(\cdot,\cdot))$ on $\overline{\boldsymbol{U}}_h$, i.e. (35), the definitions of the norms $\|\cdot\|_{\overline{\boldsymbol{U}}_h}$, $\|\cdot\|_W$, $\|\cdot\|_{\overline{P}_h}$, and $|||\cdot|||_{\overline{\overline{\boldsymbol{X}}}_h}$, see (18), (14b), (19b) and (22b), respectively, and applying the Cauchy–Schwarz

inequality and also estimate (40), one gets

$$\begin{aligned}
\overline{\overline{\mathcal{A}}}_h(\overline{\overline{\boldsymbol{x}}}_h, \overline{\overline{\boldsymbol{y}}}_h) :=& a_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)) - (p_h, \mathrm{div}\boldsymbol{v}_h) + (R^{-1}\boldsymbol{w}_h, \boldsymbol{z}_h) \\
&- b((p_h, \hat{p}_h), \boldsymbol{z}_h) - (\mathrm{div}\boldsymbol{u}_h, q_h) - b((q_h, \hat{q}_h), \boldsymbol{w}_h) - (Sp_h, q_h) \\
\le& C_a \|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\overline{U}_h} \|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\overline{U}_h} + \lambda^{-1/2}\|p_h\|_0 \lambda^{1/2}\|\mathrm{div}\boldsymbol{v}_h\|_0 \\
&+ R^{-1/2}\|\boldsymbol{w}_h\|_0 R^{-1/2}\|\boldsymbol{z}_h\|_0 + C_b R^{1/2}\|(p_h, \hat{p}_h)\|_{\mathrm{HDG}} R^{-1/2}\|\boldsymbol{z}_h\|_0 \\
&+ \lambda^{1/2}\|\mathrm{div}\boldsymbol{u}_h\|_0 \lambda^{-1/2}\|q_h\|_0 + C_b R^{1/2}\|(q_h, \hat{q}_h)\|_{\mathrm{HDG}} R^{-1/2}\|\boldsymbol{w}_h\|_0 \\
&+ S^{1/2}\|p_h\|_0 S^{1/2}\|q_h\|_0 \\
\le& C_a \|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\overline{U}_h} \|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\overline{U}_h} + \|(p_h, \hat{p}_h)\|_{\overline{P}_h} \|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\overline{U}_h} \\
&+ \|\boldsymbol{w}_h\|_{W^-} \|\boldsymbol{z}_h\|_{W^-} + C_b \|(p_h, \hat{p}_h)\|_{\overline{P}_h} \|\boldsymbol{z}_h\|_{W^-} \\
&+ \|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\overline{U}_h} \|(q_h, \hat{q}_h)\|_{\overline{P}_h} + C_b \|(q_h, \hat{q}_h)\|_{\overline{P}_h} \|\boldsymbol{w}_h\|_{W^-} \\
&+ \|(p_h, \hat{p}_h)\|_{\overline{P}_h} \|(q_h, \hat{q}_h)\|_{\overline{P}_h} \\
\le& \overline{\overline{C}} \left(\|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\overline{U}_h} + \|\boldsymbol{w}_h\|_{W^-} + \|(p_h, \hat{p}_h)\|_{\overline{P}_h} \right) \\
&\times \left(\|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\overline{U}_h} + \|\boldsymbol{z}_h\|_{W^-} + \|(q_h, \hat{q}_h)\|_{\overline{P}_h} \right).
\end{aligned} \tag{41}$$

Next we prove the inf–sup condition

$$\inf_{\overline{\overline{\boldsymbol{x}}}_h \in \overline{\overline{\boldsymbol{X}}}_h} \sup_{\overline{\overline{\boldsymbol{y}}}_h \in \overline{\overline{\boldsymbol{X}}}_h} \frac{\overline{\overline{\mathcal{A}}}_h(\overline{\overline{\boldsymbol{x}}}_h, \overline{\overline{\boldsymbol{y}}}_h)}{|||\overline{\overline{\boldsymbol{x}}}_h|||_{\overline{\overline{\boldsymbol{X}}}_h} |||\overline{\overline{\boldsymbol{y}}}_h|||_{\overline{\overline{\boldsymbol{X}}}_h}} \ge \overline{\overline{\beta}} > 0 \tag{42}$$

which immediately follows if for all $\overline{\overline{\boldsymbol{x}}}_h \in \overline{\overline{\boldsymbol{X}}}_h$ we can find $\overline{\overline{\boldsymbol{y}}}_h = \overline{\overline{\boldsymbol{y}}}_h(\overline{\overline{\boldsymbol{x}}}_h)$ such that

$$|||\overline{\overline{\boldsymbol{y}}}_h|||_{\overline{\overline{\boldsymbol{X}}}_h} \le \overline{\overline{C}}_b |||\overline{\overline{\boldsymbol{x}}}_h|||_{\overline{\overline{\boldsymbol{X}}}_h} \tag{43}$$

and the coercivity estimate

$$\overline{\overline{\mathcal{A}}}_h(\overline{\overline{\boldsymbol{x}}}_h, \overline{\overline{\boldsymbol{y}}}_h) \ge \overline{\overline{C}}_c |||\overline{\overline{\boldsymbol{x}}}_h|||^2_{\overline{\overline{\boldsymbol{X}}}_h} \tag{44}$$

are simultaneously satisfied with constants $\overline{\overline{C}}_b$ and $\overline{\overline{C}}_c$ independent of all problem parameters.

Now let $\overline{\overline{\boldsymbol{x}}}_h \in \overline{\overline{\boldsymbol{X}}}_h$ be arbitrary but fixed. Then we choose $\overline{\overline{\boldsymbol{y}}}_h := ((\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h), \boldsymbol{z}_h, (q_h, \hat{q}_h))$ by setting

$$(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h) := \delta(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h) - \frac{1}{\sqrt{\lambda_0}}(\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0}), \tag{45a}$$

$$\boldsymbol{z}_h := \delta\boldsymbol{w}_h + R\boldsymbol{w}_{h,0}, \tag{45b}$$

$$(q_h, \hat{q}_h) := -\delta(p_h, \hat{p}_h), \tag{45c}$$

where $(\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0}) \in \overline{\boldsymbol{U}}_h$ is such that

$$\mathrm{div}\boldsymbol{u}_{h,0} = \frac{1}{\sqrt{\lambda_0}} p_h, \tag{46a}$$

$$\|(\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0})\|_{\mathrm{HDG}} \le \bar{\beta}^{-1}_{S,d} \frac{1}{\sqrt{\lambda_0}} \|p_h\|_0 \tag{46b}$$

and $\boldsymbol{w}_{h,0}$ is such that

$$-b((p_h, \hat{p}_h), \boldsymbol{w}_{h,0}) = \|(p_h, \hat{p}_h)\|^2_{\mathrm{HDG}}, \tag{47a}$$

$$\|\boldsymbol{w}_{h,0}\|_0 \le \overline{\beta}^{-1}_{D,d} \|(p_h, \hat{p}_h)\|_{\mathrm{HDG}}. \tag{47b}$$

Note that the existence of $(\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0})$ and $\boldsymbol{w}_{h,0}$ satisfying the estimates (46) and (47) follows from the discrete inf–sup conditions (29) and (38). With this particular choice, we first verify (43). To begin with

$$\begin{aligned}
\|\frac{1}{\sqrt{\lambda_0}}(\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0})\|^2_{\overline{U}_h} =& \|\frac{1}{\sqrt{\lambda_0}}(\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0})\|^2_{\mathrm{HDG}} + \lambda_0(\mathrm{div}\left(\frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_{h,0}\right), \mathrm{div}\left(\frac{1}{\sqrt{\lambda_0}}\boldsymbol{u}_{h,0}\right)) \\
\overset{(46b)}{\leq}& \frac{1}{\lambda_0}\overline{\beta}^{-2}_{S,d}\frac{1}{\lambda_0}\|p_{h,0}\|^2_0 + \frac{1}{\lambda_0}\|p_h\|^2_0 \\
\leq& \left(\frac{1}{\lambda_0}\overline{\beta}^{-2}_{S,d} + 1\right)\gamma\|p_h\|^2_0 \\
\leq& \left(\frac{1}{\lambda_0}\overline{\beta}^{-2}_{S,d} + 1\right)\|(p_h, \hat{p}_h)\|^2_{\overline{P}_h},
\end{aligned}$$

from which we conclude

$$\|(\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)\|_{\overline{U}_h} \leq \delta\|(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)\|_{\overline{U}_h} + (\overline{\beta}^{-2}_{S,d} + 1)^{\frac{1}{2}}\|(p_h, \hat{p}_h)\|_{\overline{P}_h}. \tag{48}$$

Next,

$$\begin{aligned}
\|\boldsymbol{z}_h\|_{W^-} \leq& \delta\|\boldsymbol{w}_h\|_{W^-} + R\|\boldsymbol{w}_{h,0}\|_{W^-} \\
\leq& \delta\|\boldsymbol{w}_h\|_{W^-} + \sqrt{R}\|\boldsymbol{w}_{h,0}\|_0 \\
\overset{(47b)}{\leq}& \delta\|\boldsymbol{w}_h\|_{W^-} + \sqrt{R}\overline{\beta}^{-1}_{D,d}\|(p_h, \hat{p}_h)\|_{HDG} \\
\leq& \delta\|\boldsymbol{w}_h\|_{W^-} + \overline{\beta}^{-1}_{D,d}\|(p_h, \hat{p}_h)\|_{\overline{P}_h}.
\end{aligned} \tag{49}$$

Finally,

$$\|(q_h, \hat{q}_h)\|_{\overline{P}_h} \leq \delta\|(p_h, \hat{p}_h)\|_{\overline{P}_h}. \tag{50}$$

The bounds (48)–(50) together imply (43) with $\overline{\overline{C}}_b = [2(\delta^2 + \overline{\beta}^{-2}_{S,d} + \overline{\beta}^{-2}_{D,d} + 1)]^{\frac{1}{2}}$.

What remains is to verify (44):

$$\begin{aligned}
\overline{\overline{\mathcal{A}}}_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), \boldsymbol{w}_h, (p_h, \hat{p}_h)) =& a^{\mathrm{HDG}}_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)) + \lambda(\mathrm{div}\boldsymbol{u}_h, \mathrm{div}\boldsymbol{v}_h) \\
&- (p_h, \mathrm{div}\boldsymbol{v}_h) + R^{-1}(\boldsymbol{w}_h, \boldsymbol{z}_h) - b((p_h, \hat{p}_h), \boldsymbol{z}_h) - (\mathrm{div}\boldsymbol{u}_h, q_h) \\
&- b((q_h, \hat{q}_h), \boldsymbol{w}_h) - (Sp_h, q_h) \\
=& \delta a^{\mathrm{HDG}}_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)) - \frac{1}{\sqrt{\lambda_0}}a^{HDG}_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0})) \\
&+ \delta\lambda(\mathrm{div}\boldsymbol{u}_h, \mathrm{div}\boldsymbol{u}_h) - \frac{\lambda}{\sqrt{\lambda_0}}(\mathrm{div}\boldsymbol{u}_h, \mathrm{div}\boldsymbol{u}_{h,0}) - \delta(p_h, \mathrm{div}\boldsymbol{u}_h) \\
&+ \frac{1}{\sqrt{\lambda_0}}(p_h, \mathrm{div}\boldsymbol{u}_{h,0}) + \delta R^{-1}(\boldsymbol{w}_h, \boldsymbol{w}_h) + (\boldsymbol{w}_h, \boldsymbol{w}_{h,0}) \\
&+ R\|(p_h, \hat{p}_h)\|^2_{\mathrm{HDG}} + \delta(\mathrm{div}\boldsymbol{u}_h, p_h) + \delta(Sp_h, p_h) \\
\geq& \delta a^{\mathrm{HDG}}_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)) - \frac{1}{2}\frac{1}{\lambda_0}\varepsilon_1^{-1}a^{\mathrm{HDG}}_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)) \\
&- \frac{1}{2}\varepsilon_1 a^{\mathrm{HDG}}_h((\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0}), (\boldsymbol{u}_{h,0}, \hat{\boldsymbol{u}}_{h,0})) + \delta\lambda(\mathrm{div}\boldsymbol{u}_h, \mathrm{div}\boldsymbol{u}_h) \\
&- \frac{1}{2}\varepsilon_2^{-1}\lambda(\mathrm{div}\boldsymbol{u}_h, \mathrm{div}\boldsymbol{u}_h) - \frac{1}{2}\varepsilon_2\frac{\lambda}{\lambda_0}(\mathrm{div}\boldsymbol{u}_{h,0}, \mathrm{div}\boldsymbol{u}_{h,0}) \\
&+ \frac{1}{\lambda_0}(p_h, p_h) + \delta R^{-1}(\boldsymbol{w}_h, \boldsymbol{w}_h) - \frac{1}{2}\varepsilon_3^{-1}R^{-1}(\boldsymbol{w}_h, \boldsymbol{w}_h) \\
&- \frac{1}{2}\varepsilon_3 R(\boldsymbol{w}_{h,0}, \boldsymbol{w}_{h,0}) + R\|(p_h, \hat{p}_h)\|^2_{\mathrm{HDG}} + \delta(Sp_h, p_h) \\
\geq& \left(\delta - \frac{1}{2}\frac{1}{\lambda_0}\varepsilon_1^{-1}\right)a^{\mathrm{HDG}}_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)) + \left(\delta - \frac{1}{2}\varepsilon_2^{-1}\right)\lambda(\mathrm{div}\boldsymbol{u}_h, \mathrm{div}\boldsymbol{u}_h)
\end{aligned}$$

$$+\left(\delta S+\frac{1}{\lambda_0}-\frac{1}{2}\varepsilon_2\frac{\lambda}{\lambda_0^2}-\frac{1}{2}\frac{1}{\lambda_0}\varepsilon_1 C_a\overline{\beta}_{S,d}^{-2}\right)(p_h,p_h)$$
$$+\left(\delta-\frac{1}{2}\varepsilon_3^{-1}\right)R^{-1}(\boldsymbol{w}_h,\boldsymbol{w}_h)+\left(1-\frac{1}{2}\varepsilon_3\overline{\beta}_{D,d}^{-2}\right)R\|(p_h,\hat{p}_h)\|_{\mathrm{HDG}}^2.$$

By choosing $\varepsilon_1=\frac{1}{2}C_a^{-1}\overline{\beta}_{S,d}^2$, $\varepsilon_2=\frac{1}{2}$, $\varepsilon_3=\overline{\beta}_{D,d}^2$ the last inequality becomes

$$\begin{aligned}\overline{\overline{\mathcal{A}}}_h((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),\boldsymbol{w}_h,(p_h,\hat{p}_h))\geq&(\delta-C_a\overline{\beta}_{S,d}^{-2})a_h^{\mathrm{HDG}}((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),(\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h))+(\delta-1)\lambda(\mathrm{div}\boldsymbol{u}_h,\mathrm{div}\boldsymbol{u}_h)\\&+\left(\delta S+\frac{1}{\lambda_0}-\frac{1}{4}\frac{1}{\lambda_0}-\frac{1}{4}\frac{1}{\lambda_0}\right)(p_h,p_h)\\&+\left(\delta-\frac{1}{2}\frac{1}{\overline{\beta}_{D,d}^2}\right)R^{-1}(\boldsymbol{w}_h,\boldsymbol{w}_h)+\left(1-\frac{1}{2}\right)R\|(p_h,\hat{p}_h)\|_{\mathrm{HDG}}^2.\end{aligned}$$

For $\delta\geq\max\left\{\frac{3}{2},\frac{1}{2C_c}+C_a\overline{\beta}_{S,d}^{-2},\frac{1}{2}+\frac{\overline{\beta}_{D,d}^{-2}}{2}\right\}$, we finally obtain

$$\begin{aligned}\overline{\overline{\mathcal{A}}}_h((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h),\boldsymbol{w}_h,(p_h,\hat{p}_h))\geq&\frac{1}{2}\left(\|(\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h)\|_{\mathrm{HDG}}^2+\lambda\|\mathrm{div}\boldsymbol{u}_h\|^2\right.\\&\left.+\left(S+\frac{1}{\lambda_0}\right)\|p_h\|_0^2+R\|(p_h,\hat{p}_h)\|_{\mathrm{HDG}}^2+R^{-1}\|\boldsymbol{w}_h\|_0^2\right)\\=&\frac{1}{2}\left(\|(\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h)\|_{\overline{U}_h}^2+\|(p_h,\hat{p}_h)\|_{\overline{P}_h}^2+\|\boldsymbol{w}_h\|_{W^-}^2\right),\end{aligned}$$

utilizing $a_h^{\mathrm{HDG}}((\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h))\geq C_c\|(\boldsymbol{u}_h,\hat{\boldsymbol{u}}_h)\|_{\mathrm{HDG}}^2$.

4.2. Uniform preconditioners

The results from the previous subsection imply a "mapping property" that is the basis for defining uniform preconditioners. Here, we discuss norm-equivalent (block-diagonal) preconditioners which fall into this category.

Consider a uniformly well-posed problem of the form (23) where $\mathcal{A}:\boldsymbol{X}\to\boldsymbol{X}'$ is a linear operator, i.e., $\mathcal{A}\in\mathcal{L}(\boldsymbol{X},\boldsymbol{X}')$, $\mathcal{F}\in\boldsymbol{X}'$ for a given Hilbert space $\boldsymbol{X}$, e.g., $\boldsymbol{X}:=\boldsymbol{U}\times\boldsymbol{W}\times\boldsymbol{P}$ or $\boldsymbol{X}:=\overline{\boldsymbol{X}}_h=\overline{\boldsymbol{U}}_h\times\boldsymbol{W}_h\times P_h$, or $\boldsymbol{X}:=\overline{\overline{\boldsymbol{X}}}_h=\overline{\boldsymbol{U}}_h\times\boldsymbol{W}_h\times\overline{P}_h$. Here we assume that $\mathcal{A}$ and $\mathcal{F}$ are defined via the bilinear and linear forms $\mathcal{A}(\cdot,\cdot)$, $\mathcal{F}(\cdot)$, or $\overline{\mathcal{A}}_h(\cdot,\cdot)$, $\overline{\mathcal{F}}_h(\cdot)$, or $\overline{\overline{\mathcal{A}}}_h(\cdot,\cdot)$, $\overline{\overline{\mathcal{F}}}_h(\cdot)$, cf. (23), (28), (37). Let us write Eq. (23) in operator form, i.e.,

$$\mathcal{A}\boldsymbol{x}=\mathcal{F}\in\boldsymbol{X}' \tag{51}$$

and define the linear operator $\mathcal{B}:\boldsymbol{X}'\to\boldsymbol{X}$, i.e., $\mathcal{B}\in\mathcal{L}(\boldsymbol{X}',\boldsymbol{X})$ by

$$(\mathcal{B}\mathcal{G},\boldsymbol{y})_X=\langle\mathcal{G},\boldsymbol{y}\rangle,\qquad\forall\mathcal{G}\in\boldsymbol{X}',\boldsymbol{y}\in\boldsymbol{X}, \tag{52}$$

where $(\cdot,\cdot)_X$ is the inner product inducing the norm $\|\cdot\|_X$, that is, $\|\boldsymbol{y}\|_X=(\boldsymbol{y},\boldsymbol{y})_X^{\frac{1}{2}}$, or, equivalently, $\mathcal{B}^{-1}:\boldsymbol{X}\to\boldsymbol{X}'$, $\mathcal{B}^{-1}\in\mathcal{L}(\boldsymbol{X},\boldsymbol{X}')$ by

$$\langle\mathcal{B}^{-1}\boldsymbol{x},\boldsymbol{y}\rangle=(\boldsymbol{x},\boldsymbol{y})_X,\qquad\forall\boldsymbol{x},\boldsymbol{y}\in\boldsymbol{X}, \tag{53}$$

which implies

$$\langle\mathcal{B}^{-1}\boldsymbol{x},\boldsymbol{x}\rangle=(\boldsymbol{x},\boldsymbol{x})_X=\|\boldsymbol{x}\|_X^2,\qquad\forall\boldsymbol{x}\in\boldsymbol{X}. \tag{54}$$

In practice, the latter relation is often replaced by the weaker condition

$$\langle\mathcal{B}^{-1}\boldsymbol{x},\boldsymbol{x}\rangle\eqsim\|\boldsymbol{x}\|_X^2, \tag{55}$$

for which reason the preconditioner $\mathcal{B}$ is also referred to as a norm-equivalent preconditioner, cf. [56]. The symbol "$\eqsim$" stands for a norm equivalence, uniform with respect to all problem parameters.

Since (24) and (25) are in the norm $\|\cdot\|_X$, we conclude for the operators $\mathcal{B}\mathcal{A}\in\mathcal{L}(\boldsymbol{X},\boldsymbol{X})$ and $(\mathcal{B}\mathcal{A})^{-1}\in\mathcal{L}(\boldsymbol{X},\boldsymbol{X})$ the following bounds:

$$\|\mathcal{B}\mathcal{A}\|_{\mathcal{L}(X,X)}=\sup_{\boldsymbol{x},\boldsymbol{y}}\frac{(\mathcal{B}\mathcal{A}\boldsymbol{x},\boldsymbol{y})_X}{\|\boldsymbol{x}\|_X\|\boldsymbol{y}\|_X}=\sup_{\boldsymbol{x},\boldsymbol{y}}\frac{\langle\mathcal{A}\boldsymbol{x},\boldsymbol{y}\rangle}{\|\boldsymbol{x}\|_X\|\boldsymbol{y}\|_X}=\sup_{\boldsymbol{x},\boldsymbol{y}}\frac{\mathcal{A}(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}\|_X\|\boldsymbol{y}\|_X}\leq C, \tag{56}$$

$$\left(\|(\mathcal{BA})^{-1}\|_{\mathcal{L}(\boldsymbol{X},\boldsymbol{X})}\right)^{-1} = \inf_{\boldsymbol{x}} \left(\frac{1}{\sup_{\boldsymbol{y}} \frac{((\mathcal{BA})^{-1}\boldsymbol{x},\boldsymbol{y})_X}{\|\boldsymbol{x}\|_X \|\boldsymbol{y}\|_X}} \right) = \inf_{\boldsymbol{x}} \sup_{\boldsymbol{y}} \frac{(\mathcal{BA}\boldsymbol{x},\boldsymbol{y})_X}{\|\boldsymbol{x}\|_X \|\boldsymbol{y}\|_X}$$
$$= \inf_{\boldsymbol{x}} \sup_{\boldsymbol{y}} \frac{\langle \mathcal{A}\boldsymbol{x},\boldsymbol{y}\rangle}{\|\boldsymbol{x}\|_X \|\boldsymbol{y}\|_X} = \inf_{\boldsymbol{x}} \sup_{\boldsymbol{y}} \frac{\mathcal{A}(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}\|_X \|\boldsymbol{y}\|_X} \geq \beta. \tag{57}$$

Finally, (56) and (57) together imply that the condition number κ of the preconditioned operator $\mathcal{BA} \in \mathcal{L}(\boldsymbol{X}, \boldsymbol{X})$ is uniformly bounded by a constant that does not depend on any problem parameters, i.e.,

$$\kappa(\mathcal{BA}) := \|\mathcal{BA}\|_{\mathcal{L}(\boldsymbol{X},\boldsymbol{X})} \|(\mathcal{BA})^{-1}\|_{\mathcal{L}(\boldsymbol{X},\boldsymbol{X})} \leq \frac{C}{\beta}. \tag{58}$$

4.3. Optimal error estimates

The uniform well-posedness that we have established in Theorem 5 for the hybridized/hybrid mixed discretization implies near best approximation estimates, which we state next. For the following statements let $(\boldsymbol{u}, \boldsymbol{w}, p)$ be the exact solution of the continuous problem (3) assuming that

$$\boldsymbol{u} \in \boldsymbol{U} := \boldsymbol{H}_0^1(\Omega) \cap \boldsymbol{H}^2(\mathcal{T}_h), \quad \boldsymbol{w} \in \boldsymbol{W} := \boldsymbol{H}_0(\mathrm{div}, \Omega) \cap \boldsymbol{H}^1(\mathcal{T}_h), \quad \text{and} \quad p \in P := H^1(\Omega) \cap H^2(\mathcal{T}_h), \tag{59}$$

where $\boldsymbol{H}^m(\mathcal{T}_h) := \{v \in \boldsymbol{L}^2(\Omega) : v|_T \in \boldsymbol{H}^m(T) \ \forall T \in \mathcal{T}_h\}$ is the broken Sobolev space of order m. Further let $\overline{\boldsymbol{u}} := (\boldsymbol{u}, \hat{\boldsymbol{u}})$ and $\overline{p} := (p, \hat{p})$ with $\hat{\boldsymbol{u}} := \boldsymbol{u}|_{\mathcal{F}_h}$ and $\hat{p} := p|_{\mathcal{F}_h}$.

Theorem 6. *Consider problem (36)–(37) as a discretization of the continuous problem (3) in three-field formulation and assume that the exact solution fulfills (59). Then the following near-best approximation result holds with constants $\overline{\overline{C}}_{uv}, \overline{\overline{C}}_p$ independent of all problem parameters:*

$$\|\overline{\boldsymbol{u}} - \overline{\boldsymbol{u}}_h\|_{\overline{\boldsymbol{U}}_h} + \|\boldsymbol{w} - \boldsymbol{w}_h\|_{\boldsymbol{W}^-} \leq \overline{\overline{C}}_{uv} \left(\inf_{\overline{\boldsymbol{v}}_h \in \overline{\boldsymbol{U}}_h} \|\overline{\boldsymbol{u}} - \overline{\boldsymbol{v}}_h\|_{\overline{\boldsymbol{U}}_h} + \inf_{\boldsymbol{z}_h \in \boldsymbol{W}_h^-} \|\boldsymbol{w} - \boldsymbol{z}_h\|_{\boldsymbol{W}^-} \right) \tag{60}$$

$$\|\overline{p} - \overline{p}_h\|_{\overline{P}_h} \leq \overline{\overline{C}}_p \left(\inf_{\overline{\boldsymbol{v}}_h \in \overline{\boldsymbol{U}}_h} \|\overline{\boldsymbol{u}} - \overline{\boldsymbol{v}}_h\|_{\overline{\boldsymbol{U}}_h} + \inf_{\boldsymbol{z}_h \in \boldsymbol{W}_h^-} \|\boldsymbol{w} - \boldsymbol{z}_h\|_{\boldsymbol{W}^-} + \inf_{\overline{q}_h \in \overline{P}_h} \|\overline{p} - \overline{q}_h\|_{\overline{P}_h} \right) \tag{61}$$

Proof. Let $\Pi^U : \boldsymbol{U} \to \overline{\boldsymbol{U}}_h$, and $\Pi^W : \boldsymbol{W} \to \boldsymbol{W}_h^-$, and $\Pi^P : P \to \overline{P}_h$ denote the canonical interpolation operators onto the discrete spaces. Next, let us define $\overline{\boldsymbol{u}}_I = (\boldsymbol{u}_I, \hat{\boldsymbol{u}}_I) := \Pi^U \overline{\boldsymbol{u}}$, and $\boldsymbol{w}_I := \Pi^W \boldsymbol{w}$, and $\overline{p}_I = (p_I, \hat{p}_I) := \Pi^P \overline{p}$. Then, using the consistency of a_h^{HDG} and following the arguments of the proof of [23, Theorem 5.2], it can easily be seen that due to discrete stability there hold the estimates

$$\|\overline{\boldsymbol{u}}_I - \overline{\boldsymbol{u}}_h\|_{\overline{\boldsymbol{U}}_h} + \|\boldsymbol{w}_I - \boldsymbol{w}_h\|_{\boldsymbol{W}^-} \leq C \left(\sup_{\overline{\boldsymbol{v}}_h \in \overline{\boldsymbol{U}}_h} \frac{a_h^{\mathrm{HDG}}(\overline{\boldsymbol{u}}_I - \overline{\boldsymbol{u}}, \overline{\boldsymbol{v}}_h)}{\|\overline{\boldsymbol{v}}_h\|_{\overline{\boldsymbol{U}}_h}} + \sup_{\boldsymbol{z}_h \in \boldsymbol{W}_h^-} \frac{R^{-1}(\boldsymbol{w}_I - \boldsymbol{w}, \boldsymbol{z}_h)}{\|\boldsymbol{z}_h\|_{\boldsymbol{W}_h^-}} + \sup_{\overline{q}_h \in \overline{P}_h} \frac{b(\overline{q}_h, \boldsymbol{w} - \boldsymbol{w}_I)}{\|\overline{q}_h\|_{\overline{P}_h}} \right),$$

$$\|\overline{p}_I - \overline{p}_h\|_{\overline{\boldsymbol{U}}_h} \leq C \left(\sup_{\overline{\boldsymbol{v}}_h \in \overline{\boldsymbol{U}}_h} \frac{a_h^{\mathrm{HDG}}(\overline{\boldsymbol{u}}_I - \overline{\boldsymbol{u}}, \overline{\boldsymbol{v}}_h)}{\|\overline{\boldsymbol{v}}_h\|_{\overline{\boldsymbol{U}}_h}} + \sup_{\boldsymbol{z}_h \in \boldsymbol{W}_h^-} \frac{R^{-1}(\boldsymbol{w}_I - \boldsymbol{w}, \boldsymbol{z}_h)}{\|\boldsymbol{z}_h\|_{\boldsymbol{W}_h^-}} + \sup_{\overline{q}_h \in \overline{P}_h} \frac{b(\overline{q}_h, \boldsymbol{w} - \boldsymbol{w}_I)}{\|\overline{q}_h\|_{\overline{P}_h}} \right).$$

Finally, applying triangle inequality and using continuity of $a_h^{\mathrm{HDG}}(\cdot,\cdot)$, the Cauchy–Schwarz inequality and the fact that $b(\overline{q}_h, \boldsymbol{w} - \boldsymbol{w}_I) = 0$, which is a consequence of the properties of the Raviart–Thomas interpolator Π^W, the assertions of the theorem follow.

Remark 2. An analogous result to Theorem 6 is also valid for the discrete problem (27)–(28) if one replaces the spaces $\boldsymbol{W}^-$, $\boldsymbol{W}_h^-$, $\overline{P}_h$ by $\boldsymbol{W}$, $\boldsymbol{W}_h$ and P_h and the corresponding norms $\|\cdot\|_{\boldsymbol{W}^-}$ and $\|\cdot\|_{\overline{P}_h}$ by $\|\cdot\|_{\boldsymbol{W}}$ and $\|\cdot\|_P$. The result is then a consequence of Theorem 3.

In the following let $\overline{\Pi}_{P_h}(\cdot) = (\Pi_{P_h}(\cdot), \Pi_{\hat{P}_h}(\cdot)) \in \overline{P}_h$ be the standard element and facet-wise L^2-projection. Using the proper, well known (see [40,42,57]) interpolation operators and standard arguments, one can derive the following optimal error estimates from the above best approximation results.

Theorem 7. *Consider problem (36)–(37) as a discretization of the continuous problem (3) in three-field formulation. Beside (59) we assume that the exact solution fulfills the regularity estimate* $(\boldsymbol{u}, \underline{\boldsymbol{w}}, p) \in \boldsymbol{H}^m(\mathcal{T}_h) \times \boldsymbol{H}^{m-1}(\mathcal{T}_h) \times H^{m-1}(\mathcal{T}_h)$. *Then there hold the following error estimates with a constants* $\overline{\overline{C}}_{e,uv}, \overline{\overline{C}}_{e,p}$ *independent of all problem parameters:*

$$\|\overline{\boldsymbol{u}} - \overline{\boldsymbol{u}}_h\|_{\overline{U}_h} + \|\boldsymbol{w} - \boldsymbol{w}_h\|_{W^-} + \|\overline{\Pi}_{P_h}\overline{p} - \overline{p}_h\|_{\overline{P}_h} \leq \overline{\overline{C}}_{e,uv} h^s (|\boldsymbol{u}|_{\boldsymbol{H}^{s+1}(\mathcal{T}_h)} + \lambda^{\frac{1}{2}} |\operatorname{div}(\boldsymbol{u})|_{\boldsymbol{H}^s(\mathcal{T}_h)} + R^{-\frac{1}{2}} |\boldsymbol{w}|_{\boldsymbol{H}^s(\mathcal{T}_h)})$$

and

$$\begin{aligned}\|\overline{p} - \overline{p}_h\|_{\overline{P}_h} \leq \overline{\overline{C}}_{e,p} h^{s-1} (&|\boldsymbol{u}|_{\boldsymbol{H}^s(\mathcal{T}_h)} + \lambda^{\frac{1}{2}} |\operatorname{div}(\boldsymbol{u})|_{\boldsymbol{H}^{s-1}(\mathcal{T}_h)} \\ &+ R^{-\frac{1}{2}} |\boldsymbol{w}|_{\boldsymbol{H}^{s-1}(\mathcal{T}_h)} + R^{\frac{1}{2}} |p|_{H^s(\mathcal{T}_h)} + \gamma^{\frac{1}{2}} |p|_{H^{s-1}(\mathcal{T}_h)}).\end{aligned}$$

where $s := \min\{l, m-1\}$.

Proof. The proof is based on Theorem 6 and the application of the Bramble–Hilbert Lemma to bound the interpolation errors in the right-hand sides of Eqs. (60) and (61) , cf. [58, Proposition 2.3.10]. Similar estimates can also be found in [44].

Remark 3. Assuming enough regularity of the exact solution, Theorem 7 shows that the projected error $\|\overline{\Pi}_{P_h}\overline{p} - \overline{p}_h\|_{\overline{P}_h}$ converges with one order higher than $\|\overline{p} - \overline{p}_h\|_{\overline{P}_h}$. This super convergence property of (hybrid) mixed methods is well known in the literature, see for example [28,29].

4.4. Implementation aspects and static condensation

In order to solve the discrete system, we employ static condensation of the local element-wise degrees of freedom. These are given by the dof introduced through the discontinuous approximation spaces $\boldsymbol{W}_h^-$ and P_h. One can also eliminate the local $\boldsymbol{H}(\operatorname{div})$-conforming element bubbles of the space $\boldsymbol{U}_h$. However, for ease of representation, we only consider the lowest order case $l = 1$, hence, no bubbles for the displacement are present. In the following, we use the same symbols $\overline{\boldsymbol{u}}_h := (\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h)$, $\boldsymbol{w}_h$, p_h and $\hat{p}_h$ for the representation of the coefficients of the corresponding discrete finite element solutions. Then (13) can be written as

$$\begin{pmatrix} A_{\overline{u}} & 0 & B_u^{\mathrm{T}} & 0 \\ 0 & M_w & B_w^{\mathrm{T}} & \hat{B}_w^{\mathrm{T}} \\ B_u & B_w & -M_p & 0 \\ 0 & \hat{B}_w & 0 & 0 \end{pmatrix} \begin{pmatrix} \overline{\boldsymbol{u}}_h \\ \boldsymbol{w}_h \\ p_h \\ \hat{p}_h \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_h \\ \boldsymbol{0} \\ g_h \\ 0 \end{pmatrix},$$

where $\boldsymbol{f}_h$ represent the corresponding vector of the right hand side $(\boldsymbol{f}, \boldsymbol{v}_h)$ and g_h the vector of (g, q_h). Further, $A_{\overline{u}}$, B_u, M_w, M_p, B_w and $\hat{B}_w$ denote the operators, or their corresponding matrix representations, defined via the bilinear forms $a_h((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h))$, $(-\operatorname{div}\boldsymbol{u}_h, q_h)$, $(R^{-1}w_h, z_h)$, (Sp_h, q_h), $b((q_h, 0), \boldsymbol{w}_h)$ and $b((0, \hat{q}_h), \boldsymbol{w}_h)$, respectively. From the second line we see that we can eliminate $\boldsymbol{w}_h$ using $\boldsymbol{w}_h = M_w^{-1}(-B_w^{\mathrm{T}} p_h - \hat{B}_w^{\mathrm{T}} \hat{p}_h)$. Then the third line gives $p_h = -(M_p + B_w M_w^{-1} B_w^{\mathrm{T}})^{-1}(-B_u \overline{\boldsymbol{u}}_h + B_w M_w^{-1} \hat{B}_w^{\mathrm{T}} \hat{p}_h)$. Thus, we have the following system to solve

$$\begin{pmatrix} A & B^{\mathrm{T}} \\ B & -C \end{pmatrix} \begin{pmatrix} \overline{\boldsymbol{u}}_h \\ \hat{p}_h \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_h \\ g_h \end{pmatrix}, \tag{62}$$

with

$$\begin{aligned} A &:= A_{\overline{u}} + B_u^{\mathrm{T}}(M_p + B_w M_w^{-1} B_w^{\mathrm{T}})^{-1} B_u, \\ B &:= -\hat{B}_w M_w^{-1} B_w^{\mathrm{T}}(M_p + B_w M_w^{-1} B_w^{\mathrm{T}})^{-1} B_u, \\ C &:= \hat{B}_w M_w^{-1} B_w^{\mathrm{T}}(M_p + B_w M_w^{-1} B_w^{\mathrm{T}})^{-1} B_w M_w^{-1} \hat{B}_w^{\mathrm{T}} + \hat{B}_w M_w^{-1} \hat{B}_w^{\mathrm{T}}. \end{aligned}$$

Note that M_w, M_p and $(M_p + B_w M_w^{-1} B_w^{\mathrm{T}})$ are all block diagonal and locally (element-wise) invertible. Since these systems are very small they are solved by a direct method. Further, the operator C is equivalent to a (scaled) H^1-like norm on $\hat{P}_h$. By means of norm equivalent preconditioning, cf. Eq. (55), we now follow two different approaches. The first preconditioner we investigate is based on a block system that decouples mechanics from the flow problem, and, additionally, the velocity from the fluid pressure. The latter is achieved by introducing an HDG bilinear form on $\overline{P}_h$ for the discretization of $\mathrm{div}(R\nabla p)$ as given in the original equation (1) (where K was replaced due to scaling by R). Henceforth, let $\tilde{M}_p$ denote the matrix representation of the scaled bilinear form $(\gamma p_h, q_h)$. Then we define the operator

$$\mathcal{B} := \begin{pmatrix} A_{\overline{u}} & 0 & 0 & 0 \\ 0 & M_w & 0 & 0 \\ 0 & 0 & -\tilde{M}_p - A_p & -B_p^{\mathrm{T}} \\ 0 & 0 & -B_p & -A_{\hat{p}} \end{pmatrix}^{-1}.$$

where A_p, B_p and $A_{\hat{p}}$ correspond to the bilinear forms given by

$$a_p(p_h, q_h) := R \sum_{T\in\mathcal{T}_h} \int_T \nabla p_h \cdot \nabla q_h \mathrm{d}x + \int_{\partial T} (-\nabla p_h \cdot n q_h - \nabla q_h \cdot n p_h) + \eta_p l^2 h^{-1} \eta_p p_h q_h \mathrm{d}s,$$

$$b_p(p_h, \hat{q}_h) := R \sum_{T\in\mathcal{T}_h} \int_{\partial T} \nabla p_h \cdot n \hat{q}_h - \eta_p l^2 h^{-1} p_h \hat{q}_h \mathrm{d}s,$$

$$a_{\hat{p}}(\hat{p}_h, \hat{q}_h) := R \sum_{T\in\mathcal{T}_h} \int_{\partial T} \eta_p l^2 h^{-1} \hat{p}_h \hat{q}_h \mathrm{d}s,$$

respectively, where η_p is again a sufficiently large stabilization parameter. Note that the combined bilinear form $a_p(p_h, q_h) + b_p(p_h, \hat{q}_h) + b_p(q_h, \hat{p}_h) + a_{\hat{p}}(\hat{p}_h, \hat{q}_h)$ is the HDG bilinear form mentioned above which is continuous and elliptic with respect to $R\|\cdot\|_{\mathrm{HDG}}$. Similarly, as before, we can eliminate the local (element-wise) variables to obtain the following preconditioner

$$\begin{pmatrix} A_{\overline{u}} & 0 \\ 0 & -(A_{\hat{p}} + B_p(\tilde{M}_p^{-1} + A_p^{-1})B_p^T) \end{pmatrix} \tag{63}$$

for the condensed system (62), where we have again made use of A_p being block diagonal and invertible. Further, note that both blocks on the diagonal are H^1-type systems. Thus, standard solvers, such as, for example, an algebraic multigrid method for the lowest order system and a element-level "balancing domain decomposition with constraints"(BDDC) preconditioner (see [59]), the latter featuring only a polylogarithmic dependence on the polynomial degree l, can be used.

The second block diagonal preconditioner we test still satisfies the norm equivalence (55), but decouples only the mechanics and flow problems, hence, keeps the hybrid mixed formulation of the velocity pressure system. The block diagonal operator preconditioner is then given by

$$\mathcal{B} := \begin{pmatrix} A_{\overline{u}} & 0 & 0 & 0 \\ 0 & M_w & B_w^{\mathrm{T}} & \hat{B}_w^{\mathrm{T}} \\ 0 & B_w & -\tilde{M}_p & 0 \\ 0 & \hat{B}_w & 0 & 0 \end{pmatrix}^{-1}.$$

Following similar steps as above, the preconditioner for the condensed system is

$$\begin{pmatrix} A_{\overline{u}} & 0 \\ 0 & -\tilde{C} \end{pmatrix}, \tag{64}$$

where $\tilde{C}$ is the same as C with M_p replaced by $\tilde{M}_p$. The advantage of the preconditioner defined by (64), as demonstrated below in Section 5.2, is that the subsystem for the pressure variable does not require a stabilization parameter η_p which in general affects the condition number.

5. Numerical results

In this section, we present several numerical examples to validate our theoretical findings. First, we test for the expected orders of convergence for a problem with a constructed solution increasing the degree of the FE approximation. Second, we study the parameter-robustness of the proposed preconditioners. Finally, we discuss the cost efficiency of our modified methods. All numerical examples are implemented within the finite element library Netgen/NGSolve, see [60,61] and www.ngsolve.org.

5.1. Convergence of the hybridized/hybrid mixed method

Here we discuss the convergence orders of the errors of the methods introduced in this work. Note, however, that we only consider the discretization given by (13) since the solution is the same as of (10).

5.1.1. 2D example

We solve problem (13) on the spatial domain $\Omega = (0,1)^2$ and choose the right hand side $\boldsymbol{f}$ and g such that the exact solutions are given by

$$\boldsymbol{u} := (-\partial_y \phi, \partial_x \phi), \quad p := \sin(\pi x)\sin(\pi y) - p_0,$$

with the potential $\phi = x^2(1-x)^2y^2(1-y)^2$ and $p_0 \in \mathbb{R}$ is chosen such that $p \in L^2_0(\Omega)$. For simplicity, we choose the constants $K = 1$, $\mu = 1$, $S_0 = 1$, and $\alpha = 1$. Further, we set $\lambda = c$ with an arbitrary constant $c \in \mathbb{R}^+$ since the exact and discrete solutions are exactly divergence-free.

In Table 1 we have displayed several discrete errors and their estimated order of convergence (eoc) for the discretization of problem (13) for varying polynomial orders $l = 1, 2, 3, 4$. Whereas the H^1-seminorm error of the displacement $\boldsymbol{u}_h$ and the pressure $\boldsymbol{p}_h$ converge with the expected (see Theorem 7) order $\mathcal{O}(h^l)$ and $\mathcal{O}(h^{l-1})$, respectively, the corresponding L^2-norm errors converge with order $\mathcal{O}(h^{l+1})$ and $\mathcal{O}(h^l)$. This can be shown by a standard Aubin–Nitsche duality argument whenever the considered problem is sufficiently regular, see for example [40]. Note also that the L^2-norm error $\|\nabla p + R^{-1}\boldsymbol{w}_h\|_0$ of the discrete velocity $\boldsymbol{w}_h$ converges with optimal order $\mathcal{O}(h^l)$. In the lowest order case where we have a piece-wise constant approximation of the pressure p_h, we do not present the H^1-seminorm error of the pressure since the gradient ∇p_h vanishes locally on each element.

5.1.2. 3D example

We solve problem (13) on the spatial domain $\Omega = (0,1)^3$ and choose the right hand side $\boldsymbol{f}$ and g such that the exact solutions are given by

$$\boldsymbol{u} := \mathrm{curl}(\phi, \phi, \phi), \quad p := \sin(\pi x)\sin(\pi y)\sin(\pi z) - p_0,$$

with the potential $\phi = x^2(1-x)^2y^2(1-y)^2z^2(1-z)^2$ and $p_0 \in \mathbb{R}$ is chosen such that $p \in L^2_0(\Omega)$. The parameters $\lambda, \mu, S_0, \alpha, K$ are chosen as in the two-dimensional example.

Again, we present in Table 2 several discrete errors and their estimated orders of convergence for varying polynomial degree $l = 1, 2, 3$. We make the same observations as for the two-dimensional example, that is, all errors converge with optimal order as predicted by Theorem 7.

5.2. Parameter-robustness of the preconditioners

In this section, we demonstrate the robustness of the preconditioners defined in Section 4.2 with respect to varying physical parameters. Again, we solve the example given in Section 5.1.1 on a fixed triangulation with 384 elements. The system is solved by means of the minimal residual method (MinRes) with a fixed tolerance of 10^{-10} and for different polynomial degrees $l = 1, 2, 3, 4$. In Fig. 1 we plot the number of iterations for the preconditioner defined in (63) with a fixed stabilization parameter $\eta_p = 10$ for variations of the parameters R^{-1}, λ, S. In Fig. 2 we plot the number of iterations for the same example using the preconditioner defined in (64). Although both preconditioners show the expected robustness as predicted by the analysis presented in Section 4.1, we see that the results with (64) demonstrate improvement upon those with (63). While a different (smaller) choice of η_p in (63) might lead to better results – we have fixed $\eta_p = 10$ here – the analysis unfortunately only shows that η_p has to

Table 1
The H^1-seminorm and the L^2-norm errors of the discrete displacement $\boldsymbol{u}_h$ and the discrete pressure p_h and the L^2-norm errors of the discrete velocity $\boldsymbol{w}_h$ for different polynomial degrees $l = 1, 2, 3, 4$ for the two-dimensional example.

$\lvert\mathcal{T}\rvert$	$\lVert\nabla u - \nabla u_h\rVert_0$	(eoc)	$\lVert u - u_h\rVert_0$	(eoc)	$\lVert\nabla p - \nabla p_h\rVert_0$	(eoc)	$\lVert p - p_h\rVert_0$	(eoc)	$\lVert\nabla p + R^{-1}\boldsymbol{w}_h\rVert_0$	(eoc)
$l = 1$										
6	$5.3\cdot 10^{-2}$	–	$5.5\cdot 10^{-3}$	–	–	–	$1.8\cdot 10^{-1}$	–	$9.7\cdot 10^{-1}$	–
24	$4.9\cdot 10^{-2}$	0.1	$4.9\cdot 10^{-3}$	0.2	–	–	$1.8\cdot 10^{-1}$	0.0	$5.7\cdot 10^{-1}$	0.8
96	$2.2\cdot 10^{-2}$	1.1	$1.1\cdot 10^{-3}$	2.2	–	–	$8.1\cdot 10^{-2}$	1.1	$2.8\cdot 10^{-1}$	1.0
384	$1.1\cdot 10^{-2}$	1.0	$2.6\cdot 10^{-4}$	2.1	–	–	$4.0\cdot 10^{-2}$	1.0	$1.4\cdot 10^{-1}$	1.0
1536	$5.4\cdot 10^{-3}$	1.0	$6.3\cdot 10^{-5}$	2.0	–	–	$2.0\cdot 10^{-2}$	1.0	$7.1\cdot 10^{-2}$	1.0
6144	$2.7\cdot 10^{-3}$	1.0	$1.6\cdot 10^{-5}$	2.0	–	–	$1.0\cdot 10^{-2}$	1.0	$3.6\cdot 10^{-2}$	1.0
$l = 2$										
6	$4.6\cdot 10^{-2}$	–	$5.1\cdot 10^{-3}$	–	1.7	–	$1.5\cdot 10^{-1}$	–	$2.9\cdot 10^{-1}$	–
24	$1.1\cdot 10^{-2}$	2.1	$5.9\cdot 10^{-4}$	3.1	$7.4\cdot 10^{-1}$	1.2	$3.1\cdot 10^{-2}$	2.3	$8.2\cdot 10^{-2}$	1.8
96	$3.0\cdot 10^{-3}$	1.8	$7.4\cdot 10^{-5}$	3.0	$3.8\cdot 10^{-1}$	1.0	$7.9\cdot 10^{-3}$	2.0	$2.7\cdot 10^{-2}$	1.6
384	$7.7\cdot 10^{-4}$	2.0	$9.3\cdot 10^{-6}$	3.0	$1.9\cdot 10^{-1}$	1.0	$2.0\cdot 10^{-3}$	2.0	$6.9\cdot 10^{-3}$	2.0
1536	$1.9\cdot 10^{-4}$	2.0	$1.2\cdot 10^{-6}$	3.0	$9.6\cdot 10^{-2}$	1.0	$4.9\cdot 10^{-4}$	2.0	$1.8\cdot 10^{-3}$	2.0
6144	$4.9\cdot 10^{-5}$	2.0	$1.5\cdot 10^{-7}$	3.0	$4.8\cdot 10^{-2}$	1.0	$1.2\cdot 10^{-4}$	2.0	$4.4\cdot 10^{-4}$	2.0
$l = 3$										
6	$8.8\cdot 10^{-3}$	–	$5.1\cdot 10^{-4}$	–	$2.7\cdot 10^{-1}$	–	$6.2\cdot 10^{-3}$	–	$9.7\cdot 10^{-2}$	–
24	$2.3\cdot 10^{-3}$	1.9	$6.7\cdot 10^{-5}$	2.9	$1.4\cdot 10^{-1}$	0.9	$3.9\cdot 10^{-3}$	0.7	$1.0\cdot 10^{-2}$	3.3
96	$3.1\cdot 10^{-4}$	2.9	$4.3\cdot 10^{-6}$	4.0	$3.7\cdot 10^{-2}$	2.0	$4.6\cdot 10^{-4}$	3.1	$1.3\cdot 10^{-3}$	2.9
384	$3.7\cdot 10^{-5}$	3.1	$2.5\cdot 10^{-7}$	4.1	$9.2\cdot 10^{-3}$	2.0	$5.7\cdot 10^{-5}$	3.0	$1.7\cdot 10^{-4}$	3.0
1536	$4.6\cdot 10^{-6}$	3.0	$1.6\cdot 10^{-8}$	4.0	$2.3\cdot 10^{-3}$	2.0	$7.1\cdot 10^{-6}$	3.0	$2.2\cdot 10^{-5}$	3.0
6144	$5.7\cdot 10^{-7}$	3.0	$9.6\cdot 10^{-10}$	4.0	$5.7\cdot 10^{-4}$	2.0	$8.8\cdot 10^{-7}$	3.0	$2.7\cdot 10^{-6}$	3.0
$l = 4$										
6	$3.3\cdot 10^{-3}$	–	$2.6\cdot 10^{-4}$	–	$2.0\cdot 10^{-1}$	–	$2.3\cdot 10^{-3}$	–	$1.0\cdot 10^{-2}$	–
24	$2.8\cdot 10^{-4}$	3.6	$1.0\cdot 10^{-5}$	4.7	$1.9\cdot 10^{-2}$	3.4	$1.1\cdot 10^{-4}$	4.4	$8.3\cdot 10^{-4}$	3.7
96	$1.6\cdot 10^{-5}$	4.1	$2.9\cdot 10^{-7}$	5.1	$2.5\cdot 10^{-3}$	3.0	$7.4\cdot 10^{-6}$	3.9	$5.3\cdot 10^{-5}$	4.0
384	$9.8\cdot 10^{-7}$	4.0	$8.9\cdot 10^{-9}$	5.0	$3.1\cdot 10^{-4}$	3.0	$4.7\cdot 10^{-7}$	4.0	$3.3\cdot 10^{-6}$	4.0
1536	$6.1\cdot 10^{-8}$	4.0	$2.8\cdot 10^{-10}$	5.0	$3.9\cdot 10^{-5}$	3.0	$2.9\cdot 10^{-8}$	4.0	$2.1\cdot 10^{-7}$	4.0
6144	$3.8\cdot 10^{-9}$	4.0	$8.7\cdot 10^{-12}$	5.0	$4.9\cdot 10^{-6}$	3.0	$1.8\cdot 10^{-9}$	4.0	$1.3\cdot 10^{-8}$	4.0

Table 2
The H^1-seminorm and the L^2-norm errors of the discrete displacement $\boldsymbol{u}_h$ and the discrete pressure p_h and the L^2-norm errors of the discrete velocity $\boldsymbol{w}_h$ for different polynomial degrees $l = 1, 2, 3$ for the three-dimensional example.

$\lvert\mathcal{T}\rvert$	$\lVert\nabla u - \nabla u_h\rVert_0$	(eoc)	$\lVert u - u_h\rVert_0$	(eoc)	$\lVert\nabla p - \nabla p_h\rVert_0$	(eoc)	$\lVert p - p_h\rVert_0$	(eoc)	$\lVert\nabla p + R^{-1}\boldsymbol{w}_h\rVert_0$	(eoc)
$l = 1$										
48	$5.2\cdot 10^{-3}$	–	$4.9\cdot 10^{-4}$	–	–	–	$1.8\cdot 10^{-1}$	–	$9.4\cdot 10^{-1}$	–
384	$2.6\cdot 10^{-3}$	1.0	$1.7\cdot 10^{-4}$	1.5	–	–	$9.6\cdot 10^{-2}$	0.9	$4.9\cdot 10^{-1}$	0.9
3 072	$1.3\cdot 10^{-3}$	1.0	$5.2\cdot 10^{-5}$	1.7	–	–	$4.9\cdot 10^{-2}$	1.0	$2.5\cdot 10^{-1}$	1.0
24 576	$6.6\cdot 10^{-4}$	1.0	$1.4\cdot 10^{-5}$	1.9	–	–	$2.5\cdot 10^{-2}$	1.0	$1.3\cdot 10^{-1}$	1.0
$l = 2$										
48	$4.8\cdot 10^{-3}$	–	$4.7\cdot 10^{-4}$	–	1.1	–	$6.4\cdot 10^{-2}$	–	$2.8\cdot 10^{-1}$	–
384	$8.6\cdot 10^{-4}$	2.5	$4.2\cdot 10^{-5}$	3.5	$5.8\cdot 10^{-1}$	0.9	$1.7\cdot 10^{-2}$	1.9	$7.4\cdot 10^{-2}$	1.9
3 072	$1.9\cdot 10^{-4}$	2.2	$3.5\cdot 10^{-6}$	3.6	$3.0\cdot 10^{-1}$	1.0	$4.4\cdot 10^{-3}$	2.0	$1.9\cdot 10^{-2}$	2.0
24 576	$4.6\cdot 10^{-5}$	2.1	$3.4\cdot 10^{-7}$	3.4	$1.5\cdot 10^{-1}$	1.0	$1.1\cdot 10^{-3}$	2.0	$4.8\cdot 10^{-3}$	2.0
$l = 3$										
48	$1.2\cdot 10^{-3}$	–	$7.4\cdot 10^{-5}$	–	$4.4\cdot 10^{-1}$	–	$8.5\cdot 10^{-3}$	–	$5.3\cdot 10^{-2}$	–
384	$1.3\cdot 10^{-4}$	3.2	$3.6\cdot 10^{-6}$	4.4	$1.2\cdot 10^{-1}$	1.9	$1.0\cdot 10^{-3}$	3.0	$6.8\cdot 10^{-3}$	3.0
3 072	$1.5\cdot 10^{-5}$	3.0	$2.1\cdot 10^{-7}$	4.1	$3.1\cdot 10^{-2}$	2.0	$1.3\cdot 10^{-4}$	3.0	$8.6\cdot 10^{-4}$	3.0
24 576	$1.9\cdot 10^{-6}$	3.0	$1.2\cdot 10^{-8}$	4.1	$7.9\cdot 10^{-3}$	2.0	$1.6\cdot 10^{-5}$	3.0	$1.1\cdot 10^{-4}$	3.0

be chosen sufficiently large (see [57]), and its optimal choice is difficult. Therefore, it is obvious that the mixed formulation, which is known to result in a minimal stabilization, as used in (64), is preferable.

Further note that although the definition (63) includes a proper scaling (in terms of l and h) of the interior penalty stabilization parameter given by $\eta_p l^2 h^{-1}$, Fig. 1 shows a mild dependence on l of the preconditioner (63), whereas (64) seems to be more robust (see Fig. 2). Note, however, that we do not claim uniform robustness with respect to l.

Remark 4. Since $R = \frac{2\tilde{\mu}\tau}{\alpha^2}$, the robustness in R, see Figs. 1 and 2, also implies robustness with respect to the time step τ. In a time-dependent problem, with constant time step, only the right hand side will change, which does not deteriorate the performance of the iterative solver. Varying the time step leads to a variation of R, which is studied in Figs. 1–2.

5.3. Cost-efficiency of the new family of hybridized discretizations

5.3.1. DG vs. HDG

In a first step we only illustrate the effect of hybridization introduced Section 3.2. To this end we consider the model problem: Find $\boldsymbol{u} \in \boldsymbol{H}_0^1(\Omega)$ such that

$$-\operatorname{div}(\boldsymbol{\epsilon}(\boldsymbol{u})) = \boldsymbol{f},$$

with a given right hand side $\boldsymbol{f}$ and $\Omega = (0,1)^3$. We solve this problem on a given triangulation with 166 elements either with an $H(\mathrm{div})$-conforming DG or HDG method, i.e. setting $\lambda = 0$ we have the problems: Find $\boldsymbol{u}_h \in \boldsymbol{U}_h$ such that

$$a_h^{\mathrm{DG}}(\boldsymbol{u}_h, \boldsymbol{v}_h) = (f, \boldsymbol{v}_h) \quad \forall \boldsymbol{v}_h \in \boldsymbol{u}_h, \tag{65}$$

and find $(\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h) \in \overline{\boldsymbol{U}}_h$ such that

$$a_h^{\mathrm{HDG}}((\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h), (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h)) = (f, \boldsymbol{v}_h), \quad \forall (\boldsymbol{v}_h, \hat{\boldsymbol{v}}_h) \in \overline{\boldsymbol{U}}_h. \tag{66}$$

In Table 3 we compare the values

`dof`: number of unknowns,
`cdof`: number of coupling unknowns,
`nze`: number of non-zero entries in thousands of the resulting system matrix,

for varying polynomial degrees $l = 1, \ldots, 6$ which correspond to the local order of approximation of $\boldsymbol{u}_h, \hat{\boldsymbol{u}}_h$ in $\mathrm{BDM}_\ell(T)/\mathrm{P}_\ell^\perp(F)$, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_h$. Here $\mathrm{P}_\ell^\perp(F)$ is the space of polynomials of order l that are orthogonal to the normal vector, see the definition of the space $\widehat{\boldsymbol{U}}_h$ in Section 3.2.

First, note that, due to the coupling between element unknowns in the DG method, no static condensation can be applied, i.e. `dof = cdof`. When solving the linear system one is particularly interested in the number of non-zero entries. As we can see, the HDG method clearly outperforms the DG method in case of higher order approximation ($l \geq 4$). In the low order cases the additional facet unknowns dominate and thus no improvement can be expected.

Remark 5. The HDG method can further be improved by means of another technique, called "projected jumps", which was introduced in [42]. This modification allows to further decrease the coupling of the HDG method without affecting its approximation properties. This essentially compensates the overhead of the HDG method in the low order cases by reducing the polynomial degree of the space of $\hat{\boldsymbol{u}}_h$ to $\mathrm{P}_{\ell-1}^\perp(F)$ and adding consistent projections in the bilinear form. Although we do not discuss these modifications here, we include the corresponding numbers in Table 3 in the rows denoted by PHDG. Note that the well-posedness theory and the robustness of the preconditioners obtained in this work also hold for the PHDG method.

5.3.2. Mixed vs. hybrid mixed methods

The aim of this subsection is to compare the sparsity structure of the matrix arising from the application of a hybrid mixed method to the flow subproblem to that of the matrix resulting from its discretization by a standard mixed method, when both are used as building blocks of mass conserving discretizations of the Biot problem. Although hybridization introduces additional degrees of freedom in the linear system, the number of coupling

Table 3
`dof`, `cdof` and `nze` of the system matrix of the DG, HDG and PHDG methods for different polynomial degrees l.

	`dof`	`cdof`	`nze`	`dof`	`cdof`	`nze`	`dof`	`cdof`	`nze`
	$l=1$			$l=2$			$l=3$		
DG	834	834	65	2664	2664	454	6100	6100	1945
HDG	2502	2502	193	6000	5004	770	11 660	8340	2140
PHDG	1390	1390	59	4332	3336	342	9436	6116	1151
	$l=4$			$l=5$			$l=6$		
DG	11 640	11 640	6238	19 782	19 782	16 512	31 024	31 024	38 106
HDG	19 980	12 510	4815	31 458	17 514	9438	46 592	23 352	16 779
PHDG	17 200	9730	2913	28 122	14 178	6185	42 700	19 460	11 652

degrees of freedom is typically smaller when following this approach, which is a crucial advantage for its cost-efficient iterative solution. Hence, we study the effect of the modifications introduced in Section 3.3 on the following Darcy model problem: Find $(\boldsymbol{w}, p) \in \boldsymbol{H}(\mathrm{div})(\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} \boldsymbol{w} + \nabla p &= 0, \\ \mathrm{div}(\boldsymbol{w}) &= g, \end{aligned}$$

for a given right hand side g on the domain $\Omega = (0,1)^3$. We use the same mesh as in the previous section, and consider the problems: Find $(\boldsymbol{w}_h, p_h) \in \boldsymbol{W}_h \times P_h$, such that

$$(\boldsymbol{w}_h, \boldsymbol{z}_h) - (p_h, \mathrm{div}\boldsymbol{z}_h) = 0, \qquad \forall \boldsymbol{z}_h \in \boldsymbol{W}_h, \tag{67a}$$

$$-(\mathrm{div}\boldsymbol{w}_h, q_h) = -(g, q_h), \qquad \forall q_h \in P_h. \tag{67b}$$

and find $(\boldsymbol{w}_h, (p_h, \hat{p}_h)) \in \boldsymbol{W}_h^- \times \overline{P}_h$, such that

$$(\boldsymbol{w}_h, \boldsymbol{z}_h) - b((p_h, \hat{p}_h), \boldsymbol{z}_h) = 0, \qquad \forall \boldsymbol{z}_h \in \boldsymbol{W}_h^-, \tag{68a}$$

$$-b((q_h, \hat{q}_h), \boldsymbol{w}_h) = -(g, q_h), \quad \forall (q_h, \hat{q}_h) \in \overline{P}_h. \tag{68b}$$

Note that system (67) only allows a static condensation of the following degrees of freedom: all local (element-associated) degrees of freedom of the space $\boldsymbol{W}_h$, i.e. element-wise basis functions with a vanishing normal trace, and all high-order (considering a standard L^2-Dubiner basis) basis functions of P_h such that element-wise constant basis functions remain in the system. In contrast to this, system (68) allows us to eliminate all degrees of freedom associated with the basis functions of the spaces $\boldsymbol{W}_h^-$ and P_h. In Table 4, we again present the corresponding numbers as discussed above, where M represents the discretization of (67), and HM of (68). Here, the order l corresponds to the local polynomial degree of $\boldsymbol{w}_h$, p_h, $\hat{p}_h$ in $\mathrm{RT}_\ell(T)/\mathrm{P}_\ell(T)/\mathrm{P}_\ell(F)$, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_h$. Further, we observe that the hybrid mixed method produces always a smaller number of non-zero entries than the standard mixed method although the difference is negligible. However, the main purposes of hybridization are a reduction of the number of coupling dof and obtaining a condensed system with a symmetric and definite Schur complement, see also [41,62]. This allows us to use preconditioners for H^1-elliptic problems like standard algebraic multigrid methods.

6. Concluding remarks

We have introduced a family of higher-order hybridized/hybrid mixed strongly mass conserving discretizations of the three-field formulation of Biot's model of consolidation and proven their uniform well-posedness. The construction relies on a hybridized $\boldsymbol{H}(\mathrm{div})$-conforming discontinuous Galerkin method for the mechanics and a hybrid mixed method for the flow subproblems. The hybridization approach offers the advantages of reducing the number of coupling degrees of freedom and the possibility of a static (element-wise) condensation of all other degrees of freedom, in particular, of all flux degrees of freedom. Additionally, we have constructed parameter-robust norm-equivalent preconditioners that avoid the solution of an $\boldsymbol{H}(\mathrm{div})$ subproblem which is typically the most time consuming part of the application of preconditioners that rely on such. A generalization and application of the presented methodology to non-linear extensions of Biot's model are subject of ongoing research.

Table 4
`dof`, `cdof` and `nze` of the system matrix for different polynomial degrees l in the discretizations of problems (67), (68).

	`dof`	`cdof`	`nze`	`dof`	`cdof`	`nze`	`dof`	`cdof`	`nze`
	$l = 0$			$l = 1$			$l = 2$		
M	552	552	4k	2320	1324	26k	5968	2482	94k
HM	1108	278	2k	3988	834	21k	9304	1668	86k
	$l = 3$			$l = 4$			$l = 5$		
M	12 160	4026	251k	21 560	5956	555k	34 832	8272	1077k
HM	17 720	2780	238k	29 900	4170	535k	46 508	5838	1049k

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The first and third authors acknowledge the support of this work by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the project "Physics-oriented solvers for multicompartmental poromechanics" under grant number 456235063. The second and the last authors acknowledge the support by the Austrian Science Fund (FWF) through the research program "Taming complexity in partial differential systems" (F65) - project "Automated discretization in multiphysics" (P10).

References

[1] M. Biot, General theory of three-dimensional consolidation, J. Appl. Phys. 12 (2) (1941) 155–164.
[2] M. Biot, Theory of elasticity and consolidation for a porous anisotropic solid, J. Appl. Phys. 26 (2) (1955) 182–185.
[3] N. Sebaa, Z. Fellah, M. Fellah, E. Ogam, F. Mitri, C. Depollier, W. Laurikis, Application of the Biot model to ultrasound in bone: Inverse problem, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 55 (7) (2008) 1516–1523.
[4] L. Guo, J. Vardakis, T. Lassila, M. Mitolo, N. Ravikumar, D. Chou, M. Lange, A. Sarrami-Foroushani, B. Tully, Z. Taylor, S. Varma, A. Venneri, A. Frangi, Y. Ventikos, Subject specific multi-poroelastic model for exploring the risk factors associated with the early stages of alzheimer's disease, Interf. Focus 8 (1) (2018) 20170019.
[5] M. Murad, A. Loula, Improved accuracy in finite element analysis of Biot's consolidation problem, Comput. Methods Appl. Mech. Engrg. 95 (1992) 359–382.
[6] M. Murad, A. Loula, On stability and convergence of finite element approximations of Biot's consolidation problem, Internat. J. Numer. Methods Engrg. 37 (1994) 645–667.
[7] C. Rodrigo, X. Hu, P. Ohm, J. Adler, F. Gaspar, L. Zikatanov, New stabilized discretizations for poroelasticity and the Stokes' equations, Comput. Methods Appl. Mech. Engrg. 341 (2018) 467–484.
[8] J. Nordbotten, Stable cell-centered finite volume discretization for Biot equations, SIAM J. Numer. Anal. 54 (2016) 942–968.
[9] J. Lee, K.-A. Mardal, R. Winther, Parameter-robust discretization and preconditioning of Biot's consolidation model, SIAM J. Sci. Comput. 39 (2017) A1–A24.
[10] R. Oyarzúa, R. Ruiz-Baier, Locking-free finite element methods for poroelasticity, SIAM J. Numer. Anal. 54 (2016) 2951–2973.
[11] S. Kumar, R. Oyarzúa, R. Ruiz-Baier, R. Sandilya, Conservative discontinuous finite volume and mixed schemes for a new four-field formulation in poroelasticity, Esaim Math. Model. Numer. Anal. 54 (1) (2020) 273–299.
[12] V. Girault, X. Lu, M. Wheeler, A posteriori error estimates for Biot system using Enriched Galerkin for flow, Comput. Methods Appl. Mech. Engrg. 369 (2020) 113185.
[13] S. Lee, M. Wheeler, Enriched Galerkin methods for two-phase flow in porous media with capillary pressure, J. Comput. Phys. 367 (2018) 65–86.
[14] M. Bause, F. Radu, U. Köcher, Space-time finite element approximation of the Biot poroelasticity system with iterative coupling, Comput. Methods Appl. Mech. Engrg. 320 (2017) 745–768.
[15] F. Radu, K. Kumar, J. Nordbotten, I. Pop, A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities, IMA J. Numer. Anal. 38 (2018) 884–920.
[16] J. Lee, E. Piersanti, K.-A. Mardal, M. Rognes, A mixed finite element method for nearly incompressible multiple-network poroelasticity, SIAM J. Sci. Comput. 41 (2019) A722–A747.
[17] Q. Hong, J. Kraus, M. Lymbery, M. Wheeler, Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems, Multiscale Model. Simul. 18 (2) (2020) 916–941.

[18] P. Phillips, M. Wheeler, A coupling of mixed and continuous Galerkin finite element methods for poroelasticity. I. The continuous in time case, Comput. Geosci. 11 (2) (2007) 131–144.
[19] P. Phillips, M. Wheeler, A coupling of mixed and continuous Galerkin finite element methods for poroelasticity. II. The discrete-in-time case, Comput. Geosci. 11 (2) (2007) 145–158.
[20] P. Phillips, M. Wheeler, A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity, Comput. Geosci. 12 (4) (2008) 417–435.
[21] S.-Y. Yi, A coupling of nonconforming and mixed finite element methods for Biot's consolidation model, Numer. Methods Partial Differ. Equ. 29 (5) (2013) 1749–1777.
[22] X. Hu, C. Rodrigo, F. Gaspar, L. Zikatanov, A nonconforming finite element method for the Biot's consolidation model in poroelasticity, J. Comput. Appl. Math. 310 (2017) 143–154.
[23] Q. Hong, J. Kraus, Parameter-robust stability of classical three-field formulation of Biot's consolidation model, ETNA - Electron. Trans. Numer. Anal. 48 (2018) 202–226.
[24] G. Kanschat, B. Riviere, A finite element method with strong mass conservation for Biot's linear consolidation model, J. Sci. Comput. 77 (2018) 1762–1779.
[25] B. Cockburn, G. Kanschat, D. Schotzau, A locally conservative LDG method for the incompressible Navier-Stokes equations, Math. Comp. 74 (251) (2005) 1067–1095.
[26] B. Cockburn, G. Kanschat, D. Schötzau, C. Schwab, Local discontinuous Galerkin methods for the Stokes system, SIAM J. Numer. Anal. 40 (1) (2002) 319–343.
[27] B. Cockburn, G. Kanschat, D. Schötzau, A note on discontinuous Galerkin divergence-free solutions of the Navier–Stokes equations, J. Sci. Comput. 31 (1–2) (2007) 61–73.
[28] J. Könnö, R. Stenberg, Numerical computations with H(div)-finite elements for the Brinkman problem, Comput. Geosci. 16 (1) (2012) 139–158.
[29] J. Könnö, R. Stenberg, H(div)-conforming finite elements for the brinkman problem, Math. Models Methods Appl. Sci. 21 (11) (2011) 2227–2248.
[30] G. Fu, A high-order HDG method for the Biot's consolidation model, Comput. Math. Appl. 77 (1) (2019) 237–252.
[31] C. Niu, H. Rui, X. Hu, A stabilized hybrid mixed finite element method for poroelasticity, Comput. Geosci. (2020).
[32] D.N. Arnold, F. Brezzi, Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates, RAIRO Modél. Math. Anal. Numér. 19 (1) (1985) 7–32.
[33] J.H. Adler, F.J. Gaspar, X. Hu, C. Rodrigo, L.T. Zikatanov, Robust block preconditioners for Biot's model, in: Domain Decomposition Methods in Science and Engineering XXIV, Springer International Publishing, Cham, 2018, pp. 3–16.
[34] M. Borregales, F.A. Radu, K. Kumar, J.M. Nordbotten, Robust iterative schemes for non-linear poromechanics, Comput. Geosci. 22 (4) (2018) 1021–1038.
[35] M. Rahrah, F. Vermolen, A moving finite element framework for fast infiltration in nonlinear poroelastic media, Comput. Geosci. 25 (2) (2021) 793–804.
[36] M.A. Borregales Reverón, K. Kumar, J.M. Nordbotten, F.A. Radu, Iterative solvers for Biot model under small and large deformations, Comput. Geosci. 25 (2) (2021) 687–699.
[37] M. Neunteufel, A.S. Pechstein, J. Schöberl, Three-field mixed finite element methods for nonlinear elasticity, Comput. Methods Appl. Mech. Engrg. 382 (2021) 113857.
[38] Q. Hong, J. Kraus, M. Lymbery, F. Philo, Parameter-robust uzawa-type iterative methods for double saddle point problems arising in Biot's consolidation and multiple-network poroelasticity models, Math. Models Methods Appl. Sci. 30 (13) (2020) 2523–2555.
[39] K. Terzaghi, Erdbaumechanik Auf Bodenphysikalischer Grundlage, F. Deuticke, 1925.
[40] D. Boffi, F. Brezzi, M. Fortin, Mixed Finite Element Methods and Applications, in: Springer Ser. Comput. Math., vol. 44, Springer, Heidelberg, 2013, p. xiv+685.
[41] B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems, SIAM J. Numer. Anal. 47 (2) (2009) 1319–1365.
[42] C. Lehrenfeld, J. Schöberl, High order exactly divergence-free hybrid discontinuous Galerkin methods for unsteady incompressible flows, Comput. Methods Appl. Mech. Engrg. 307 (2016) 339–361.
[43] P.L. Lederer, C. Lehrenfeld, J. Schöberl, Hybrid discontinuous Galerkin methods with relaxed H(div)-conformity for incompressible flows. Part II, ESAIM Math. Model. Numer. Anal. 53 (2) (2019) 503–522.
[44] P.L. Lederer, C. Lehrenfeld, J. Schöberl, Hybrid discontinuous Galerkin methods with relaxed H(div)-conformity for incompressible flows. Part I, SIAM J. Numer. Anal. 56 (4) (2018) 2070–2094.
[45] Q. Hong, J. Kraus, M. Lymbery, F. Philo, Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models, Numer. Linear Algebra Appl. (2019) e2242.
[46] A. Ženišek, The existence and uniqueness theorem in Biot's consolidation theory, Apl. Mat. 29 (3) (1984) 194–211.
[47] A. Ženišek, Finite element methods for coupled thermoelasticity and coupled consolidation of clay, RAIRO Anal. Numér. 18 (2) (1984) 183–205.
[48] R. Showalter, Diffusion in poro-elastic media, J. Math. Anal. Appl. 251 (1) (2000) 310–340.
[49] E. Hairer, C. Lubich, M. Roche, The Numerical Solution of Differential-Algebraic Systems by Runge-KUtta Methods, in: Lecture Notes in Mathematics, vol. 1409, Springer-Verlag, Berlin, 1989, p. viii+139.
[50] I. Babuška, Error-bounds for finite element method, Numer. Math. 16 (1970/71) 322–333.
[51] Q. Hong, J. Kraus, J. Xu, L. Zikatanov, A robust multigrid method for discontinuous Galerkin discretizations of Stokes and linear elasticity equations, Numer. Math. 132 (2016) 23–49.

[52] Q. Hong, J. Kraus, Uniformly stable discontinuous Galerkin discretization and robust iterative solution methods for the brinkman problem, SIAM J. Numer. Anal. 54 (5) (2016) 2750–2774.
[53] J. Gopalakrishnan, P.L. Lederer, J. Schöberl, A mass conserving mixed stress formulation for the Stokes equations, IMA J. Numer. Anal. 40 (3) (2019) 1838–1874.
[54] J. Gopalakrishnan, P.L. Lederer, J. Schöberl, A mass conserving mixed stress formulation for Stokes flow with weakly imposed stress symmetry, SIAM J. Numer. Anal. 58 (1) (2020) 706–732.
[55] B. Fraeijs de Veubeke, A Course in Elasticity, in: Applied Mathematical Sciences, vol. 29, Springer, Heidelberg, 1979, p. 330.
[56] K.-A. Mardal, R. Winther, Preconditioning discretizations of systems of partial differential equations, Numer. Linear Algebra Appl. 18 (1) (2011) 1–40.
[57] D.N. Arnold, F. Brezzi, B. Cockburn, L.D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM J. Numer. Anal. 39 (5) (2002) 1749–1779.
[58] C. Lehrenfeld, Hybrid discontinuous Galerkin methods for solving incompressible flow problems, Rheinisch-Westfal. Techn. Hochschule Aachen (2010).
[59] J. Schöberl, C. Lehrenfeld, Domain decomposition preconditioning for high order hybrid discontinuous Galerkin methods on tetrahedral meshes, in: T. Apel, O. Steinbach (Eds.), Advanced Finite Element Methods and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 27–56.
[60] J. Schöberl, NETGEN an advancing front 2d/3D-mesh generator based on abstract rules, Comput. Vis. Sci. 1 (1) (1997) 41–52.
[61] J. Schöberl, C++11 Implementation of Finite Elements in NGSolve, Institute for Analysis and Scientific Computing, Vienna University of Technology, 2014.
[62] B. Cockburn, J. Gopalakrishnan, A characterization of hybridized mixed methods for second order elliptic problems, SIAM J. Numer. Anal. 42 (1) (2004) 283–301.

A NEW PRACTICAL FRAMEWORK FOR THE STABILITY ANALYSIS OF PERTURBED SADDLE-POINT PROBLEMS AND APPLICATIONS

MATHEMATICS OF COMPUTATION
https://doi.org/10.1090/mcom/3795
Article electronically published on November 17, 2022

A NEW PRACTICAL FRAMEWORK FOR THE STABILITY ANALYSIS OF PERTURBED SADDLE-POINT PROBLEMS AND APPLICATIONS

QINGGUO HONG, JOHANNES KRAUS, MARIA LYMBERY, AND FADI PHILO

Abstract. In this paper we prove a new abstract stability result for perturbed saddle-point problems based on a norm fitting technique. We derive the stability condition according to Babuška's theory from a small inf-sup condition, similar to the famous Ladyzhenskaya-Babuška-Brezzi (LBB) condition, and the other standard assumptions in Brezzi's theory, in a combined abstract norm. The construction suggests to form the latter from individual *fitted* norms that are composed from proper seminorms.

This abstract framework not only allows for simpler (shorter) proofs of many stability results but also guides the design of parameter-robust norm-equivalent preconditioners. These benefits are demonstrated on mixed variational formulations of generalized Poisson, Stokes, vector Laplace and Biot's equations.

1. Introduction

Saddle-point problems (SPPs) arise in various areas of computational science and engineering ranging from computational fluid dynamics [26, 27, 57], elasticity [4, 15, 24], and electromagnetics [12, 48] to computational finance [40]. Moreover, SPPs play a vital role in the context of image reconstruction [29], model order reduction [56], constrained optimization [25], optimal control [8], and parameter identification [18], to mention only a few but important applications.

In the mathematical modeling of multiphysics phenomena described by (initial-) boundary-value problems for systems of partial differential equations, SPPs often naturally arise and are frequently posed in a variational formulation. Mixed finite element methods and other discretization techniques can be and have been successfully used for their discretization and numerical solution, see, e.g. [9, 12, 17, 23] and the references therein.

The pioneering works laying the foundations of the solution theory for SPPs have been conducted by Ivo Babuška, Franco Brezzi, Olga Ladyzhenskaya, and Jindřich Nečas [5, 16, 42, 51], see also the contributions [6, 41].

Designing and analyzing discretizations and solvers for SPPs require a careful study of the mapping properties of the underlying operators. Of particular interest are their continuity and stability, which not only guarantee the well-posedness of (continuous and discrete) mathematical models but also provide the basis for error

Received by the editor May 27, 2021, and, in revised form, April 27, 2022, and August 28, 2022.

2020 *Mathematics Subject Classification.* Primary 65N12, 65J05, 65F08, 65N30.

The second author and the third author were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the project "Physics-oriented solvers for multicompartmental poromechanics" under grant number 456235063.

estimates and a convergence analysis of iterative methods and preconditioners, see, e.g. [12, 23, 46, 47], for a review see also [9].

Saddle-point problems/systems are of a two-by-two block form and characterized by an operator/matrix of the form

$$\mathcal{A} = \begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix}, \tag{1.1}$$

where A and C denote positive semidefinite operators/matrices and B_1^T the adjoint/transpose of an operator/matrix B_1. We consider the symmetric case in this paper where $A^T = A$, $C^T = C$, and $B_1 = B_2 = B$. Problems in which $C \neq 0$ are often referred to as *perturbed saddle-point problems*.

In [63] a technique has been proposed to determine norms for parameter-dependent SPPs providing necessary and sufficient conditions for their well-posedness and leading to robust estimates of the solution in terms of the data. A drawback of this approach, however, is that these conditions are often hard to verify in practice as the operators inducing the norms are defined only implicitly.

More general SPPs in which A (and C) are allowed to be non-symmetric and $B_1 \neq B_2$ have also been studied by many authors, see, e.g., [12, 17, 52], and the references therein. Their analysis, in general is more complicated and is mostly done following the *monolithic* approach, i.e., imposing conditions on $\mathcal{A}$ rather than on A, B_1, B_2, and C separately.

Our work is motivated by the stability analysis of variational problems occurring in poromechanics (cf. [20]), a subarea of continuum mechanics which originates from the early works of Terzaghi and Biot [10, 58]. Various formulations of Biot's consolidation model have been considered and analyzed since it had been introduced in [10, 11], including two-field [49, 50], three-field [19, 32, 43, 53–55], and four-field-formulations [39, 44, 62], for generalizations to several fluid networks as considered in multiple network poroelastic theory (MPET), see also [7, 28, 33–35, 45, 60].

Although they typically relate more than two physical fields, or quantities of interest (except for the two-field formulation of Biot's model), the variational problems arising from the above-mentioned formulations–subject to a proper grouping or rather aggregation of variables–result in symmetric two-by-two block systems of saddle point form characterized by a self-adjoint operator $\mathcal{A}$.

The abstract framework presented in the next section of this paper applies to such saddle-point operators. After introducing some notation, we recall the classical stability results of Babuška and Brezzi for classical (unperturbed) SPPs. Next, we focus on perturbed (symmetric) SPPs, initially summarizing some of the additional conditions which, together with the Ladyzhenskaya-Babuška-Brezzi (LBB) condition (small inf-sup condition), imply the necessary and sufficient stability condition of Babuška (big inf-sup condition). Our main theoretical result then follows in Section 2.3 where we propose a generalization of the classical Brezzi conditions for the analysis of perturbed SPPs with $C \neq 0$. These new conditions imply the Babuška condition. The fitted norms on which they are based provide a constructive tool for designing norm-equivalent preconditioners.

This paper does not discuss discretizations and discrete variants of inf-sup conditions. However, the proposed framework directly translates to discrete settings where it also allows for shorter and simpler proofs of the well-posedness of discrete models and error-estimates for stable discretizations.

2. Abstract framework

2.1. **Notation and problem formulation.** Consider two Hilbert spaces V and Q equipped with the norms $\|\cdot\|_V$ and $\|\cdot\|_Q$ induced by the scalar products $(\cdot,\cdot)_V$ and $(\cdot,\cdot)_Q$, respectively. We denote their product space by $Y := V \times Q$ and endow it with the norm $\|\cdot\|_Y$ defined by

(2.1)
$$\|y\|_Y^2 = (y,y)_Y = (v,v)_V + (q,q)_Q = \|v\|_V^2 + \|q\|_Q^2 \qquad \forall y = (v;q) := \begin{pmatrix} v \\ q \end{pmatrix} \in Y.$$

Next, we introduce an abstract bilinear form $\mathcal{A}((\cdot\,;\cdot),(\cdot\,;\cdot))$ on $Y \times Y$ defined by

$$\mathcal{A}((u;p),(v;q)) := a(u,v) + b(v,p) + b(u,q) - c(p,q) \tag{2.2}$$

for some symmetric positive semidefinite (SPSD) bilinear forms $a(\cdot,\cdot)$ on $V \times V$, $c(\cdot,\cdot)$ on $Q \times Q$, i.e.,

$$a(u,v) = a(v,u) \qquad \forall u, v \in V, \tag{2.3}$$

$$a(v,v) \geq 0 \qquad \forall v \in V, \tag{2.4}$$

$$c(p,q) = c(q,p) \qquad \forall p, q \in Q, \tag{2.5}$$

$$c(q,q) \geq 0 \qquad \forall q \in Q, \tag{2.6}$$

and a bilinear form $b(\cdot,\cdot)$ on $V \times Q$.

We assume that $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are continuous with respect to the norms $\|\cdot\|_V$ and $\|\cdot\|_Q$, i.e.,

$$a(u,v) \leq \bar{C}_a \|u\|_V \|v\|_V \qquad \forall u, v \in V, \tag{2.7}$$

$$b(v,q) \leq \bar{C}_b \|v\|_V \|q\|_Q \qquad \forall v \in V, \forall q \in Q, \tag{2.8}$$

$$c(p,q) \leq \bar{C}_c \|p\|_Q \|q\|_Q \qquad \forall p, q \in Q. \tag{2.9}$$

Then each of these bilinear forms defines a bounded linear operator as follows:

$$A : V \to V' : \langle Au, v\rangle_{V'\times V} = a(u,v), \qquad \forall u, v \in V, \tag{2.10a}$$

$$C : Q \to Q' : \langle Cp, q\rangle_{Q'\times Q} = c(p,q), \qquad \forall p, q \in Q, \tag{2.10b}$$

$$B : V \to Q' : \langle Bv, q\rangle_{Q'\times Q} = b(v,q), \qquad \forall v \in V, \forall q \in Q, \tag{2.10c}$$

$$B^T : Q \to V' : \langle v, B^T q\rangle_{V\times V'} = b(v,q), \qquad \forall v \in V, \forall q \in Q. \tag{2.10d}$$

Here, V' and Q' denote the dual spaces of V and Q and $\langle\cdot,\cdot\rangle$ the corresponding duality pairings.

Associated with the bilinear form $\mathcal{A}$ defined in (2.2) we consider the following abstract perturbed saddle-point problem

$$\mathcal{A}((u;p),(v;q)) = \mathcal{F}((v;q)) \qquad \forall v \in V, \forall q \in Q \tag{2.11}$$

which can also be written as

$$\mathcal{A}(x,y) = \mathcal{F}(y) \qquad \forall y \in Y,$$

thereby using the definitions $x = (u;p)$ and $y = (v;q)$, or, in operator form

$$\mathcal{A}x = \mathcal{F}, \tag{2.12}$$

where

$$\mathcal{A} : Y \to Y' : \langle \mathcal{A}x, y\rangle_{Y'\times Y} = \mathcal{A}(x,y), \qquad \forall x, y \in Y \tag{2.13}$$

and $\mathcal{F} \in Y'$, i.e., $\mathcal{F} : Y \to \mathbb{R} : \mathcal{F}(y) = \langle \mathcal{F}, y\rangle_{Y'\times Y}$ for all $y \in Y$.

The operator $\mathcal{A}$ can also be represented in block form by

$$\mathcal{A} = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}. \tag{2.14}$$

Problem (2.11) (and (2.12)) is called a perturbed saddle-point problem (in operator form) when $c(\cdot,\cdot) \not\equiv 0$ and a classical saddle-point problem in the case $c(\cdot,\cdot) \equiv 0$.

2.2. **Babuška's and Brezzi's conditions for stability of saddle-point problems.** As is well-known from [5], the abstract variational problem (2.11) is well-posed under the following necessary and sufficient conditions 2.15 and 2.16 given in Theorem 2.1.

Theorem 2.1 (Babuška [5])**.** *Let $\mathcal{F} \in Y'$ be a bounded linear functional. Then the saddle-point problem* (2.11) *is well-posed if and only if there exist positive constants $\bar{C}$ and α for which the conditions*

$$\mathcal{A}(x,y) \le \bar{C}\|x\|_Y\|y\|_Y \qquad \forall x, y \in Y, \tag{2.15}$$

$$\inf_{x\in Y}\sup_{y\in Y}\frac{\mathcal{A}(x,y)}{\|x\|_Y\|y\|_Y} \ge \underline{\alpha} > 0 \tag{2.16}$$

hold. The solution x then satisfies the stability estimate

$$\|x\|_Y \le \frac{1}{\underline{\alpha}}\sup_{y\in Y}\frac{\mathcal{F}(y)}{\|y\|_Y} =: \frac{1}{\underline{\alpha}}\|\mathcal{F}\|_{Y'}.$$

Remark 2.2. Estimate (2.15) ensures continuity, that is, boundedness of the operator $\mathcal{A}$ from above, whereas (2.16) is a stability condition, sometimes referred to as Babuška condition, which grants boundedness of $\mathcal{A}$ from below.

Using the operator notations introduced in (2.14), the conditions (2.15) and (2.16) can be rewritten as

$$\underline{\alpha}\|y\|_Y \le \|\mathcal{A}y\|_{Y'} \le \bar{C}\|y\|_Y \text{ for all } y \in Y. \tag{2.17}$$

In [63], the condition (2.17) is characterized by two equivalent conditions as stated in Theorem 2.3.

Theorem 2.3 (Zulehner [63])**.** *If there are constants $\underline{\gamma}_v, \bar{\gamma}_v, \underline{\gamma}_q, \bar{\gamma}_q > 0$ such that*

$$\underline{\gamma}_v\|v\|_V^2 \le a(v,v) + \left(\sup_{q\in Q}\frac{b(v,q)}{\|q\|_Q}\right)^2 \le \bar{\gamma}_v\|v\|_V^2 \quad \forall v \in V, \tag{2.18}$$

and

$$\underline{\gamma}_q\|q\|_Q^2 \le c(q,q) + \left(\sup_{v\in V}\frac{b(v,q)}{\|v\|_V}\right)^2 \le \bar{\gamma}_q\|q\|_Q^2 \quad \forall q \in Q, \tag{2.19}$$

then (2.17) *is satisfied with constants $\underline{\alpha}, \bar{C} > 0$ that depend only on $\underline{\gamma}_v, \bar{\gamma}_v, \underline{\gamma}_q, \bar{\gamma}_q$. And, vice versa, if the estimates* (2.17) *are satisfied with constants $\underline{\alpha}, \bar{C} > 0$, then the estimates* (2.18) *and* (2.19) *are satisfied with constants $\underline{\gamma}_v, \bar{\gamma}_v, \underline{\gamma}_q, \bar{\gamma}_q > 0$ that depend only on $\underline{\alpha}, \bar{C} > 0$.*

Remark 2.4. The two conditions (2.15) and (2.16), entangling the bilinear forms $a(\cdot,\cdot), b(\cdot,\cdot)$ and $c(\cdot,\cdot)$, are equivalent to the two conditions (2.18) and (2.19) which entangle the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ and the bilinear forms $c(\cdot,\cdot)$ and $b(\cdot,\cdot)$,

respectively. However, verifying the two big conditions (2.18) and (2.19) is sometimes difficult or even impractical. Our aim is to propose a framework untangling the bilinear forms $a(\cdot,\cdot), b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ and impose Brezzi-type conditions, in particular a small coercivity condition on $a(\cdot,\cdot)$ and a small inf-sup condition on $b(\cdot,\cdot)$.

For the classical saddle-point problem, i.e., $c(\cdot,\cdot) \equiv 0$, Theorem 2.5 which we formulate under the conditions that $a(\cdot,\cdot)$ is symmetric positive semidefinite and

$$\operatorname{Ker}(B^T) := \{q \in Q : b(v,q) = 0 \ \forall v \in V\} = \emptyset \tag{2.20}$$

has been proven in [16], see also [12, 17].

Theorem 2.5 (Brezzi [16])**.** *Assume that the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are continuous on $V \times V$ and on $V \times Q$, respectively, $a(\cdot,\cdot)$ is symmetric positive semidefinite, and also that*

$$a(v,v) \geq \underline{C}_a \|v\|_V^2 \qquad \forall v \in \operatorname{Ker}(B), \tag{2.21}$$

$$\inf_{q\in Q} \sup_{v\in V} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \beta > 0 \tag{2.22}$$

hold. Then the classical saddle-point problem (problem (2.11) *with $c(\cdot,\cdot) \equiv 0$) is well-posed.*

Remark 2.6. Note that if $\operatorname{Ker}(B^T) \neq \emptyset$, the statement of Theorem 2.5 remains valid if we identify any two elements q_1, q_2 for which $q_0 := q_1 - q_2$ is an element of $\operatorname{Ker}(B^T)$, i.e., replacing the space Q with the quotient space $Q/\operatorname{Ker}(B^T)$ and also the norm $\|\cdot\|_Q$ with $\|\cdot\|_{Q/\operatorname{Ker}(B^T)}$, the latter being defined by

$$\|q\|_{Q/\operatorname{Ker}(B^T)} = \inf_{q_0 \in \operatorname{Ker}(B^T)} \|q + q_0\|_Q.$$

In this case, the solution p is only unique up to an arbitrary element $p_0 \in \operatorname{Ker}(B^T)$.

For the classical saddle-point problem Brezzi's stability condition (2.22) and the continuity of $a(\cdot,\cdot)$ imply Babuška's stability condition (2.16), see [21], where it has also been shown that from (2.16) it follows (2.22) and the inf-sup condition for $a(\cdot,\cdot)$ in the kernel of B, the latter being equivalent to the coercivity estimate (2.21) if $a(\cdot,\cdot)$ is symmetric positive semidefinite.

Obviously, the stability condition (2.16) directly applies to perturbed saddle-point problems, a reason why they can be studied using Babuška's theory. However, conditions (2.21) and (2.22) together with the continuity of $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ in general are not sufficient to guarantee the stability condition (2.16) when $c(\cdot,\cdot) \not\equiv 0$. Additional conditions to ensure (2.16) have been studied, for example, in [12,14,17].

In [17] it has been shown that a condition on the kernel of B^T can be used as an additional assumption to ensure well-posedness of the perturbed saddle-point problem, that is, in particular, for Babuška's inf-sup condition (2.16) to hold. This condition is expressed in terms of the following auxiliary problem

$$\epsilon(p_0,q)_Q + c(p_0,q) = -c(p^\perp,q), \qquad \forall q \in \operatorname{Ker}(B^T), \tag{2.23}$$

and requires the following general assumption:

Assumption 2.7. There exists a $\gamma_0 > 0$ such that for every $p^\perp \in (\operatorname{Ker}(B^T))^\perp$ and for every $\epsilon > 0$ it holds that the norm of the solution $p_0 \in \operatorname{Ker}(B^T)$ of (2.23) is bounded by $\|p_0\|_Q \leq \frac{1}{\gamma_0}\|p^\perp\|_Q$.

The theorem then reads as follows:

Theorem 2.8 (Brezzi and Fortin [17]). *Assume that $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are continuous bilinear forms on $V\times V$, on $V\times Q$, and on $Q\times Q$, respectively. Further assume that $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are symmetric positive semidefinite. Finally, let* (2.21), (2.22) *from Theorem* 2.5 *and Assumption* 2.7 *be satisfied. Then for every $f\in V'$ and every $g\in \mathrm{Im}(B)$ problem* (2.11) *with $\mathcal{A}$ as defined in* (2.2) *and $\mathcal{F}(y)=\langle f,v\rangle_{V'\times V}+\langle g,q\rangle_{Q'\times Q}$ has a unique solution $x=(u;p)$ in $Y=V\times Q/M$ where*

$$M=\mathrm{Ker}(B^T)\cap\mathrm{Ker}(C).$$

Moreover, the estimate

$$\|u\|_V+\|p\|_{Q/\mathrm{Ker}(B^T)}\le K(\|f\|_{V'}+\|g\|_{Q'})$$

holds with a constant K only depending on $\bar{C}_a$, $\bar{C}_c$, $\underline{C}_a$, β and γ_0.

Remark 2.9. The result in [17] is more general than Theorem 2.8 in that it applies also to non-symmetric but positive semidefinite $a(\cdot,\cdot)$. We are considering only the case of symmetric positive semidefinite $a(\cdot,\cdot)$ in this paper.

In order to ensure the boundedness (continuity) of the symmetric positive semidefinite bilinear form $c(\cdot,\cdot)$ with respect to the norm $\|\cdot\|_Q$, and more generally the boundedness of $\mathcal{A}(\cdot,\cdot)$ with respect to the combined norm $\|\cdot\|_Y=(\|\cdot\|_V^2+\|\cdot\|_Q^2)^{1/2}$, it is natural to include the contribution of $c(\cdot,\cdot)$ in the norm $\|\cdot\|_Q$, e.g., by defining $\|\cdot\|_Q$ via

$$\|q\|_Q^2=|q|_Q^2+t^2c(q,q),\qquad \forall q\in Q, \tag{2.24}$$

for a proper seminorm or norm $|\cdot|_Q$ and a parameter $t\in[0,1]$.

As it has been shown in [14] the stability of the perturbed saddle-point problem then can be proven under Brezzi's conditions for the classical saddle-point problem and the additional condition

$$\inf_{u\in V}\sup_{(v;q)\in V\times Q}\frac{a(u,v)+b(u,q)}{|||(v;q)|||}\ge\gamma>0, \tag{2.25}$$

where $|||\cdot|||$ is defined by

$$|||(v;q)|||^2:=\|v\|_V^2+|q|_Q^2+t^2c(q,q),\qquad t\in[0,1], \tag{2.26}$$

and provides a specific choice of $\|\cdot\|_Y$, i.e., $\|\cdot\|_Y=|||\cdot|||$. The corresponding theorem then reads as:

Theorem 2.10 (Braess [14]). *Assume that the classical saddle-point problem with $a(\cdot,\cdot)$ being SPSD and $c(\cdot,\cdot)\equiv 0$ is stable, i.e., Brezzi's conditions* (2.21) *and* (2.22) *are fulfilled. If in addition condition* (2.25) *holds with $\gamma>0$ and we choose $t>0$ in* (2.26) *for $c(\cdot,\cdot)\not\equiv 0$, then the perturbed saddle-point problem* (2.11) *is stable under the norm $\|\cdot\|_Y:=|||\cdot|||$ and the stability constant $\underline{\alpha}$ in* (2.16) *depends only on β, $\underline{C}_a$ and γ and the choice of t.*

Remark 2.11. Note that as it can easily be seen if $a(\cdot,\cdot)$ is symmetric positive semidefinite, condition (2.25) is equivalent to the condition that there exists a constant $\gamma'>0$ such that

$$\frac{a(u,u)}{\|u\|_V}+\sup_{q\in Q}\frac{b(u,q)}{|q|_Q+tc(q,q)}\ge\gamma'\|u\|_V\qquad\forall u\in V. \tag{2.27}$$

Moreover, as shown in [14], then (2.27) is also equivalent to the condition that there exists a constant $\gamma'' > 0$ such that

$$\sup_{(v;q)\in Y} \frac{\mathcal{A}((u;0),(v;q))}{|||(v;q)|||} \geq \gamma''\|u\|_V \qquad \forall u \in V. \tag{2.28}$$

Since (2.25) is an inf-sup condition for $(a(\cdot,\cdot)+b(\cdot,\cdot))$ which can be interpreted as a big inf-sup condition on $\mathcal{A}$ for $p = 0$ under the specific norm $\|\cdot\|_Y = |||\cdot|||$, see (2.28), Theorem 2.10 still does not provide us with the desired stability result in terms of conditions on $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ separately. On the other hand, Theorem 2.8 requires the solution of the auxiliary problem (2.23) on $\mathrm{Ker}(B^T)$ for which one has to verify Assumption 2.7 which, in some situations, is a difficult task.

Our aim is to avoid the latter and still impose Brezzi-type conditions, in particular a small inf-sup condition on $b(\cdot,\cdot)$. In the next section, we will prove a theorem (Theorem 2.15) which ensures the stability of the perturbed saddle-point problem 2.11 under conditions which are equivalent to the conditions in Brezzi's theorem (Theorem 2.5) when the perturbation term vanishes. Moreover, our approach provides a framework suited for finding norms in which stability can be shown and allows for simplifying and shortening proofs based on the result of Babuška.

2.3. **A new framework for the stability analysis of perturbed saddle-point problems.** The key idea for studying and verifying the stability of perturbed saddle-point problems we follow in this paper is to construct proper norms as part of an abstract framework which applies to a variational formulation of various multiphysics models. As we have already observed in the previous subsection, a norm-splitting of the form (2.24) is quite natural if the symmetric positive semi-definite perturbation form $c(\cdot,\cdot)$ is not identical to zero. For fixed $t > 0$, the norm defined in (2.24) is equivalent to the norm defined by

$$\|q\|_Q^2 := |q|_Q^2 + c(q,q) =: \langle \bar{Q}q, q\rangle_{Q'\times Q}. \tag{2.29}$$

Note that the assumption that $\|\cdot\|_Q$ is a full norm induced by an inner product under which Q is a Hilbert space implies that the seminorm $|\cdot|_Q$ corresponds to an SPSD bilinear form $d(\cdot,\cdot) : Q \times Q \to \mathbb{R}$, i.e., $|q|_Q^2 = d(q,q)$. Consequently, the form $c(p,q) + d(p,q)$ is symmetric positive definite (SPD) and defines a linear operator $\bar{Q} : Q \to Q'$ by $\langle \bar{Q}p, q\rangle := c(p,q) + d(p,q)$.

Now we introduce the following splitting of the norm $\|\cdot\|_V$ defined by

$$\|v\|_V^2 := |v|_V^2 + |v|_b^2, \tag{2.30}$$

where $|\cdot|_V$ is a proper seminorm, which is a norm on $\mathrm{Ker}(B)$ satisfying

$$|v|_V^2 \eqsim a(v,v) \qquad \forall v \in \mathrm{Ker}(B)$$

and $|\cdot|_b$ is defined by

$$|v|_b^2 := \langle Bv, \bar{Q}^{-1}Bv\rangle_{Q'\times Q} = \|Bv\|_{Q'}^2. \tag{2.31}$$

Here, $\bar{Q}^{-1} : Q' \to Q$ is an isometric isomorphism (Riesz isomorphism) since $\bar{Q}$ is an isometric isomorphism, i.e.,

$$\begin{aligned}\|Bv\|_{Q'}^2 &= \|\bar{Q}^{-1}Bv\|_Q^2 = (\bar{Q}^{-1}Bv, \bar{Q}^{-1}Bv)_Q = \langle \bar{Q}\bar{Q}^{-1}Bv, \bar{Q}^{-1}Bv\rangle_{Q'\times Q}\\ &= \langle Bv, \bar{Q}^{-1}Bv\rangle_{Q'\times Q}.\end{aligned}$$

Remark 2.12. Note that both $|\cdot|_V$ and $|\cdot|_b$ can be seminorms as long as they add up to a full norm. Likewise, only the sum of the seminorms $|\cdot|_Q$ and $c(\cdot,\cdot)$ has to define a norm. In some particular situations, it is also useful to identify certain of the involved seminorms with 0, in which case the corresponding splitting becomes a trivial splitting. The splitting (2.30) is closely related to a Schur complement type operator, corresponding to the modified (regularized) bilinear form resulting from $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ by replacing $c(\cdot,\cdot)$ with $(\cdot,\cdot)_Q$.

In order to present our main theoretical result, we give Definition 2.13.

Definition 2.13. For two Hilbert spaces V and Q, a norm $\|\cdot\|_V$ on V and a norm $\|\cdot\|_Q$ on Q are called fitted if they satisfy the splittings (2.29) and (2.30), respectively, where $|\cdot|_Q$ is a seminorm on Q and $|\cdot|_V$ and $|\cdot|_b$ are seminorms on V, the latter defined according to (2.31).

Remark 2.14. Note that the norm fitting can also be performed by first fixing the full norm on V (instead of the full norm on Q as described above). Exploiting the structure of the problem, in the latter case one uses the following norm splittings

$$\|v\|_V := |v|_V^2 + a(v,v) =: \langle \bar{V}v, v\rangle_{V'\times V},$$
$$\|q\|_Q := |q|_Q^2 + |q|_b^2,$$

where $\bar{V} : V \to V'$ is a linear operator, $|q|_Q^2$ is equivalent to $c(q,q)$ and $|q|_b^2 =: \langle B^T q, \bar{V}^{-1}B^T q\rangle_{V'\times V}$.

Theorem 2.15. *Let $\|\cdot\|_V$ and $\|\cdot\|_Q$ be fitted norms according to Definition* 2.13*, which immediately implies the continuity of $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ in these norms with $\bar{C}_b = 1$ and $\bar{C}_c = 1$, cf.* (2.8)–(2.9)*. Consider the bilinear form $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ defined in* (2.2) *where $a(\cdot,\cdot)$ is continuous, i.e.,* (2.7) *holds, and $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are symmetric positive semidefinite. Assume, further, that $a(\cdot,\cdot)$ satisfies the coercivity estimate*

$$a(v,v) \geq \underline{C}_a |v|_V^2, \qquad \forall v \in V, \tag{2.32}$$

and that there exists a constant $\underline{\beta} > 0$ such that

$$\sup_{\substack{v\in V\\ v\neq 0}} \frac{b(v,q)}{\|v\|_V} \geq \underline{\beta}|q|_Q, \qquad \forall q \in Q. \tag{2.33}$$

Then the bilinear form $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ is continuous and inf-sup stable under the combined norm $\|\cdot\|_Y$ defined in (2.1)*, i.e., the conditions* (2.15) *and* (2.16) *hold.*

Before presenting the proof of Theorem 2.15, we show an auxiliary result and make some remarks.

Lemma 2.16. *The inf-sup condition: there exists a constant $\underline{\beta} > 0$ such that*

$$\sup_{\substack{v\in V\\ v\neq 0}} \frac{b(v,q)}{\|v\|_V} \geq \underline{\beta}|q|_Q, \qquad \forall q \in Q, \tag{2.34}$$

is equivalent to the condition: for any $q \in Q$, there exists $v \in V$, such that

$$b(v,q) = |q|_Q^2 \quad \text{and} \quad \|v\|_V \leq \underline{\beta}^{-1}|q|_Q. \tag{2.35}$$

Proof. Obviously, (2.35) implies (2.34). Hence, it remains to prove that (2.34) implies (2.35). Let $K_Q = \{q \in Q : |q|_Q = 0\}$ and define $\|q\|_{Q/K_Q} := |q|_Q$ which is a norm on the quotient space Q/K_Q. For any $q \in Q$, where the class in Q/K_Q which q belongs to is also denoted by q, there exists $f \in (Q/K_Q)'$ s.t. $f(q) = \|q\|^2_{Q/K_Q}$ and $\|f\|_{(Q/K_Q)'} = \|q\|_{Q/K_Q}$. Since B is onto, we can find v s.t. $Bv = f$ and by the open mapping theorem, we can find v with $\|v\|_V \le \underline{\beta}^{-1}\|f\|_{(Q/K_Q)'} = \underline{\beta}^{-1}\|q\|_{Q/K_Q} = \underline{\beta}^{-1}|q|_Q$ and $b(v,q) = \langle Bv, q\rangle = f(q) = \|q\|^2_{Q/K_Q} = |q|^2_Q$. □

Remark 2.17. The continuity of $b(\cdot,\cdot)$ readily follows from

$$b(v,q) = \langle Bv, q\rangle_{Q'\times Q} = \langle \bar{Q}\bar{Q}^{-1}Bv, q\rangle_{Q'\times Q} = (\bar{Q}^{-1}Bv, q)_Q \le \|\bar{Q}^{-1}Bv\|_Q\|q\|_Q \le \|v\|_V\|q\|_Q.$$

If $|\cdot|_V$ is induced by the bilinear form $a(\cdot,\cdot)$ then the continuity of $a(\cdot,\cdot)$ also follows directly from the definition of the fitted norms.

Remark 2.18. Theorem 2.15 is a generalization of Theorem 2.5 in the sense that given two norms $\|\cdot\|_{Q,\mathrm{eqv}}$ and $\|\cdot\|_{V,\mathrm{eqv}}$ under which the conditions of Theorem 2.5 are satisfied, one can always find two fitted equivalent norms $\|\cdot\|_Q \eqsim \|\cdot\|_{Q,\mathrm{eqv}}$ and $\|\cdot\|_V \eqsim \|\cdot\|_{V,\mathrm{eqv}}$ such that the conditions of Theorem 2.15 are satisfied in these fitted norms when $c(\cdot,\cdot) \equiv 0$.

More specifically, for $c(\cdot,\cdot) \equiv 0$, we have $|\cdot|_Q = ||\cdot||_Q = \|q\|_{Q,\mathrm{eqv}}$ and $\bar{Q} = I$. If we define the fitted norm $\|\cdot\|_V$ by choosing

$$|v|^2_V = a(v,v), \tag{2.36}$$

then (2.32) obviously holds. In addition, there exists a constant α_0 such that (see [12, Proposition 4.3.4])

$$\alpha_0\|v\|^2_{V,\mathrm{eqv}} \le a(v,v) + \|Bv\|^2_{Q'} = \|v\|^2_V. \tag{2.37}$$

At the same time, under the conditions of Theorem 2.5, the continuity of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ in the norms $\|\cdot\|_{V,\mathrm{eqv}}$ and $\|\cdot\|_{Q,\mathrm{eqv}}$, we have

$$\|v\|^2_V = a(v,v) + \|Bv\|^2_{Q'} \le C\|v\|^2_{V,\mathrm{eqv}}. \tag{2.38}$$

Thus, the fitted norm $\|\cdot\|_V$ is equivalent to the norm $\|\cdot\|_{V,\mathrm{eqv}}$, and (2.33) is induced by (2.22).

Remark 2.19. Note that under the conditions of Theorem 2.15 the coercivity of $a(\cdot,\cdot)$ on $\mathrm{Ker}(B)$ in the (semi-) norms $|\cdot|_V$ and $\|\cdot\|_V$ are equivalent since $|v|_V = \|v\|_V$ for all $v \in \mathrm{Ker}(B)$. The inf-sup condition (2.33), however, uses the seminorm $|\cdot|_Q$ instead of $\|\cdot\|_Q$ as in Brezzi's condition (2.22).

Proof of Theorem 2.15. Demonstrating (2.15) is straightforward since

$$\begin{aligned}\mathcal{A}((w;r),(v;q)) =& a(w,v) + b(v,r) + b(w,q) - c(r,q)\\ \le& \bar{C}_a\|w\|_V\|v\|_V + \|v\|_V\|r\|_Q + \|w\|_V\|q\|_Q + \|r\|_Q\|q\|_Q\\ \le& \bar{C}(\|w\|_V + \|r\|_Q)(\|v\|_V + \|q\|_Q) \le 2\bar{C}\|(w;r)\|_Y\|(v;q)\|_Y\end{aligned} \tag{2.39}$$

with $\bar{C} := \max\{\bar{C}_a, 1\}$.

In order to prove (2.16) for a positive constant δ, which will be selected later, and for a given arbitrary pair $(w,r) \in V \times Q$, we choose

$$v := \delta w + w_0, \tag{2.40}$$

where, by Lemma 2.16, $w_0 \in V$ can be chosen such that

$$b(w_0, r) = |r|_Q^2, \tag{2.41a}$$

$$\|w_0\|_V \leq \underline{\beta}^{-1} |r|_Q, \tag{2.41b}$$

and

$$q := -\delta r + r_0, \tag{2.42}$$

where

$$r_0 := \bar{Q}^{-1} B w. \tag{2.43}$$

Note that the existence of an element w_0 satisfying (2.41) follows from (2.33).

Then we have

$$\begin{aligned} \|v\|_V &\leq \|\delta w\|_V + \|w_0\|_V \leq \delta \|w\|_V + \underline{\beta}^{-1} |r|_Q \leq \delta \|w\|_V + \underline{\beta}^{-1} \|r\|_Q, \\ \|q\|_Q &\leq \delta \|r\|_Q + \|r_0\|_Q = \delta \|r\|_Q + (\bar{Q}^{-1} B w, \bar{Q}^{-1} B w)_Q^{1/2} = \delta \|r\|_Q + |w|_b, \end{aligned} \tag{2.44}$$

and, consequently,

$$\|(v; q)\|_Y^2 = \|v\|_V^2 + \|q\|_Q^2 \leq 2(\delta^2 + 1) \|w\|_V^2 + 2(\underline{\beta}^{-2} + \delta^2) \|r\|_Q^2. \tag{2.45}$$

Hence, it follows that

$$\|(v; q)\|_Y \leq \left(2 \max\{(\delta^2 + 1), (\underline{\beta}^{-2} + \delta^2)\}\right)^{\frac{1}{2}} \|(w; r)\|_Y. \tag{2.46}$$

Moreover, for the same choice of v and q, we obtain

$$\begin{aligned} &\mathcal{A}((w; r), (v; q)) \\ &= a(w, \delta w + w_0) + b(\delta w + w_0, r) - b(w, \delta r - r_0) + c(r, \delta r - r_0) \\ &\geq \delta a(w, w) + a(w, w_0) + \delta b(w, r) + b(w_0, r) - \delta b(w, r) + \delta c(r, r) \\ &+ \langle B w, \bar{Q}^{-1} B w \rangle_{Q' \times Q} - c(r, \bar{Q}^{-1} B w) \\ &\geq \delta a(w, w) - \frac{1}{2} \epsilon^{-1} a(w, w) - \frac{1}{2} \epsilon a(w_0, w_0) + |r|_Q^2 + \delta c(r, r) \\ &\quad + |w|_b^2 - \frac{1}{2} c(r, r) - \frac{1}{2} c(\bar{Q}^{-1} B w, \bar{Q}^{-1} B w) \\ &\geq \left(\delta - \frac{1}{2} \epsilon^{-1}\right) a(w, w) - \frac{1}{2} \epsilon \bar{C}_a \underline{\beta}^{-2} |r|_Q^2 + |r|_Q^2 + \delta c(r, r) + |w|_b^2 - \frac{1}{2} c(r, r) - \frac{1}{2} |w|_b^2 \\ &\geq \left(\delta - \frac{1}{2} \epsilon^{-1}\right) \underline{C}_a |w|_V^2 + \left(1 - \frac{1}{2} \epsilon \bar{C}_a \underline{\beta}^{-2}\right) |r|_Q^2 + \delta c(r, r) - \frac{1}{2} c(r, r) + \frac{1}{2} |w|_b^2 \end{aligned}$$

and, hence, for $\epsilon = \frac{1}{2} \bar{C}_a^{-1} \underline{\beta}^2$ and $\delta = \max\{\frac{1}{4} \underline{C}_a^{-1} + \bar{C}_a \underline{\beta}^{-2}, \frac{3}{4}\}$, we have

$$\begin{aligned} \mathcal{A}((w; r), (v; q)) &\geq \left(\delta - \bar{C}_a \underline{\beta}^{-2}\right) \underline{C}_a |w|_V^2 + \frac{1}{4} |r|_Q^2 + \left(\delta - \frac{1}{2}\right) c(r, r) + \frac{1}{2} |w|_b^2 \\ &\geq \frac{1}{4} \left(\|w\|_V^2 + \|r\|_Q^2\right) = \frac{1}{4} \|(w; r)\|_Y^2. \end{aligned} \tag{2.47}$$

Together, (2.46) and (2.47) imply the inf-sup condition (2.16) which can equivalently be formulated as

$$\sup_{(v; q) \in Y} \frac{\mathcal{A}((w; r), (v; q))}{\|(v; q)\|_Y} \geq \underline{\alpha} \|(w; r)\|_Y \qquad \forall (w; r) \in Y \tag{2.48}$$

because the supremum on the left-hand side of (2.48) is bounded from below by

$$\frac{\mathcal{A}((w;r),(v;q))}{\|(v;q)\|_Y}$$

if we insert any fixed $(v;q)$, in particular the choice we made and for which we proved

$$\frac{\mathcal{A}((w;r),(v;q))}{\|(v;q)\|_Y} \geq \frac{\frac{1}{4}\|(w;r)\|_Y^2}{(2\max\{(\delta^2+1),(\underline{\beta}^{-2}+\delta^2)\})^{1/2}\|(w;r)\|_Y}.$$

□

Remark 2.20. The statement of Theorem 2.15 remains valid if the norms $\|\cdot\|_Q$ and $\|\cdot\|_V$, as defined in (2.29) and (2.30), are replaced with equivalent norms $\|\cdot\|_{Q,\mathrm{eqv}} \eqsim \|\cdot\|_Q$ and $\|\cdot\|_{V,\mathrm{eqv}} \eqsim \|\cdot\|_V$, hence using $\|\cdot\|_{V,\mathrm{eqv}}$ in (2.33) in this case. The proof remains unchanged and the only difference in the final result is that the inf-sup constant $\underline{\alpha}$ in (2.48) with respect to the (new) equivalent combined norm has to be scaled by the quotient of the constants in the norm equivalence relation for the combined norms. For that reason, without loss of generality, we can use the fitted norms defined by (2.29) and (2.30) directly in the formulation of Theorem 2.15.

Remark 2.21. An advantage of the new framework is that Theorem 2.15 provides sufficient conditions that are easy to verify in practice. These are given in form of an LBB-type condition on the basis of the proposed norm fitting technique, which allows, contrary to the technique in [63], to choose the norms subsequently, see also Remark 2.14.

3. Applications of the framework

In this section four different classes of problems are analyzed by means of the proposed framework demonstrating its versatility and ease of application. Here, we use bold letters to denote vector-valued functions and the spaces to which they belong which means that we identify certain non-bold symbols from the abstract framework in the previous section with bold symbols, e.g., $v = \boldsymbol{v}$. To prove stability of the exemplified mixed variational formulations, we assume that proper boundary conditions are imposed. In certain cases, we will also make use of the following classical inf-sup conditions, see [16], also [12, 17], for the pairs of spaces $(\boldsymbol{V}, Q)$: there exist constants β_d and β_s such that

$$\inf_{q\in Q}\sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{(\mathrm{div}\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\mathrm{div}}\|q\|} \geq \beta_d > 0, \tag{3.1}$$

$$\inf_{q\in Q}\sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{(\mathrm{div}\boldsymbol{v},q)}{\|\boldsymbol{v}\|_1\|q\|} \geq \beta_s > 0, \tag{3.2}$$

where the norms $\|\cdot\|_{\mathrm{div}}$, $\|\cdot\|_1$ and $\|\cdot\|$ denote the standard $\boldsymbol{H}(\mathrm{div})$, $\boldsymbol{H}^1$ and L^2 norms and $(\cdot,\cdot)$ is the L^2-inner product.

3.1. **Generalized Poisson and generalized Stokes equations.**

Example 3.1. The first example, see [13], is the following mixed variational problem resulting from a weak formulation of a generalized Poisson equation: find $(\boldsymbol{u}, p) \in \boldsymbol{H}(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} (\boldsymbol{u}, \boldsymbol{v}) + (p, \mathrm{div}\boldsymbol{v}) &= 0, \qquad \forall \boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}, \Omega), \\ (\mathrm{div}\boldsymbol{u}, q) - t(p, q) &= -(f, q), \qquad \forall q \in L^2(\Omega), \end{aligned} \tag{3.3}$$

where $t \geq 0$ is a parameter.

The bilinear forms generating $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ are given by

$$a(\boldsymbol{u}, \boldsymbol{v}) := (\boldsymbol{u}, \boldsymbol{v}), \quad b(\boldsymbol{v}, q) := (\mathrm{div}\boldsymbol{v}, q), \quad c(p, q) = t(p, q), \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}, \forall p, q \in Q,$$

where $Q := L^2(\Omega)$, $\boldsymbol{V} := \boldsymbol{H}(\mathrm{div}, \Omega)$. Using the norm fitting technique, we define $|\cdot|_Q$, $|\cdot|_V$ by

$$|q|_Q^2 := (q, q) \quad \forall q \in Q \quad \text{and} \quad |\boldsymbol{v}|_V^2 := (\boldsymbol{v}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{V}.$$

Obviously, $B := \mathrm{div} : \boldsymbol{V} \to Q'$, and (2.29) and (2.30) take the form

$$\|q\|_Q^2 = |q|_Q^2 + c(q, q) = (q, q) + t(q, q) = ((1 + t)q, q) = \langle (1 + t)Iq, q\rangle_{Q' \times Q},$$

$$\|\boldsymbol{v}\|_V^2 = |\boldsymbol{v}|_V^2 + |\boldsymbol{v}|_b^2 = (\boldsymbol{v}, \boldsymbol{v}) + \langle \mathrm{div}\boldsymbol{v}, \frac{1}{(1+t)} I \mathrm{div}\boldsymbol{v}\rangle_{Q' \times Q} = (\boldsymbol{v}, \boldsymbol{v}) + \frac{1}{(1+t)}(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}),$$

respectively. Since $a(\boldsymbol{v}, \boldsymbol{v}) = (\boldsymbol{v}, \boldsymbol{v}) = |\boldsymbol{v}|_V^2$ for all $\boldsymbol{v} \in \boldsymbol{V}$, condition (2.32) in Theorem 2.15 is satisfied with $\underline{C}_a = 1$.

Finally, we have to verify condition (2.33) in Theorem 2.15, namely

$$\sup_{\boldsymbol{v} \in V} \frac{(\mathrm{div}\boldsymbol{v}, q)}{(\|\boldsymbol{v}\|^2 + \frac{1}{(1+t)}\|\mathrm{div}\boldsymbol{v}\|^2)^{1/2}} \geq \underline{\beta}|q|_Q =: \underline{\beta}\|q\|, \qquad \forall q \in Q,$$

which follows directly from the classical inf-sup condition (3.1) on the spaces $(\boldsymbol{V}, Q)$ since $t \geq 0$.

As a result, we obtain that the preconditioner

$$\mathcal{B} := \begin{bmatrix} (I - (1 + t)^{-1}\nabla\mathrm{div})^{-1} & \\ & ((1 + t)I)^{-1} \end{bmatrix}$$

is norm-equivalent for the combined norm, cf. [47]. To solve the $\boldsymbol{H}(\mathrm{div})$ subproblem, one can use various preconditioners, see, e.g., [30, 38], multigrid, see, e.g., [2, 61], and domain decomposition methods [59].

Remark 3.2. Note that in Example 3.1 as well as the ones which follow, the perturbation term $c(\cdot,\cdot)$, due to the presence of various parameters, can dominate the problem. In this situation stability often cannot be proven using Theorem 2.8 because $\|q\|_Q$ has to bound $c(q, q)^{\frac{1}{2}}$ which for dominating perturbation conflicts satisfying the classical LBB condition (2.22) uniformly. A way to overcome this problem is to work either with the Braess inf-sup condition, see Theorem 2.10, or with the Babuška inf-sup condition, see Theorem 2.1, or, alternatively use the simpler to verify inf-sup condition provided in Theorem 2.15.

Example 3.3. The second example we consider is taken from [43]. Its mixed variational formulation reads: find $(\boldsymbol{u}, p) \in \boldsymbol{H}_0^1(\Omega) \times (H^1(\Omega) \cap L_0^2(\Omega))$ such that

$$\begin{aligned} (\nabla\boldsymbol{u}, \nabla\boldsymbol{v}) - (p, \mathrm{div}\boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega), \\ -(\mathrm{div}\boldsymbol{u}, q) - (\kappa\nabla p, \nabla q) &= (g, q), \qquad \forall q \in H^1(\Omega) \cap L_0^2(\Omega), \end{aligned} \tag{3.4}$$

where $\kappa \geq 0$ is a parameter.

The bilinear forms defining $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ here are given by

$$a(\boldsymbol{u},\boldsymbol{v}) := (\nabla\boldsymbol{u},\nabla\boldsymbol{v}), \quad b(\boldsymbol{v},q) := -(\mathrm{div}\boldsymbol{v},q),$$
$$c(p,q) = (\kappa\nabla p,\nabla q), \quad \forall \boldsymbol{u},\boldsymbol{v}\in\boldsymbol{V},\forall p,q\in Q.$$

In Example 3.3, we set $Q := H^1(\Omega)\cap L^2_0(\Omega)$, $\boldsymbol{V} := \boldsymbol{H}^1_0(\Omega)$, and $|\cdot|_Q$, $|\cdot|_V$ to be

$$|q|_Q^2 := (q,q) \quad \forall q\in Q \quad \text{and} \quad |\boldsymbol{v}|_V^2 := (\nabla\boldsymbol{v},\nabla\boldsymbol{v}) \qquad \forall \boldsymbol{v}\in\boldsymbol{V}.$$

Then the operator B is defined by $B:\boldsymbol{V}\to Q'$, $B := -\mathrm{div}$ and the norm splittings (2.29) and (2.30) are given by

$$\|q\|_Q^2 = |q|_Q^2 + c(q,q) = (q,q) + (\kappa\nabla q,\nabla q) = \langle\bar{Q}q,q\rangle_{Q'\times Q} \tag{3.5}$$

and

$$\|\boldsymbol{v}\|_V^2 = |\boldsymbol{v}|_V^2 + |\boldsymbol{v}|_b^2 = (\nabla\boldsymbol{v},\nabla\boldsymbol{v}) + \langle B\boldsymbol{v},\bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q'\times Q}.$$

Condition (2.32) is automatically satisfied with $\underline{C}_a = 1$. To show (2.33) we first note that using (3.5) we obtain

$$\begin{aligned}\langle B\boldsymbol{v},\bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q'\times Q} = \|B\boldsymbol{v}\|_{Q'}^2 &= \left(\sup_{q\neq 0}\frac{b(\boldsymbol{v},q)}{\|q\|_Q}\right)^2 = \left(\sup_{q\neq 0}\frac{(\mathrm{div}\boldsymbol{v},q)}{\|q\|_Q}\right)^2\\ &\le \left(\sup_{q\neq 0}\frac{\|\mathrm{div}\boldsymbol{v}\|\|q\|}{\|q\|_Q}\right)^2\\ &\le (\mathrm{div}\boldsymbol{v},\mathrm{div}\boldsymbol{v}).\end{aligned}$$

Thus,

$$\begin{aligned}\|\boldsymbol{v}\|_V^2 &= (\nabla\boldsymbol{v},\nabla\boldsymbol{v}) + \langle B\boldsymbol{v},\bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q'\times Q} \le (\nabla\boldsymbol{v},\nabla\boldsymbol{v}) + (\mathrm{div}\boldsymbol{v},\mathrm{div}\boldsymbol{v}) \le 2(\nabla\boldsymbol{v},\nabla\boldsymbol{v})\\ &\le 2\|\boldsymbol{v}\|_1^2.\end{aligned} \tag{3.6}$$

Now, we choose $\boldsymbol{v}_0$ such that $-\mathrm{div}\boldsymbol{v}_0 = q$ and hereby obtain from the Stokes inf-sup condition (3.2) the estimate $\|\boldsymbol{v}_0\|_1 \le \frac{1}{\beta_s}\|q\|$, and, finally,

$$\sup_{\boldsymbol{v}\in V}\frac{b(\boldsymbol{v},q)}{\|\boldsymbol{v}\|_V} \ge \frac{b(\boldsymbol{v}_0,q)}{\|\boldsymbol{v}_0\|_V} = \frac{\|q\|^2}{\|\boldsymbol{v}_0\|_V} \ge \frac{1}{\sqrt{2}}\frac{\|q\|^2}{\|\boldsymbol{v}_0\|_1} \ge \frac{\beta_s}{\sqrt{2}}\frac{\|q\|^2}{\|q\|} =: \underline{\beta}\|q\| = \underline{\beta}|q|_Q, \quad \forall q\in Q.$$

The induced norm-equivalent preconditioner in Example 3.3 reads as

$$\mathcal{B} := \begin{bmatrix}(-\Delta - \nabla(I-\mathrm{div}\kappa\nabla)^{-1}\mathrm{div})^{-1} & \\ & (I-\mathrm{div}\kappa\nabla)^{-1}\end{bmatrix} \eqsim \begin{bmatrix}-\Delta^{-1} & \\ & (I-\mathrm{div}\kappa\nabla)^{-1}\end{bmatrix},$$

where the equivalence is due to (3.6).

3.2. **Stokes Darcy problem.**

Example 3.4. Let $\Omega = \Omega_S\cup\Omega_D$ and $\Gamma = \partial\Omega_S\cap\partial\Omega_D$. We assume that $\Gamma_S^D\cup\Gamma_S^N\cup\Gamma$ forms a disjoint decomposition of $\partial\Omega_S$ and, similarly, $\Gamma_D^D\cup\Gamma_D^N\cup\Gamma$ is a partition of $\partial\Omega_D$. Denote

$$H^1_{\Gamma_i^D}(\Omega_i) = \left\{w\in H^1(\Omega_i) : w|_{\Gamma_i^D} = 0\right\},\ i = S \text{ or } D.$$

The classical formulation of Stokes Darcy problem follows [22]: Find $(\boldsymbol{u}, p_S, p_D) \in \boldsymbol{H}^1_{\Gamma_S^D}(\Omega_S) \times L^2(\Omega_S) \times H^1_{\Gamma_D^D}(\Omega_D)$ such that

$$\begin{aligned}
&(2\mu\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{v}))_{\Omega_S} + \beta_\tau (\boldsymbol{\tau}\cdot\boldsymbol{u}, \boldsymbol{\tau}\cdot\boldsymbol{v})_\Gamma \\
&\qquad - (p_S, \nabla\cdot\boldsymbol{v})_{\Omega_S} + (p_D, \boldsymbol{n}\cdot\boldsymbol{v})_\Gamma = (\boldsymbol{f}_S, \boldsymbol{v})_{\Omega_S}, \qquad \forall \boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_S^D}(\Omega_S), \\
&\qquad\qquad - (\nabla\cdot\boldsymbol{u}, q_S)_{\Omega_S} = 0, \qquad\qquad \forall q_S \in L^2(\Omega_S), \\
&\qquad (\boldsymbol{n}\cdot\boldsymbol{u}, q_D)_\Gamma - (\kappa\nabla p_D, \nabla q_D)_{\Omega_D} = (f_D, q_D)_{\Omega_D}, \quad \forall q_D \in H^1_{\Gamma_D^D}(\Omega_D),
\end{aligned}$$

where $\boldsymbol{n} := \boldsymbol{n}_S$ is the outer normal of the Stokes domain and $\boldsymbol{\tau} := \boldsymbol{I} - (\boldsymbol{n}\otimes\boldsymbol{n})$ is the projection onto the tangent bundle of the interface Γ.

The bilinear forms defining $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ here are given by

$$\begin{aligned}
a(\boldsymbol{u}, \boldsymbol{v}) &:= (2\mu\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{v}))_{\Omega_S} + \beta_\tau (\boldsymbol{\tau}\cdot\boldsymbol{u}, \boldsymbol{\tau}\cdot\boldsymbol{v})_\Gamma, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}, \\
b(\boldsymbol{v}, \boldsymbol{q}) &:= -(q_S, \nabla\cdot\boldsymbol{v})_{\Omega_S} + (q_D, \boldsymbol{n}\cdot\boldsymbol{v})_\Gamma, \quad \forall \boldsymbol{v} \in \boldsymbol{V}, \forall \boldsymbol{q} \in \boldsymbol{Q}, \\
c(\boldsymbol{p}, \boldsymbol{q}) &:= (\kappa\nabla p_D, \nabla q_D)_{\Omega_D}, \quad \forall \boldsymbol{p}, \boldsymbol{q} \in \boldsymbol{Q},
\end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{H}^1_{\Gamma_S^D}(\Omega_S), \boldsymbol{Q} = L^2(\Omega_S) \times H^1_{\Gamma_D^D}(\Omega_D)$ and $\boldsymbol{p} = (p_S, p_D), \boldsymbol{q} = (q_S, q_D)$. We fix $|\cdot|_Q$, $|\cdot|_V$ to be

$$\begin{aligned}
|\boldsymbol{q}|_Q^2 &:= (2\mu)^{-1}(q_S, q_S)_{\Omega_S} + (2\mu)^{-1}\|q_D\|^2_{-\frac{1}{2},\Gamma}, \quad \forall \boldsymbol{q} \in \boldsymbol{Q}, \\
|\boldsymbol{v}|_V^2 &:= (2\mu\boldsymbol{\epsilon}(\boldsymbol{v}), \boldsymbol{\epsilon}(\boldsymbol{v}))_{\Omega_S} + \beta_\tau (\boldsymbol{\tau}\cdot\boldsymbol{v}, \boldsymbol{\tau}\cdot\boldsymbol{v})_\Gamma, \quad \forall \boldsymbol{u} \in \boldsymbol{V}.
\end{aligned}$$

Noting that $a(\boldsymbol{v}, \boldsymbol{v}) = |\boldsymbol{v}|^2_{\boldsymbol{V}}$ for all $\boldsymbol{v} \in \boldsymbol{V}$, that is, (2.32) is satisfied with $\underline{C}_a = 1$. In addition, we have

$$\begin{aligned}
\|\boldsymbol{q}\|_Q^2 &:= |\boldsymbol{q}|_Q^2 + c(\boldsymbol{q}, \boldsymbol{q}) = \langle \bar{Q}\boldsymbol{q}, \boldsymbol{q}\rangle_{Q'\times Q}, \ \forall \boldsymbol{q} \in \boldsymbol{Q}, \\
\|\boldsymbol{v}\|^2_{\boldsymbol{V}} &:= |\boldsymbol{v}|^2_{\boldsymbol{V}} + \langle B\boldsymbol{v}, \bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q'\times Q} = |\boldsymbol{v}|^2_{\boldsymbol{V}} + |\boldsymbol{v}|^2_b, \ \forall \boldsymbol{v} \in \boldsymbol{V}.
\end{aligned}$$

The continuity of B is shown by the following calculation, utilizing the Cauchy-Schwarz inequality and a trace inequality:

$$\begin{aligned}
\langle B\boldsymbol{v}, \boldsymbol{q}\rangle_{\boldsymbol{Q}'\times\boldsymbol{Q}} &= -(\nabla\cdot\boldsymbol{v}, p_S)_{\Omega_S} + (\boldsymbol{n}\cdot\boldsymbol{v}, q_D)_\Gamma \\
&\le \|\nabla\cdot\boldsymbol{v}\|_{\Omega_S}\|q_S\|_{\Omega_S} + \|\boldsymbol{n}\cdot\boldsymbol{v}\|_{\frac{1}{2},\Gamma}\|q_D\|_{-\frac{1}{2},\Gamma} \\
&\lesssim \|\boldsymbol{\epsilon}(\boldsymbol{v})\|_{\Omega_S}\left(\|q_S\|_{\Omega_S} + \|q_D\|_{-\frac{1}{2},\Gamma}\right) \\
&\lesssim (2\mu)^{\frac{1}{2}}\|\boldsymbol{\epsilon}(\boldsymbol{v})\|_{\Omega_S}(2\mu)^{-\frac{1}{2}}\left(\|q_S\|^2_{\Omega_S} + \|q_D\|^2_{-\frac{1}{2},\Gamma}\right)^{\frac{1}{2}} \\
&\le |\boldsymbol{v}|_{\boldsymbol{V}}\,|q|_{\boldsymbol{Q}} \le |\boldsymbol{v}|_V\|q\|_{\boldsymbol{Q}}.
\end{aligned} \tag{3.7}$$

Furthermore, by (2.31), we have

$$|\boldsymbol{v}|_b = \|B\boldsymbol{v}\|_{\boldsymbol{Q}'} = \sup_{\boldsymbol{q}\in\boldsymbol{Q}} \frac{\langle B\boldsymbol{v}, \boldsymbol{q}\rangle_{\boldsymbol{Q}'\times\boldsymbol{Q}}}{\|q\|_{\boldsymbol{Q}}} \le |\boldsymbol{v}|_{\boldsymbol{V}}, \quad \forall \boldsymbol{v} \in \boldsymbol{V}, \tag{3.8}$$

and $\|\cdot\|_{\boldsymbol{V}}$ is equivalent to $|\cdot|_{\boldsymbol{V}}$, namely

$$\|\boldsymbol{v}\|_{\boldsymbol{V}} \eqsim |\boldsymbol{v}|_{\boldsymbol{V}}, \quad \forall \boldsymbol{v} \in \boldsymbol{V}. \tag{3.9}$$

Now we show that (2.33) is satisfied. By (3.9), it suffices to show the following inf-sup condition of B: there exists a constant $\underline{\beta} > 0$ such that

$$\sup_{\substack{\boldsymbol{v}\in\boldsymbol{V}\\ \boldsymbol{v}\neq 0}} \frac{b(\boldsymbol{v}, \boldsymbol{q})}{|\boldsymbol{v}|_{\boldsymbol{V}}} \ge \underline{\beta}|\boldsymbol{q}|_{\boldsymbol{Q}}, \quad \forall \boldsymbol{q} \in \boldsymbol{Q}. \tag{3.10}$$

For any given $\boldsymbol{q} = (q_S, q_D)$, let $\boldsymbol{v}^S \in \boldsymbol{H}^1(\Omega_S)$ be constructed, using the Stokes inf-sup condition (3.2), such that

$$\boldsymbol{v}^S|_\Gamma = 0, \quad \nabla \cdot \boldsymbol{v}^S = -q_S, \quad \|\boldsymbol{\epsilon}\left(\boldsymbol{v}^S\right)\|_{\Omega_S} \le \beta_s \|q_S\|_{\Omega_S}. \tag{3.11}$$

On the other hand, let $\phi \in H^{\frac{1}{2}}(\Gamma)$ be the Riesz representative of $q_D|_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$. We then define $\boldsymbol{v}^D \in \boldsymbol{H}^1(\Omega_S)$ as the bounded extension that satisfies

$$\boldsymbol{v}^D|_\Gamma = \phi \boldsymbol{n}, \quad \nabla \cdot \boldsymbol{v}^D = 0, \quad \|\boldsymbol{\epsilon}\left(\boldsymbol{v}^D\right)\|_{\Omega_S} \le \beta_0 \|\phi\|_{\frac{1}{2},\Gamma} = \beta_0 \|q_D\|_{-\frac{1}{2},\Gamma}. \tag{3.12}$$

We now set the test function $\boldsymbol{v}_0 := (2\mu)^{-1}\left(\boldsymbol{v}^S + \boldsymbol{v}^D\right)$. Noting that $\boldsymbol{\tau} \cdot \boldsymbol{v}_0 = 0$ on Γ, this function satisfies

$$\begin{aligned} b(\boldsymbol{v}_0, \boldsymbol{q}) &= -(2\mu)^{-1}\left(\nabla \cdot \boldsymbol{v}^S, q_S\right)_{\Omega_D} + (2\mu)^{-1}\left(\boldsymbol{n} \cdot \boldsymbol{v}^D, q_D\right)_\Gamma \\ &= (2\mu)^{-1}\|q_S\|^2_{\Omega_S} + (2\mu)^{-1}\|q_D\|^2_{-\frac{1}{2},\Gamma} = |\boldsymbol{q}|^2_{\boldsymbol{Q}}, \\ |\boldsymbol{v}_0|_{\boldsymbol{V}} &= (2\mu)^{\frac{1}{2}}\|\boldsymbol{\epsilon}\left((2\mu)^{-1}\left(\boldsymbol{v}^S + \boldsymbol{v}^D\right)\right)\|_{\Omega_S} \\ &\le (2\mu)^{-\frac{1}{2}}\left(\|\boldsymbol{\epsilon}\left(\boldsymbol{v}^S\right)\|_{\Omega_S} + \|\boldsymbol{\epsilon}\left(\boldsymbol{v}^D\right)\|_{\Omega_S}\right) \\ &\lesssim (2\mu)^{-\frac{1}{2}}\left(\|q_S\|_{\Omega_S} + \|q_D\|_{-\frac{1}{2},\Gamma}\right) = |\boldsymbol{q}|_{\boldsymbol{Q}}. \end{aligned}$$

Hence, condition (3.10) is fulfilled.

3.3. **Vector Laplace equation.**

Example 3.5. We consider the following mixed variational formulation of the vector Laplace equation [3, 37]: find $\boldsymbol{p} \in \boldsymbol{H}_0(\mathrm{curl}, \Omega), \boldsymbol{u} \in \boldsymbol{H}_0(\mathrm{div}, \Omega)$, such that

$$\begin{aligned} (\alpha \boldsymbol{p}, \boldsymbol{q}) - (\boldsymbol{u}, \mathrm{curl}\boldsymbol{q}) &= 0, \qquad \forall \boldsymbol{q} \in \boldsymbol{H}_0(\mathrm{curl}, \Omega), \\ -(\mathrm{curl}\boldsymbol{p}, \boldsymbol{v}) - (\mathrm{div}\boldsymbol{u}, \mathrm{div}\boldsymbol{v}) &= (f, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{div}, \Omega), \end{aligned} \tag{3.13}$$

where α is a positive scalar. Here, $\boldsymbol{H}_0(\mathrm{curl}, \Omega) = \{\boldsymbol{q} \in \boldsymbol{L}^2(\Omega) : \mathrm{curl}\boldsymbol{q} \in \boldsymbol{L}^2(\Omega),\ \boldsymbol{q} \times \boldsymbol{n} = 0 \text{ on } \partial\Omega\}$ and $\boldsymbol{H}_0(\mathrm{div}, \Omega) = \{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega) : \mathrm{div}\boldsymbol{v} \in L^2(\Omega),\ \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega\}$. We rewrite the above equations as

$$\begin{aligned} (\mathrm{div}\boldsymbol{u}, \mathrm{div}\boldsymbol{v}) + (\mathrm{curl}\boldsymbol{p}, \boldsymbol{v}) &= -(f, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{div}, \Omega), \\ (\boldsymbol{u}, \mathrm{curl}\boldsymbol{q}) - (\alpha \boldsymbol{p}, \boldsymbol{q}) &= 0, \qquad \forall \boldsymbol{q} \in \boldsymbol{H}_0(\mathrm{curl}, \Omega). \end{aligned} \tag{3.14}$$

The bilinear forms that define $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ are

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) &:= (\mathrm{div}\boldsymbol{u}, \mathrm{div}\boldsymbol{v}), \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}, \\ b(\boldsymbol{v}, \boldsymbol{p}) &:= (\mathrm{curl}\boldsymbol{p}, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{V}, \forall \boldsymbol{p} \in \boldsymbol{Q}, \\ c(\boldsymbol{p}, \boldsymbol{q}) &:= (\alpha \boldsymbol{p}, \boldsymbol{q}), \qquad \forall \boldsymbol{p}, \boldsymbol{q} \in \boldsymbol{Q}, \end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{H}_0(\mathrm{div}, \Omega), \boldsymbol{Q} = \boldsymbol{H}_0(\mathrm{curl}, \Omega)$. We fix $|\cdot|_Q$ and $|\cdot|_V$ to be

$$\begin{aligned} |\boldsymbol{q}|^2_Q &:= ((\alpha + 1)\mathrm{curl}\boldsymbol{q}, \mathrm{curl}\boldsymbol{q}), \qquad \forall \boldsymbol{q} \in \boldsymbol{Q}, \\ |\boldsymbol{v}|^2_V &:= (\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{V}. \end{aligned}$$

As in the previous examples, $a(\boldsymbol{v}, \boldsymbol{v}) = |\boldsymbol{v}|^2_V$ for all $\boldsymbol{v} \in \boldsymbol{V}$, that is, (2.32) is satisfied with $\underline{C}_a = 1$. In addition, noting that $B := \mathrm{curl}^* : \boldsymbol{V} \to \boldsymbol{Q}'$, we have

$$\begin{aligned} \|\boldsymbol{q}\|^2_Q &:= |\boldsymbol{q}|^2_Q + c(\boldsymbol{q}, \boldsymbol{q}) = ((\alpha + 1)\mathrm{curl}\boldsymbol{q}, \mathrm{curl}\boldsymbol{q}) + (\alpha \boldsymbol{q}, \boldsymbol{q}) = \langle \bar{Q}\boldsymbol{q}, \boldsymbol{q}\rangle_{Q' \times Q}, \ \forall \boldsymbol{q} \in \boldsymbol{Q}, \\ \|\boldsymbol{v}\|^2_V &:= |\boldsymbol{v}|^2_V + \langle B\boldsymbol{v}, \bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q' \times Q} \\ &= (\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + \left((\alpha I + \mathrm{curl}^*(\alpha + 1)\mathrm{curl})^{-1}\mathrm{curl}^*\boldsymbol{v}, \mathrm{curl}^*\boldsymbol{v}\right), \ \forall \boldsymbol{v} \in \boldsymbol{V}. \end{aligned}$$

Now we show that (2.33) is satisfied. For any $\boldsymbol{q} \in \boldsymbol{Q}$, choose $\boldsymbol{v}_0 = \mathrm{curl}\boldsymbol{q} \in \boldsymbol{V}$ to obtain

$$\begin{aligned}\|\boldsymbol{v}_0\|_V^2 &= (\mathrm{div}\,\mathrm{curl}\boldsymbol{q}, \mathrm{div}\,\mathrm{curl}\boldsymbol{q}) + \left((\alpha I + \mathrm{curl}^*(\alpha+1)\mathrm{curl})^{-1}\mathrm{curl}^*\mathrm{curl}\boldsymbol{q}, \mathrm{curl}^*\mathrm{curl}\boldsymbol{q}\right)\\ &= \left((\alpha I + \mathrm{curl}^*(\alpha+1)\mathrm{curl})^{-1}\mathrm{curl}^*\mathrm{curl}\boldsymbol{q}, \mathrm{curl}^*\mathrm{curl}\boldsymbol{q}\right)\\ &\le (\boldsymbol{q}, (\alpha+1)^{-1}\mathrm{curl}^*\mathrm{curl}\boldsymbol{q})\\ &= ((\alpha+1)^{-1}\mathrm{curl}\boldsymbol{q}, \mathrm{curl}\boldsymbol{q})\end{aligned}$$

and

(3.15)
$$\sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{b(\boldsymbol{v},\boldsymbol{q})}{\|\boldsymbol{v}\|_V} \ge \frac{b(\boldsymbol{v}_0,\boldsymbol{q})}{\|\boldsymbol{v}_0\|_V} = \frac{(\mathrm{curl}\boldsymbol{q}, \mathrm{curl}\boldsymbol{q})}{\|\boldsymbol{v}_0\|_V} \ge \frac{(\mathrm{curl}\boldsymbol{q}, \mathrm{curl}\boldsymbol{q})}{((\alpha+1)^{-1}\mathrm{curl}\boldsymbol{q}, \mathrm{curl}\boldsymbol{q})^{\frac{1}{2}}} = |\boldsymbol{q}|_Q.$$

Remark 3.6. Consider $\boldsymbol{H}(\mathrm{div}_0,\Omega) = \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div},\Omega) : \mathrm{div}\boldsymbol{v} = 0\}$. Then for any $\alpha > 0$ and $\boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{curl},\Omega) \cap \boldsymbol{H}(\mathrm{div}_0,\Omega)$, we have

$$\begin{aligned}((\alpha I + (\alpha+1)\mathrm{curl}^*\mathrm{curl})\boldsymbol{v}, \boldsymbol{v}) &\le (c_P\alpha + (\alpha+1))(\mathrm{curl}^*\mathrm{curl}\boldsymbol{v}, \boldsymbol{v})\\ &\le c_1((\alpha+1)\mathrm{curl}^*\mathrm{curl}\boldsymbol{v}, \boldsymbol{v}),\end{aligned}$$

where c_P denotes the Poincaré constant for the curl operator and $c_1 = c_P + 1$. Hence,

$$((\alpha I + (\alpha+1)\mathrm{curl}^*\mathrm{curl})^{-1}\boldsymbol{f}, \boldsymbol{f}) \ge c_1^{-1}((\alpha+1)^{-1}(\mathrm{curl}^*\mathrm{curl})^{-1}\boldsymbol{f}, \boldsymbol{f}),$$

for any $\boldsymbol{f} \in (\alpha I + (\alpha+1)\mathrm{curl}^*\mathrm{curl})\,(\boldsymbol{H}_0(\mathrm{curl},\Omega) \cap \boldsymbol{H}(\mathrm{div}_0,\Omega))$. Now, by using the Helmholtz decomposition $\boldsymbol{v} = \mathrm{curl}\boldsymbol{w} + \nabla z$ of $\boldsymbol{v}$ and choosing $\boldsymbol{f} = \mathrm{curl}^*\boldsymbol{v}$ we obtain

$$c_1^{-1}((\alpha+1)^{-1}\mathrm{curl}\boldsymbol{w}, \mathrm{curl}\boldsymbol{w}) \le \left((\alpha I + (\alpha+1)\mathrm{curl}^*\mathrm{curl})^{-1}\mathrm{curl}^*\boldsymbol{v}, \mathrm{curl}^*\boldsymbol{v}\right). \tag{3.16}$$

On the other hand, the Poincare's inequality for vector Laplacian (see Theorem 2.2 in [3]) implies

(3.17)
$$(\nabla z, \nabla z) \le c_{P_v}\left((\mathrm{div}(\nabla z), \mathrm{div}(\nabla z)) + (\mathrm{curl}(\nabla z), \mathrm{curl}(\nabla z))\right) = c_{P_v}(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}),$$

where c_{P_v} denotes the Poincaré constant for vector Laplacian. By multiplying (3.17) with $c_1^{-1}(\alpha+1)^{-1}$ we obtain

(3.18)
$$c_1^{-1}(\alpha+1)^{-1}(\nabla z, \nabla z) \le c_1^{-1}(\alpha+1)^{-1}c_{P_v}(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) \le c_1^{-1}c_{P_v}(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}).$$

Combining (3.16) and (3.18) and noting that $(\boldsymbol{v},\boldsymbol{v}) = (\mathrm{curl}\boldsymbol{w}, \mathrm{curl}\boldsymbol{w}) + (\nabla z, \nabla z)$, we have

$$c_1^{-1}(\alpha+1)^{-1}\|\boldsymbol{v}\|^2 \le c_1^{-1}c_{P_v}\|\mathrm{div}\boldsymbol{v}\|^2 + \left((\alpha I + (\alpha+1)\mathrm{curl}^*\mathrm{curl})^{-1}\mathrm{curl}^*\boldsymbol{v}, \mathrm{curl}^*\boldsymbol{v}\right).$$

Next we add $c_1^{-1}\|\mathrm{div}\boldsymbol{v}\|^2$ to both sides and get

$$\begin{aligned}&c_1^{-1}\left((\alpha+1)^{-1}\|\boldsymbol{v}\|^2 + \|\mathrm{div}\boldsymbol{v}\|^2\right)\\ &\quad\le c_1^{-1}(c_{P_v}+1)\|\mathrm{div}\boldsymbol{v}\|^2 + \left((\alpha I + (\alpha+1)\mathrm{curl}^*\mathrm{curl})^{-1}\mathrm{curl}^*\boldsymbol{v}, \mathrm{curl}^*\boldsymbol{v}\right).\end{aligned}$$

By multiplying with c_1 and setting $c_2 := \max\{c_1, (c_{P_v}+1)\}$ it follows that

$$(\alpha+1)^{-1}\|\boldsymbol{v}\|^2 + (\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) \le c_2\|\boldsymbol{v}\|_V^2.$$

Moreover, we have

$$\begin{aligned}\|\boldsymbol{v}\|_V^2 &= |\boldsymbol{v}|_V^2 + \langle B\boldsymbol{v}, \bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q'\times Q}\\ &= (\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + \left((\alpha I + \mathrm{curl}^*(\alpha+1)\mathrm{curl})^{-1}\mathrm{curl}^*\boldsymbol{v}, \mathrm{curl}^*\boldsymbol{v}\right)\\ &\le (\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + \left((\alpha+1)^{-1}\boldsymbol{v}, \boldsymbol{v}\right).\end{aligned}$$

Therefore we conclude that $\|\boldsymbol{v}\|_V^2$ and $(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + ((\alpha+1)^{-1}\boldsymbol{v}, \boldsymbol{v})$ are equivalent.

The corresponding norm-equivalent preconditioner is then given by

$$\mathcal{B} := \begin{bmatrix} ((\alpha+1)^{-1}I - \nabla\mathrm{div})^{-1} & \\ & (\alpha I + (\alpha+1)\mathrm{curl}^*\mathrm{curl})^{-1}\end{bmatrix}.$$

To solve the $\boldsymbol{H}(\mathrm{curl})$ subproblem, one can use the preconditioner proposed in [30], the multigrid method in [2], or domain decomposition methods [59].

3.4. **Poromechanics.**

Example 3.7. The two-field formulation of the quasi-static Biot's consolidation model after semidiscretization in time by the implicit Euler method as studied in [1, 43] reads: find $(\boldsymbol{u}, p_F) \in \boldsymbol{H}_0^1(\Omega) \times H_0^1(\Omega)$ such that

$$\begin{aligned}(\varepsilon(\boldsymbol{u}), \varepsilon(\boldsymbol{v})) + \lambda(\mathrm{div}\boldsymbol{u}, \mathrm{div}\boldsymbol{v}) - \alpha(p_F, \mathrm{div}\boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega),\\ -\alpha(\mathrm{div}\boldsymbol{u}, q_F) - c_0(p_F, q_F) - (\kappa\nabla p_F, \nabla q_F) &= (g, q_F), \qquad \forall q \in H_0^1(\Omega),\end{aligned} \tag{3.19}$$

where $\lambda \ge 0$ is a scaled Lamé coefficient, $c_0 \ge 0$ is the storage coefficient, κ is the (scaled) hydraulic conductivity, and α is the (scaled) Biot-Willis coefficient.

The bilinear forms defining $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ are given by

$$\begin{aligned}a(\boldsymbol{u}, \boldsymbol{v}) &:= (\varepsilon(\boldsymbol{u}), \varepsilon(\boldsymbol{v})) + \lambda(\mathrm{div}\boldsymbol{u}, \mathrm{div}\boldsymbol{v}), \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V},\\ b(\boldsymbol{v}, q_F) &:= -\alpha(\mathrm{div}\boldsymbol{v}, q_F), \qquad \forall \boldsymbol{v} \in \boldsymbol{V}, \forall q_F \in Q,\\ c(p_F, q_F) &:= c_0(p_F, q_F) + (\kappa\nabla p_F, \nabla q_F), \qquad \forall p_F, q_F \in Q,\end{aligned}$$

where $Q := H_0^1(\Omega)$, $\boldsymbol{V} := \boldsymbol{H}_0^1(\Omega)$. We define $|\cdot|_Q$, $|\cdot|_V$ to be

$$\begin{aligned}|q_F|_Q^2 &:= \eta(q_F, q_F), \qquad \forall q_F \in Q,\\ |\boldsymbol{v}|_V^2 &:= (\varepsilon(\boldsymbol{v}), \varepsilon(\boldsymbol{v})) + \lambda(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{V},\end{aligned}$$

where the parameter $\eta > 0$ is to be determined later. As before, $a(\boldsymbol{v}, \boldsymbol{v}) \ge |\boldsymbol{v}|_V^2$ for all $\boldsymbol{v} \in \boldsymbol{V}$, that is, (2.32) is satisfied with $\underline{C}_a = 1$. It remains to show (2.33). As in Example 3.3, it is easy to see that

$$\langle B\boldsymbol{v}, \bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q'\times Q} \le \frac{\alpha^2}{\eta}(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}),$$

where $B : \boldsymbol{V} \to Q'$, $B := -\alpha\mathrm{div}$. Therefore, we obtain

$$\begin{aligned}\|\boldsymbol{v}\|_V^2 &= (\varepsilon(\boldsymbol{v}), \varepsilon(\boldsymbol{v})) + \lambda(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + \langle B\boldsymbol{v}, \bar{Q}^{-1}B\boldsymbol{v}\rangle_{Q'\times Q}\\ &\le (\varepsilon(\boldsymbol{v}), \varepsilon(\boldsymbol{v})) + \left(\lambda + \frac{\alpha^2}{\eta}\right)(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) \le \left(1 + \lambda + \frac{\alpha^2}{\eta}\right)\|\boldsymbol{v}\|_1^2.\end{aligned} \tag{3.20}$$

We choose $\boldsymbol{v}_0$ such that $-\mathrm{div}\boldsymbol{v}_0 = \frac{1}{\sqrt{1+\lambda}} q_F$ and use (3.2) to obtain $\|\boldsymbol{v}_0\|_1 \leq \frac{1}{\beta_s}\frac{1}{\sqrt{1+\lambda}}\|q_F\|$, and finally

$$\begin{aligned}\sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{b(\boldsymbol{v}, q_F)}{\|\boldsymbol{v}\|_V} &\geq \frac{b(\boldsymbol{v}_0, q_F)}{\|\boldsymbol{v}_0\|_V} = \frac{\frac{\alpha}{\sqrt{1+\lambda}}\|q_F\|^2}{\|\boldsymbol{v}_0\|_V} \geq \frac{\frac{\alpha}{\sqrt{1+\lambda}}}{\sqrt{\left(1+\lambda+\frac{\alpha^2}{\eta}\right)}}\frac{\|q_F\|^2}{\|\boldsymbol{v}_0\|_1} \\ &\geq \frac{\beta_s\alpha}{\sqrt{\left(1+\lambda+\frac{\alpha^2}{\eta}\right)}}\frac{\|q_F\|^2}{\|q_F\|} = \frac{\beta_s\alpha}{\sqrt{\left(1+\lambda+\frac{\alpha^2}{\eta}\right)}}\frac{1}{\sqrt{\eta}}|q_F|_Q.\end{aligned} \tag{3.21}$$

For $\eta := \frac{\alpha^2}{(1+\lambda)} > 0$ the right-hand side of (3.21) is bounded from below by $\frac{\beta_s}{\sqrt{2}}|q_F|_Q$ which shows (2.33) with $\underline{\beta} = \frac{1}{\sqrt{2}}\beta_s$. Note that there are also other possible choices for η.

We conclude that

$$\mathcal{B} := \begin{bmatrix} (-\mathrm{div}\varepsilon - (1+\lambda)\nabla\mathrm{div})^{-1} & \\ & \left(\left(c_0 + \alpha^2/(1+\lambda)\right) I - \mathrm{div}\kappa\nabla\right)^{-1} \end{bmatrix}$$

provides a norm-equivalent preconditioner for the combined norm, where we have used (3.20). To solve the elasticity subproblem, one can use the multigrid method proposed in [31, 36].

Example 3.8. By introducing $p_S = -\lambda\mathrm{div}\boldsymbol{u}$ and substituting $\alpha p_F \to p_F, c_0\alpha^{-2} \to c_0, \kappa\alpha^{-2} \to \kappa, \alpha^{-1}g \to g$ in *Example* 3.7 we obtain the following three-field variational formulation of Biot's model, see [43],

$$\begin{aligned} (\varepsilon(\boldsymbol{u}), \varepsilon(\boldsymbol{v})) - (p_S + p_F, \mathrm{div}\boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega), \\ -(\mathrm{div}\boldsymbol{u}, q_S) - \lambda^{-1}(p_S, q_S) &= 0, \qquad \forall q_S \in L_0^2(\Omega), \\ -(\mathrm{div}\boldsymbol{u}, q_F) - c_0(p_F, q_F) - (\kappa\nabla p_F, \nabla q_F) &= (g, q_F), \qquad \forall q_F \in H_0^1(\Omega). \end{aligned} \tag{3.22}$$

The bilinear forms that determine $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ are

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) &:= (\varepsilon(\boldsymbol{u}), \varepsilon(\boldsymbol{v})), \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}, \\ b(\boldsymbol{v}, \boldsymbol{q}) &:= -(\mathrm{div}\boldsymbol{v}, q_S) - (\mathrm{div}\boldsymbol{v}, q_F), \qquad \forall \boldsymbol{v} \in \boldsymbol{V}, \forall \boldsymbol{q} \in \boldsymbol{Q}, \\ c(\boldsymbol{p}, \boldsymbol{q}) &:= \lambda^{-1}(p_S, q_S) + c_0(p_F, q_F) + (\kappa\nabla p_F, \nabla q_F), \qquad \forall \boldsymbol{p}, \boldsymbol{q} \in \boldsymbol{Q}, \end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{H}_0^1(\Omega)$, $\boldsymbol{Q} = L_0^2(\Omega) \times H_0^1(\Omega)$ and $\boldsymbol{p} = (p_S; p_F)$, $\boldsymbol{q} = (q_S; q_F)$. Then the operator B is given by

$$B := \begin{pmatrix} -\mathrm{div} \\ -\mathrm{div} \end{pmatrix}.$$

Moreover, we define $|\cdot|_Q$, $|\cdot|_V$ to be

$$\begin{aligned} |\boldsymbol{q}|_Q^2 &:= \left(\begin{pmatrix} I & I \\ I & I \end{pmatrix}\begin{pmatrix} q_S \\ q_{F,0} \end{pmatrix}, \begin{pmatrix} q_S \\ q_{F,0} \end{pmatrix}\right) = \|q_S + q_{F,0}\|^2, \qquad \forall \boldsymbol{q} \in \boldsymbol{Q}, \\ |\boldsymbol{v}|_V^2 &:= (\varepsilon(\boldsymbol{v}), \varepsilon(\boldsymbol{v})), \qquad \forall \boldsymbol{v} \in \boldsymbol{V}, \end{aligned}$$

where $q_{F,0} := P_0 q_F$ and P_0 is the L^2 projection from $L^2(\Omega)$ to $L^2_0(\Omega)$. Then

$$\begin{aligned}\|\boldsymbol{q}\|_Q^2 &= \left(\begin{pmatrix} I & I \\ I & I \end{pmatrix}\begin{pmatrix} q_S \\ q_{F,0}\end{pmatrix}, \begin{pmatrix} q_S \\ q_{F,0}\end{pmatrix}\right) + \left(\begin{pmatrix} \lambda^{-1}I & 0 \\ 0 & c_0 I - \mathrm{div}\kappa\nabla\end{pmatrix}\begin{pmatrix} q_S \\ q_F\end{pmatrix}, \begin{pmatrix} q_S \\ q_F\end{pmatrix}\right)\\ &= \left(\begin{pmatrix} (1+\lambda^{-1})I & P_0 \\ P_0 & P_0 + c_0 I - \mathrm{div}\kappa\nabla\end{pmatrix}\begin{pmatrix} q_S \\ q_F\end{pmatrix}, \begin{pmatrix} q_S \\ q_F\end{pmatrix}\right) = (\bar{Q}\boldsymbol{q}, \boldsymbol{q})\\ &= \langle \bar{Q}\boldsymbol{q}, \boldsymbol{q}\rangle_{\boldsymbol{Q}'\times\boldsymbol{Q}}.\end{aligned}$$

As in the previous examples, (2.32) is satisfied with $\underline{C}_a = 1$. Next, we choose $\boldsymbol{v}_0$ such that $-\mathrm{div}\boldsymbol{v}_0 = q_S + q_{F,0}$ for which we have $\|\boldsymbol{v}_0\|_1 \le \beta_s^{-1}\|q_S + q_{F,0}\|$, see (3.2). Then

$$b(\boldsymbol{v}_0, \boldsymbol{q}) = \|q_S + q_{F,0}\|^2 = |\boldsymbol{q}|_Q^2.$$

Moreover,

$$\begin{aligned}\|\boldsymbol{v}_0\|_V^2 &= (\varepsilon(\boldsymbol{v}_0), \varepsilon(\boldsymbol{v}_0)) + (\bar{Q}^{-1}B\boldsymbol{v}_0, B\boldsymbol{v}_0)\\ &= (\varepsilon(\boldsymbol{v}_0), \varepsilon(\boldsymbol{v}_0)) + \left(\bar{Q}^{-1}\begin{pmatrix} -\mathrm{div}\boldsymbol{v}_0 \\ -\mathrm{div}\boldsymbol{v}_0\end{pmatrix}, \begin{pmatrix} -\mathrm{div}\boldsymbol{v}_0 \\ -\mathrm{div}\boldsymbol{v}_0\end{pmatrix}\right)\\ &\le \|\boldsymbol{v}_0\|_1^2 + \frac{1}{4}\left(\bar{Q}^{-1}\begin{pmatrix} I & P_0 \\ P_0 & P_0\end{pmatrix}\begin{pmatrix} \mathrm{div}\boldsymbol{v}_0 \\ \mathrm{div}\boldsymbol{v}_0\end{pmatrix}, \begin{pmatrix} I & P_0 \\ P_0 & P_0\end{pmatrix}\begin{pmatrix} \mathrm{div}\boldsymbol{v}_0 \\ \mathrm{div}\boldsymbol{v}_0\end{pmatrix}\right)\\ &\le \|\boldsymbol{v}_0\|_1^2 + \frac{1}{4}\left(\begin{pmatrix} I & P_0 \\ P_0 & P_0\end{pmatrix}\begin{pmatrix} \mathrm{div}\boldsymbol{v}_0 \\ \mathrm{div}\boldsymbol{v}_0\end{pmatrix}, \begin{pmatrix} \mathrm{div}\boldsymbol{v}_0 \\ \mathrm{div}\boldsymbol{v}_0\end{pmatrix}\right) = \|\boldsymbol{v}_0\|_1^2 + (\mathrm{div}\boldsymbol{v}_0, \mathrm{div}\boldsymbol{v}_0)\\ &\le \beta_s^{-2}\|q_S + q_{F,0}\|^2 + \|q_S + q_{F,0}\|^2 = (\beta_s^{-2}+1)|\boldsymbol{q}|_Q^2.\end{aligned}$$

Now (2.33) follows directly from

$$\sup_{\boldsymbol{v}\in V}\frac{b(\boldsymbol{v}, \boldsymbol{q})}{\|\boldsymbol{v}\|_V} \ge \frac{b(\boldsymbol{v}_0, \boldsymbol{q})}{\|\boldsymbol{v}_0\|_V} \ge \frac{|\boldsymbol{q}|_Q^2}{\sqrt{(\beta_s^{-2}+1)}|\boldsymbol{q}|_Q} =: \underline{\beta}|\boldsymbol{q}|_Q, \qquad \forall \boldsymbol{q}\in\boldsymbol{Q}.$$

Using the fitted norms for the constructions of a norm-equivalent preconditioner, cf. [47], results in

$$\mathcal{B} := \begin{bmatrix} (-\mathrm{div}\varepsilon)^{-1} & \\ & \begin{pmatrix} (1+\lambda^{-1})I & P_0 \\ P_0 & P_0 + c_0 I - \mathrm{div}\kappa\nabla\end{pmatrix}^{-1}\end{bmatrix}.$$

Remark 3.9. In [43], the authors showed that the three-field formulation for Biot's model in *Example* 3.8 is not stable under the Q-seminorm defined by $|\boldsymbol{q}|_Q^2 = \|p_S\|^2 + \|p_F\|^2$.

Example 3.10. By introducing the total pressure $p_T = p_S + p_F$ in *Example* 3.8, another discrete in time three-field formulation of the quasi-static Biot's consolidation model, see [43], is obtained and has the form

$$\begin{aligned}(\varepsilon(\boldsymbol{u}), \varepsilon(\boldsymbol{v})) - (p_T, \mathrm{div}\boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v}), \qquad \forall \boldsymbol{v}\in\boldsymbol{H}_0^1(\Omega),\\ -(\mathrm{div}\boldsymbol{u}, q_T) - (\lambda^{-1}p_T, q_T) + (\alpha\lambda^{-1}p_F, q_T) &= 0, \qquad \forall q_T \in L^2(\Omega),\\ (\alpha\lambda^{-1}p_T, q_F) - ((\alpha^2\lambda^{-1} + c_0)p_F, q_F) - (\kappa\nabla p_F, \nabla q_F) &= (g, q_F), \qquad \forall q_F \in H_0^1(\Omega).\end{aligned} \tag{3.23}$$

Here, $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ is constructed from

$$\begin{aligned}
a(\boldsymbol{u},\boldsymbol{v}) :=& (\varepsilon(\boldsymbol{u}),\varepsilon(\boldsymbol{v})), \qquad \forall \boldsymbol{u},\boldsymbol{v}\in\boldsymbol{V},\\
b(\boldsymbol{v},\boldsymbol{q}) :=& -(\mathrm{div}\boldsymbol{v},q_T), \qquad \forall \boldsymbol{v}\in\boldsymbol{V},\forall \boldsymbol{q}\in\boldsymbol{Q},\\
c(\boldsymbol{p},\boldsymbol{q}) :=& (\lambda^{-1}p_T,q_T)-(\alpha\lambda^{-1}p_F,q_T)-(\alpha\lambda^{-1}p_T,q_F)+((\alpha^2\lambda^{-1}+c_0)p_F,q_F)\\
&+(\kappa\nabla p_F,\nabla q_F), \qquad \forall \boldsymbol{p},\boldsymbol{q}\in\boldsymbol{Q},
\end{aligned}$$

where $\boldsymbol{V}=\boldsymbol{H}^1_0(\Omega)$, $\boldsymbol{Q}=L^2(\Omega)\times H^1_0(\Omega)$ and $\boldsymbol{p}=(p_T;p_F)$, $\boldsymbol{q}=(q_T;q_F)$. Obviously, the operator $B:\boldsymbol{V}\to\boldsymbol{Q}'$ is defined by

$$B:=\begin{pmatrix}-\mathrm{div}\\ 0\end{pmatrix}.$$

We next set

$$\begin{aligned}
|\boldsymbol{q}|^2_Q &:= (q_{T,0},q_{T,0}), \qquad \forall \boldsymbol{q}\in\boldsymbol{Q},\\
|\boldsymbol{v}|^2_V &:= (\varepsilon(\boldsymbol{v}),\varepsilon(\boldsymbol{v})), \qquad \forall \boldsymbol{v}\in\boldsymbol{V},
\end{aligned}$$

where $q_{T,0}:=P_0q_T$ is the L^2 projection of $q_T\in L^2(\Omega)$ to $L^2_0(\Omega)$. Now we choose $\boldsymbol{v}_0\in\boldsymbol{V}=\boldsymbol{H}^1_0(\Omega)$ such that $-\mathrm{div}\boldsymbol{v}_0=q_{T,0}$ for which it holds $\|\boldsymbol{v}_0\|_1\le\beta_s^{-1}\|q_{T,0}\|=\beta_s^{-1}|\boldsymbol{q}|_Q$, see (3.2), and also

$$b(\boldsymbol{v}_0,\boldsymbol{q})=(q_{T,0},q_{T,0})=|\boldsymbol{q}|^2_Q.$$

From the definition of $\|\cdot\|_Q$, and using similar arguments as in the previous examples, we obtain

$$\begin{aligned}
\|\boldsymbol{v}_0\|^2_V &= (\varepsilon(\boldsymbol{v}_0),\varepsilon(\boldsymbol{v}_0))+\langle B\boldsymbol{v}_0,\bar{Q}^{-1}B\boldsymbol{v}_0\rangle \le (\varepsilon(\boldsymbol{v}_0),\varepsilon(\boldsymbol{v}_0))+(\mathrm{div}\boldsymbol{v}_0,\mathrm{div}\boldsymbol{v}_0)\\
&\le 2\|\boldsymbol{v}_0\|^2_1\le 2\beta_s^{-2}|\boldsymbol{q}|^2_Q.
\end{aligned}$$

Again, (2.32) is satisfied with $\underline{C}_a=1$ while (2.33) follows from

$$\sup_{\boldsymbol{v}\in V}\frac{b(\boldsymbol{v},\boldsymbol{q})}{\|\boldsymbol{v}\|_V}\ge\frac{b(\boldsymbol{v}_0,\boldsymbol{q})}{\|\boldsymbol{v}_0\|_V}\ge\beta_s\frac{|\boldsymbol{q}|^2_Q}{|\boldsymbol{q}|_Q}=:\underline{\beta}|\boldsymbol{q}|_Q, \qquad \forall \boldsymbol{q}\in\boldsymbol{Q}.$$

Thus, the fitted norms generate the norm-equivalent preconditioner

$$\mathcal{B}:=\begin{bmatrix}(-\mathrm{div}\varepsilon)^{-1} & \\ & \begin{pmatrix}\lambda^{-1}I+P_0 & -\alpha\lambda^{-1}I\\ -\alpha\lambda^{-1}I & \alpha^2\lambda^{-1}I+c_0I-\mathrm{div}\kappa\nabla\end{pmatrix}^{-1}\end{bmatrix}.$$

Remark 3.11. The arguments presented above are valid for a vanishing storage coefficient, i.e., $c_0=0$. Moreover, our analysis shows how *Example* 3.8 and *Example* 3.10 are related to each other. In fact, by the transformation $\begin{pmatrix}p_T\\ p_F\end{pmatrix}=\begin{pmatrix}I & I\\ 0 & I\end{pmatrix}\begin{pmatrix}p_S\\ p_F\end{pmatrix}$ or, equivalently, $\begin{pmatrix}p_S\\ p_F\end{pmatrix}=\begin{pmatrix}I & -I\\ 0 & I\end{pmatrix}\begin{pmatrix}p_T\\ p_F\end{pmatrix}$, we can derive the stability and preconditioners of *Example* 3.8 and *Example* 3.10 from each other.

Remark 3.12. Note that for the case $c_0 = \alpha^2\lambda^{-1}$, as considered in [43], we can further estimate

$$\begin{aligned} c(\boldsymbol{q},\boldsymbol{q}) :=& (\lambda^{-1}q_T, q_T) - 2(\alpha\lambda^{-1}q_F, q_T) + 2(\alpha^2\lambda^{-1}q_F, q_F) + (\kappa\nabla q_F, \nabla q_F) \\ \geq& \frac{1}{4}(\lambda^{-1}q_T, q_T) + \frac{2}{3}(\alpha^2\lambda^{-1}q_F, q_F) + (\kappa\nabla q_F, \nabla q_F) \\ \geq& \frac{1}{4}\left((\lambda^{-1}q_T, q_T) + (\alpha^2\lambda^{-1}q_F, q_F) + (\kappa\nabla q_F, \nabla q_F)\right). \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \|\boldsymbol{q}\|_Q^2 &= |\boldsymbol{q}|_Q^2 + c(\boldsymbol{q},\boldsymbol{q}) \\ &\geq \frac{1}{4}\left((q_{T,0}, q_{T,0}) + (\lambda^{-1}q_T, q_T) + (\alpha^2\lambda^{-1}q_F, q_F) + (\kappa\nabla q_F, \nabla q_F)\right), \end{aligned}$$

from which we conclude the stability result and preconditioner shown in [43]:

$$\mathcal{B}_0 := \begin{bmatrix} (-\mathrm{div}\varepsilon)^{-1} & \\ & \begin{pmatrix} \lambda^{-1}I + P_0 & \\ & \alpha^2\lambda^{-1}I - \mathrm{div}\kappa\nabla \end{pmatrix}^{-1} \end{bmatrix}.$$

Example 3.13. Next we consider a four-field formulation of Biot's model in which a total pressure has been introduced [39]. The variational problem reads as

$$\begin{aligned} 2\mu(\varepsilon(\boldsymbol{u}),\varepsilon(\boldsymbol{v})) + (p_T, \mathrm{div}\boldsymbol{v}) &= (\boldsymbol{f},\boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega), \\ \frac{1}{\tau\kappa}(\boldsymbol{w},\boldsymbol{z}) - (p, \mathrm{div}\boldsymbol{z}) &= 0, \qquad \forall \boldsymbol{z} \in \boldsymbol{H}_0(\mathrm{div},\Omega), \\ (\mathrm{div}\boldsymbol{u}, q_T) - \lambda^{-1}(p_T, q_T) - \frac{\alpha}{\lambda}(p, q_T) &= 0, \qquad \forall q_T \in L_0^2(\Omega), \\ -(\mathrm{div}\boldsymbol{w}, q) - \frac{\alpha}{\lambda}(p_T, q) - \left(c_0 + \frac{\alpha^2}{\lambda}\right)(p, q) &= (f, q), \qquad \forall q \in L_0^2(\Omega), \end{aligned} \tag{3.24}$$

with μ and λ the Lamé coefficients, τ the time step size, κ the hydraulic conductivity, α the Biot-Willis coefficient, and $c_0 \geq 0$ the constrained specific storage coefficient.

The bilinear forms defining $\mathcal{A}((\cdot;\cdot),(\cdot;\cdot))$ in Example 3.13 are given by

$$\begin{aligned} a(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}}) :=& 2\mu(\varepsilon(\boldsymbol{u}),\varepsilon(\boldsymbol{v})) + \frac{1}{\tau\kappa}(\boldsymbol{w},\boldsymbol{z}), \qquad \forall \bar{\boldsymbol{u}},\bar{\boldsymbol{v}} \in \boldsymbol{V}, \\ b(\bar{\boldsymbol{v}},\boldsymbol{q}) :=& (\mathrm{div}\boldsymbol{v}, q_T) - (\mathrm{div}\boldsymbol{z}, q), \qquad \forall \bar{\boldsymbol{v}} \in \boldsymbol{V}, \forall \boldsymbol{q} \in \boldsymbol{Q}, \\ c(\boldsymbol{p},\boldsymbol{q}) :=& \lambda^{-1}(p_T, q_T) + \frac{\alpha}{\lambda}(p, q_T) + \frac{\alpha}{\lambda}(p_T, q) + \left(c_0 + \frac{\alpha^2}{\lambda}\right)(p, q), \qquad \forall \boldsymbol{p},\boldsymbol{q} \in \boldsymbol{Q}, \end{aligned}$$

where $\bar{\boldsymbol{u}} := (\boldsymbol{u};\boldsymbol{w}), \bar{\boldsymbol{v}} := (\boldsymbol{v};\boldsymbol{z}) \in \boldsymbol{V} = \boldsymbol{H}_0^1(\Omega) \times \boldsymbol{H}_0(\mathrm{div},\Omega)$, $\boldsymbol{p} := (p_T;p), \boldsymbol{q} := (q_T;q) \in \boldsymbol{Q} = L_0^2(\Omega) \times L_0^2(\Omega)$. Hence, B has the form

$$B = \begin{pmatrix} \mathrm{div} & 0 \\ 0 & -\mathrm{div} \end{pmatrix}.$$

Using the norm fitting technique, we define

$$\begin{aligned} |\boldsymbol{q}|_Q^2 &:= \frac{1}{2\mu}(q_T, q_T) + \tau\kappa(q, q), \qquad \forall \boldsymbol{q} \in \boldsymbol{Q}, \\ |\bar{\boldsymbol{v}}|_V^2 &:= 2\mu(\varepsilon(\boldsymbol{v}),\varepsilon(\boldsymbol{v})) + \frac{1}{\tau\kappa}(\boldsymbol{z},\boldsymbol{z}), \qquad \forall \bar{\boldsymbol{v}} \in \boldsymbol{V}. \end{aligned}$$

For this choice of $|\cdot|_V$ the coercivity estimate (2.32) is again fulfilled with $\underline{C_a} = 1$.

Now, in view of the Stokes and Darcy inf-sup conditions (3.2) and (3.1), for any $\boldsymbol{q} = (q_T; q)$ we can choose $\bar{\boldsymbol{v}}_0 = (\boldsymbol{v}_0; \boldsymbol{z}_0)$ such that $\mathrm{div}\boldsymbol{v}_0 = \frac{1}{2\mu} q_T$ and $-\mathrm{div}\boldsymbol{z}_0 = \tau\kappa\, q$ and it holds that

$$\|\boldsymbol{v}_0\|_1 \leq \frac{\beta_s^{-1}}{2\mu}\|q_T\|, \qquad \|\boldsymbol{z}_0\|_{\mathrm{div}} \leq \beta_d^{-1}\tau\kappa\|q\|. \tag{3.25}$$

Then we have

$$b(\bar{\boldsymbol{v}}_0, \boldsymbol{q}) = (\mathrm{div}\boldsymbol{v}_0, q_T) - (\mathrm{div}\boldsymbol{z}_0, q) = \frac{1}{2\mu}(q_T, q_T) + \tau\kappa(q, q) = |\boldsymbol{q}|_Q^2 \tag{3.26}$$

and

$$\begin{aligned}
\|\bar{\boldsymbol{v}}_0\|_V^2 &= 2\mu(\varepsilon(\boldsymbol{v}_0), \varepsilon(\boldsymbol{v}_0)) + \frac{1}{\tau\kappa}(\boldsymbol{z}_0, \boldsymbol{z}_0) + \langle B\bar{\boldsymbol{v}}_0, \bar{Q}^{-1}B\bar{\boldsymbol{v}}_0\rangle \\
&= 2\mu(\varepsilon(\boldsymbol{v}_0), \varepsilon(\boldsymbol{v}_0)) + \frac{1}{\tau\kappa}(\boldsymbol{z}_0, \boldsymbol{z}_0) + \left(\bar{Q}^{-1}\begin{pmatrix}\mathrm{div}\boldsymbol{v}_0\\ -\mathrm{div}\boldsymbol{z}_0\end{pmatrix}, \begin{pmatrix}\mathrm{div}\boldsymbol{v}_0\\ -\mathrm{div}\boldsymbol{z}_0\end{pmatrix}\right) \\
&\leq 2\mu\|\boldsymbol{v}_0\|_1^2 + \frac{1}{\tau\kappa}\|\boldsymbol{z}_0\|^2 + \left(\begin{pmatrix}2\mu I & 0\\ 0 & \frac{1}{\tau\kappa}I\end{pmatrix}\begin{pmatrix}\mathrm{div}\boldsymbol{v}_0\\ -\mathrm{div}\boldsymbol{z}_0\end{pmatrix}, \begin{pmatrix}\mathrm{div}\boldsymbol{v}_0\\ -\mathrm{div}\boldsymbol{z}_0\end{pmatrix}\right) \\
&= 2\mu\|\boldsymbol{v}_0\|_1^2 + \frac{1}{\tau\kappa}\|\boldsymbol{z}_0\|^2 + 2\mu\|\mathrm{div}\boldsymbol{v}_0\|^2 + \frac{1}{\tau\kappa}\|\mathrm{div}\boldsymbol{z}_0\|^2 \\
&\leq 4\mu\|\boldsymbol{v}_0\|_1^2 + \frac{1}{\tau\kappa}\|\boldsymbol{z}_0\|_{\mathrm{div}}^2.
\end{aligned}$$

Now by (3.25) and the definition of $|\cdot|_Q$, we obtain

$$\begin{aligned}
\|\bar{\boldsymbol{v}}_0\|_V^2 &\leq 4\mu\|\boldsymbol{v}_0\|_1^2 + \frac{1}{\tau\kappa}\|\boldsymbol{z}_0\|_{\mathrm{div}}^2 \leq 4\mu\frac{\beta_s^{-2}}{4\mu^2}\|q_T\|^2 + \frac{1}{\tau\kappa}\beta_d^{-2}\tau^2\kappa^2\|q\|^2 \\
&\leq 2\max\{\beta_s^{-2}, \beta_d^{-2}\}\left(\frac{1}{2\mu}\|q_T\|^2 + \tau\kappa\|q\|^2\right) \leq 2\max\{\beta_s^{-2}, \beta_d^{-2}\}|\boldsymbol{q}|_Q^2.
\end{aligned} \tag{3.27}$$

Hence, in Example 3.14 (2.33) follows from

$$\sup_{\bar{\boldsymbol{v}}\in\boldsymbol{V}} \frac{b(\bar{\boldsymbol{v}}, \boldsymbol{q})}{\|\bar{\boldsymbol{v}}\|_V} \geq \frac{b(\bar{\boldsymbol{v}}_0, \boldsymbol{q})}{\|\bar{\boldsymbol{v}}_0\|_V} \geq \frac{|\boldsymbol{q}|_Q^2}{\sqrt{2\max\{\beta_s^{-2}, \beta_d^{-2}\}}|\boldsymbol{q}|_Q} =: \underline{\beta}|\boldsymbol{q}|_Q, \qquad \forall \boldsymbol{q} \in \boldsymbol{Q},$$

where we have used (3.26) and (3.27).

From our findings we conclude that

$$\begin{aligned}
\mathcal{B} &:= \begin{bmatrix}\left(\begin{bmatrix}-2\mu\mathrm{div}\varepsilon & \\ & (\tau\kappa)^{-1}I\end{bmatrix} + \begin{bmatrix}-\nabla & \nabla\end{bmatrix}\mathcal{C}^{-1}\begin{bmatrix}\mathrm{div} & \\ & -\mathrm{div}\end{bmatrix}\right)^{-1} & \\ & \mathcal{C}^{-1}\end{bmatrix} \\
&\eqsim \begin{bmatrix}(-2\mu\mathrm{div}\varepsilon)^{-1} & & \\ & \tau\kappa(I - \nabla\mathrm{div})^{-1} & \\ & & \mathcal{C}^{-1}\end{bmatrix}
\end{aligned}$$

provides a norm-equivalent preconditioner, where

$$\mathcal{C} := \begin{bmatrix}(\lambda^{-1} + (2\mu)^{-1})I & \alpha\lambda^{-1}I \\ \alpha\lambda^{-1}I & (c_0 + \alpha^2\lambda^{-1} + \tau\kappa)I\end{bmatrix} \quad \text{and}$$

$$\mathcal{C}^{-1} = \eta\begin{bmatrix}2\mu(\alpha^2 + \lambda(c_0 + \tau\kappa))I & -2\mu\alpha I \\ -2\mu\alpha I & (\lambda + 2\mu)I\end{bmatrix},$$

with $\eta = 1/(\alpha^2 + (\lambda + 2\mu)(c_0 + \tau\kappa))$.

Example 3.14. Finally, let us consider the classical three-field formulation of Biot's consolidation model as analyzed in [32]. After some rescaling of parameters and semidiscretization in time by the implicit Euler method the static variational problem to be solved in each time step is given by

$$
\begin{aligned}
(\varepsilon(\boldsymbol{u}),\varepsilon(\boldsymbol{v}))+\lambda_\mu(\mathrm{div}\boldsymbol{u},\mathrm{div}\boldsymbol{v})-(p,\mathrm{div}\boldsymbol{v})&=(\boldsymbol{f},\boldsymbol{v}), \qquad \forall \boldsymbol{v}\in\boldsymbol{H}^1_0(\Omega),\\
R_p^{-1}(\boldsymbol{w},\boldsymbol{z})-(p,\mathrm{div}\boldsymbol{z})&=0, \qquad \forall \boldsymbol{z}\in\boldsymbol{H}_0(\mathrm{div},\Omega),\\
-(\mathrm{div}\boldsymbol{u},q)-(\mathrm{div}\boldsymbol{w},q)-\alpha_p(p,q)&=(g,q), \qquad \forall q\in L^2_0(\Omega),
\end{aligned}
\tag{3.28}
$$

with parameters $\lambda_\mu=\lambda/(2\mu)$, $R_p^{-1}=\alpha^2\tau^{-1}\kappa^{-1}$, and $\alpha_p=c_0\alpha^{-2}$. In this example, we set $\boldsymbol{V}:=\boldsymbol{H}^1_0(\Omega)\times\boldsymbol{H}_0(\mathrm{div},\Omega)$, $Q:=L^2_0(\Omega)$. The individual bilinear forms are defined by

$$
\begin{aligned}
a(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})&=a((\boldsymbol{u};\boldsymbol{w}),(\boldsymbol{v};\boldsymbol{z}))\\
&:=(\varepsilon(\boldsymbol{u}),\varepsilon(\boldsymbol{v}))+\lambda_\mu(\mathrm{div}\boldsymbol{u},\mathrm{div}\boldsymbol{v})+R_p^{-1}(\boldsymbol{w},\boldsymbol{z}), \quad \forall\bar{\boldsymbol{u}},\bar{\boldsymbol{v}}\in\boldsymbol{V},\\
b(\bar{\boldsymbol{v}},q)&=b((\boldsymbol{v};\boldsymbol{z}),q):=-(\mathrm{div}\boldsymbol{v}+\mathrm{div}\boldsymbol{z},q)=:-(\mathrm{Div}\bar{\boldsymbol{v}},q), \quad \forall\bar{\boldsymbol{v}}\in\boldsymbol{V},\forall q\in Q,\\
c(p,q)&:=\alpha_p(p,q), \quad \forall p,q\in Q,
\end{aligned}
$$

and $B:\boldsymbol{V}\to Q'$, $B=-\mathrm{Div}$ where $\mathrm{Div}\bar{\boldsymbol{v}}=\mathrm{Div}(\boldsymbol{v};\boldsymbol{z}):=\mathrm{div}\boldsymbol{v}+\mathrm{div}\boldsymbol{z}$.

The fitted norms we use in Example 3.14 are determined by

$$
\begin{aligned}
|q|_Q^2&:=\left(R_p+\frac{1}{1+\lambda_\mu}\right)(q,q), \qquad \forall q\in Q,\\
|\bar{\boldsymbol{v}}|_V^2&:=a(\bar{\boldsymbol{v}},\bar{\boldsymbol{v}}), \qquad \forall\bar{\boldsymbol{v}}\in\boldsymbol{V}.
\end{aligned}
$$

The coercivity estimate (2.32) is again trivially fulfilled with $\underline{C}_a=1$.

Next, from the inf-sup conditions (3.1) and (3.2) we infer that for any $q\in Q$ there exist $\boldsymbol{v}_0$ and $\boldsymbol{z}_0$ such that $\bar{\boldsymbol{v}}_0:=(\boldsymbol{v}_0;\boldsymbol{z}_0)\in\boldsymbol{V}$ such that

$$
-\mathrm{div}\boldsymbol{z}_0=R_pq \quad \text{and} \quad \|\boldsymbol{z}_0\|^2_{\mathrm{div}}\le\beta_d^{-2}R_p^2\|q\|^2, \tag{3.29a}
$$

$$
-\mathrm{div}\boldsymbol{v}_0=\frac{1}{1+\lambda_\mu}q \quad \text{and} \quad \|\boldsymbol{v}_0\|_1^2\le\beta_s^{-2}\frac{1}{(1+\lambda_\mu)^2}\|q\|^2. \tag{3.29b}
$$

Now, from (3.29a) and (3.29b) we conclude

$$
\begin{aligned}
\beta_d^{-2}R_p\|q\|^2\ge R_p^{-1}\|\boldsymbol{z}_0\|^2_{\mathrm{div}}&=R_p^{-1}\left(\|\mathrm{div}\boldsymbol{z}_0\|^2+\|\boldsymbol{z}_0\|^2\right)\\
&=\|R_p^{-1/2}\mathrm{div}\boldsymbol{z}_0\|^2+\|R_p^{-1/2}\boldsymbol{z}_0\|^2,
\end{aligned}
\tag{3.30}
$$

$$
\begin{aligned}
\beta_s^{-2}\frac{1}{1+\lambda_\mu}\|q\|^2\ge(1+\lambda_\mu)\|\boldsymbol{v}_0\|_1^2&=\frac{1}{2}(1+\lambda_\mu)2\left(\|\nabla\boldsymbol{v}_0\|^2+\|\boldsymbol{v}_0\|^2\right)\\
&\ge\frac{1}{2}\left[(\varepsilon(\boldsymbol{v}_0),\varepsilon(\boldsymbol{v}_0))+\lambda_\mu(\mathrm{div}\boldsymbol{v}_0,\mathrm{div}\boldsymbol{v}_0)+(1+\lambda_\mu)(\mathrm{div}\boldsymbol{v}_0,\mathrm{div}\boldsymbol{v}_0)\right].
\end{aligned}
\tag{3.31}
$$

Choosing $\underline{\beta} := \min\{\beta_d, \frac{\beta_s}{\sqrt{2}}\}$, we obtain

$$\begin{aligned}
\underline{\beta}^{-2}|q|_Q^2 &= \underline{\beta}^{-2}\left(R_p + \frac{1}{1+\lambda_\mu}\right)\|q\|^2 \\
&\geq (\varepsilon(\boldsymbol{v}_0), \varepsilon(\boldsymbol{v}_0)) + \lambda_\mu(\mathrm{div}\boldsymbol{v}_0, \mathrm{div}\boldsymbol{v}_0) + R_p^{-1}(\boldsymbol{z}_0, \boldsymbol{z}_0) \\
&+ R_p^{-1}(\mathrm{div}\boldsymbol{z}_0, \mathrm{div}\boldsymbol{z}_0) + \left(\frac{1}{1+\lambda_\mu}\right)^{-1}(\mathrm{div}\boldsymbol{v}_0, \mathrm{div}\boldsymbol{v}_0) \\
&\geq |\bar{\boldsymbol{v}}_0|_V^2 + \frac{1}{2}(R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)^{-1}(\mathrm{Div}\bar{\boldsymbol{v}}_0, \mathrm{Div}\bar{\boldsymbol{v}}_0) \geq \frac{1}{2}\|\bar{\boldsymbol{v}}_0\|_V^2. \qquad (3.32)
\end{aligned}$$

Hence, in Example 3.13 (2.33) follows by combining (3.29a), (3.29b) and (3.32).

Remark 3.15. Noting that for any $\epsilon \in (0,1)$, by Cauchy inequality, we have

$$\begin{aligned}
\|\bar{\boldsymbol{v}}\|_V^2 &= |\bar{\boldsymbol{v}}|_V^2 + \langle B\bar{\boldsymbol{v}}, \bar{Q}^{-1}B\bar{\boldsymbol{v}}\rangle_{Q'\times Q} = a(\bar{\boldsymbol{v}}, \bar{\boldsymbol{v}}) + (R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)^{-1}(\mathrm{Div}\bar{\boldsymbol{v}}, \mathrm{Div}\bar{\boldsymbol{v}}) \\
&= (\varepsilon(\boldsymbol{v}), \varepsilon(\boldsymbol{v})) + \lambda_\mu(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + R_p^{-1}(\boldsymbol{z}, \boldsymbol{z}) \\
&+ (R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)^{-1}(\mathrm{div}\boldsymbol{v} + \mathrm{div}\boldsymbol{z}, \mathrm{div}\boldsymbol{v} + \mathrm{div}\boldsymbol{z}) \\
&\geq (\varepsilon(\boldsymbol{v}), \varepsilon(\boldsymbol{v})) + \lambda_\mu(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + R_p^{-1}(\boldsymbol{z}, \boldsymbol{z}) \\
&+ (\frac{1}{1+\lambda_\mu})^{-1}(1-\epsilon^{-1})(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + (R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)^{-1}(1-\epsilon)(\mathrm{div}\boldsymbol{z}, \mathrm{div}\boldsymbol{z}) \\
&\geq \frac{1}{2}(\varepsilon(\boldsymbol{v}), \varepsilon(\boldsymbol{v})) + \frac{1}{2}\lambda_\mu(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + R_p^{-1}(\boldsymbol{z}, \boldsymbol{z}) \\
&+ (1+\lambda_\mu)(\frac{3}{2} - \epsilon^{-1})(\mathrm{div}\boldsymbol{v}, \mathrm{div}\boldsymbol{v}) + (R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)^{-1}(1-\epsilon)(\mathrm{div}\boldsymbol{z}, \mathrm{div}\boldsymbol{z}).
\end{aligned}$$

Taking $\epsilon = \frac{2}{3}$, we have

$$\|\bar{\boldsymbol{v}}\|_V^2 \geq \frac{1}{3}\left(\|\varepsilon(\boldsymbol{v})\|^2 + \lambda_\mu\|\mathrm{div}\boldsymbol{v}\|^2 + R_p^{-1}\|\boldsymbol{z}\|^2 + (R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)^{-1}\|\mathrm{div}\boldsymbol{z}\|^2\right),$$

from which we conclude the uniform stability result for classical three-field formulation of Biot's consolidation model presented in [32].

A norm-equivalent preconditioner in this final example is, therefore, given by

$$\mathcal{B} := \begin{bmatrix} -(\mathrm{div}\varepsilon + \lambda_\mu\nabla\mathrm{div})^{-1} & & \\ & (R_p^{-1}I - \nabla(R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)^{-1}\mathrm{div})^{-1} & \\ & & ((R_p + \frac{1}{1+\lambda_\mu} + \alpha_p)I)^{-1} \end{bmatrix}.$$

4. Conclusion

In this paper we have presented a new framework for the stability analysis of perturbed saddle-point problems in variational formulation in a Hilbert space setting. Our approach is constructive and is based on a specific norm fitting technique. The main theoretical result (Theorem 2.15) is a generalization of the classical splitting theorem (Theorem 2.5) of Brezzi and allows to conclude the necessary stability condition (big inf-sup condition) according to Babuška's theory (Theorem 2.1) from conditions similar to those on which Brezzi's theorem is based also in presence of a symmetric positive semidefinite perturbation term $c(\cdot,\cdot) \not\equiv 0$.

As demonstrated on mixed formulations of generalized Poisson, Stokes, vector Laplacian, and Biot's equations, the new norm fitting technique guides the process of defining proper parameter-dependent norms and allows for simple and short proofs of the stability of perturbed saddle-point problems. Although the examples in the present paper are continuous (infinite-dimensional) models, the abstract framework suggests that the proposed technique is also quite useful when studying the stability of various mixed (finite element) discretizations and has a wide range of applications.

References

[1] J. H. Adler, F. J. Gaspar, X. Hu, C. Rodrigo, and L. T. Zikatanov, *Robust block preconditioners for Biot's model*, Domain Decomposition Methods in Science and Engineering XXIV, Lect. Notes Comput. Sci. Eng., vol. 125, Springer, Cham, 2018, pp. 3–16, DOI 10.1007/978-3-319-93873-8_1. MR3989852

[2] D. N. Arnold, R. S. Falk, and R. Winther, *Preconditioning in $H(\mathrm{div})$ and applications*, Math. Comp. **66** (1997), no. 219, 957–984, DOI 10.1090/S0025-5718-97-00826-0. MR1401938

[3] D. N. Arnold, R. S. Falk, and R. Winther, *Finite element exterior calculus, homological techniques, and applications*, Acta Numer. **15** (2006), 1–155, DOI 10.1017/S0962492906210018. MR2269741

[4] D. N. Arnold and R. Winther, *Mixed finite elements for elasticity*, Numer. Math. **92** (2002), no. 3, 401–419, DOI 10.1007/s002110100348. MR1930384

[5] I. Babuška, *Error-bounds for finite element method*, Numer. Math. **16** (1970/71), 322–333, DOI 10.1007/BF02165003. MR288971

[6] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, The Mathematical Foundations of the Finite Element Method With Applications to Partial Differential Equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972), Academic Press, New York, 1972, pp. 1–359. With the collaboration of G. Fix and R. B. Kellogg. MR0421106

[7] M. Bai, D. Elsworth, and J.-C. Roegiers, *Multiporosity/multipermeability approach to the simulation of naturally fractured reservoirs*, Water Resources Res. **29** (1993), no. 6, 1621–1633.

[8] A. Battermann and M. Heinkenschloss, *Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems*, Control and Estimation of Distributed Parameter Systems (Vorau, 1996), Internat. Ser. Numer. Math., vol. 126, Birkhäuser, Basel, 1998, pp. 15–32. MR1627643

[9] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer. **14** (2005), 1–137, DOI 10.1017/S0962492904000212. MR2168342

[10] M. A. Biot, *General theory of three-dimensional consolidation*, J. Appl. Phys. **12** (1941), no. 2, 155–164.

[11] M. A. Biot, *Theory of elasticity and consolidation for a porous anisotropic solid*, J. Appl. Phys. **26** (1955), 182–185. MR66874

[12] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics, vol. 44, Springer, Heidelberg, 2013, DOI 10.1007/978-3-642-36519-5. MR3097958

[13] W. M. Boon, M. Kuchta, K.-A. Mardal, and R. Ruiz-Baier, *Robust preconditioners for perturbed saddle-point problems and conservative discretizations of Biot's equations utilizing total pressure*, SIAM J. Sci. Comput. **43** (2021), no. 4, B961–B983, DOI 10.1137/20M1379708. MR4295052

[14] D. Braess, *Stability of saddle point problems with penalty* (English, with English and French summaries), RAIRO Modél. Math. Anal. Numér. **30** (1996), no. 6, 731–742, DOI 10.1051/m2an/1996300607311. MR1419936

[15] D. Braess, *Finite Elements*, 3rd ed., Cambridge University Press, Cambridge, 2007. Theory, fast solvers, and applications in elasticity theory; Translated from the German by Larry L. Schumaker, DOI 10.1017/CBO9780511618635. MR2322235

[16] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers* (English, with French summary), Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge **8** (1974), no. R-2, 129–151. MR365287

[17] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Series in Computational Mathematics, vol. 15, Springer-Verlag, New York, 1991, DOI 10.1007/978-1-4612-3172-1. MR1115205

[18] M. Burger and W. Mühlhuber, *Iterative regularization of parameter identification problems by sequential quadratic programming methods*, Inverse Problems **18** (2002), no. 4, 943–969, DOI 10.1088/0266-5611/18/4/301. MR1929275

[19] S. Chen, Q. Hong, J. Xu, and K. Yang, *Robust block preconditioners for poroelasticity*, Comput. Methods Appl. Mech. Engrg. **369** (2020), 113229, 21, DOI 10.1016/j.cma.2020.113229. MR4120363

[20] O. Coussy, *Poromechanics*, John Wiley & Sons, West Sussex, England, 2004.

[21] L. Demkowicz, *Babuška ⇔ Brezzi??*, Technical Report 08-06, ICE, The University of Texas at Austin, Texas Institute for Computational and Applied Mathematics, 2006.

[22] M. Discacciati, E. Miglio, and A. Quarteroni, *Mathematical and numerical models for coupling surface and groundwater flows*, Appl. Numer. Math. **43** (2002), no. 1-2, 57–74, DOI 10.1016/S0168-9274(02)00125-3. 19th Dundee Biennial Conference on Numerical Analysis (2001). MR1936102

[23] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005. MR2155549

[24] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Applied Mathematical Sciences, vol. 159, Springer-Verlag, New York, 2004, DOI 10.1007/978-1-4757-4355-5. MR2050138

[25] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London-New York, 1981. MR634376

[26] V. Girault and P.-A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Mathematics, vol. 749, Springer-Verlag, Berlin-New York, 1979. MR548867

[27] R. Glowinski, *Finite Element Methods for Incompressible Viscous Flow*, Handbook of numerical analysis, Vol. IX, Handb. Numer. Anal., IX, North-Holland, Amsterdam, 2003, pp. 3–1176. MR2009826

[28] L. Guo, Z. Li, J. Lyu, Y. Mei, J. Vardakis, D. Chen, C. Han, X. Lou, and Y. Ventikos, *On the validation of a multiple-network poroelastic model using arterial spin labeling MRI data*, Front. Comput. Neurosci. **13** (2019), PMID 31551742, PMCID PMC6733888.

[29] E. L. Hall, *Computer Image Processing and Recognition*, Computer Science and Applied Mathematics, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London-Toronto, Ont., 1979. MR553438

[30] R. Hiptmair and J. Xu, *Nodal auxiliary space preconditioning in* $\mathbf{H}(\mathbf{curl})$ *and* $\mathbf{H}(\mathrm{div})$ *spaces*, SIAM J. Numer. Anal. **45** (2007), no. 6, 2483–2509, DOI 10.1137/060660588. MR2361899

[31] Q. Hong and J. Kraus, *Uniformly stable discontinuous Galerkin discretization and robust iterative solution methods for the Brinkman problem*, SIAM J. Numer. Anal. **54** (2016), no. 5, 2750–2774, DOI 10.1137/14099810X. MR3544656

[32] Q. Hong and J. Kraus, *Parameter-robust stability of classical three-field formulation of Biot's consolidation model*, Electron. Trans. Numer. Anal. **48** (2018), 202–226, DOI 10.1553/etna_vol48s202. MR3820123

[33] Q. Hong, J. Kraus, M. Lymbery, and F. Philo, *Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models*, Numer. Linear Algebra Appl. **26** (2019), no. 4, e2242, 25, DOI 10.1002/nla.2242. MR3979954

[34] Q. Hong, J. Kraus, M. Lymbery, and F. Philo, *Parameter-robust Uzawa-type iterative methods for double saddle point problems arising in Biot's consolidation and multiple-network poroelasticity models*, Math. Models Methods Appl. Sci. **30** (2020), no. 13, 2523–2555, DOI 10.1142/S0218202520500499. MR4192882

[35] Q. Hong, J. Kraus, M. Lymbery, and M. F. Wheeler, *Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems*, Multiscale Model. Simul. **18** (2020), no. 2, 916–941, DOI 10.1137/19M1253988. MR4102708

[36] Q. Hong, J. Kraus, J. Xu, and L. Zikatanov, *A robust multigrid method for discontinuous Galerkin discretizations of Stokes and linear elasticity equations* (English, with Romanian summary), Numer. Math. **132** (2016), no. 1, 23–49, DOI 10.1007/s00211-015-0712-y. MR3439214

[37] Q. Hong, Y. Li, and J. Xu, *An extended Galerkin analysis in finite element exterior calculus*, Math. Comp. **91** (2022), no. 335, 1077–1106, DOI 10.1090/mcom/3707. MR4405489

[38] J. Kraus, R. Lazarov, M. Lymbery, S. Margenov, and L. Zikatanov, *Preconditioning heterogeneous $H(\mathrm{div})$ problems by additive Schur complement approximation and applications*, SIAM J. Sci. Comput. **38** (2016), no. 2, A875–A898, DOI 10.1137/140974092. MR3474851

[39] S. Kumar, R. Oyarzúa, R. Ruiz-Baier, and R. Sandilya, *Conservative discontinuous finite volume and mixed schemes for a new four-field formulation in poroelasticity*, ESAIM Math. Model. Numer. Anal. **54** (2020), no. 1, 273–299, DOI 10.1051/m2an/2019063. MR4058208

[40] Y. K. Kwok and W. Zheng, *Saddlepoint Approximation Methods in Financial Engineering*, SpringerBriefs in Quantitative Finance, Springer, Cham, 2018, DOI 10.1007/978-3-319-74101-7. MR3753567

[41] O. A. Ladyženskaja and V. A. Solonnikov, *Some problems of vector analysis, and generalized formulations of boundary value problems for the Navier-Stokes equation* (Russian, with English summary), Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **59** (1976), 81–116, 256. Boundary value problems of mathematical physics and related questions in the theory of functions, 9. MR0467031

[42] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Mathematics and its Applications, Vol. 2, Gordon and Breach Science Publishers, New York-London-Paris, 1969. Second English edition, revised and enlarged; Translated from the Russian by Richard A. Silverman and John Chu. MR0254401

[43] J. J. Lee, K.-A. Mardal, and R. Winther, *Parameter-robust discretization and preconditioning of Biot's consolidation model*, SIAM J. Sci. Comput. **39** (2017), no. 1, A1–A24, DOI 10.1137/15M1029473. MR3590654

[44] J. J. Lee, *Robust error analysis of coupled mixed methods for Biot's consolidation model*, J. Sci. Comput. **69** (2016), no. 2, 610–632, DOI 10.1007/s10915-016-0210-0. MR3551338

[45] J. J. Lee, E. Piersanti, K.-A. Mardal, and M. E. Rognes, *A mixed finite element method for nearly incompressible multiple-network poroelasticity*, SIAM J. Sci. Comput. **41** (2019), no. 2, A722–A747, DOI 10.1137/18M1182395. MR3922236

[46] D. Loghin and A. J. Wathen, *Analysis of preconditioners for saddle-point problems*, SIAM J. Sci. Comput. **25** (2004), no. 6, 2029–2049, DOI 10.1137/S1064827502418203. MR2086829

[47] K.-A. Mardal and R. Winther, *Preconditioning discretizations of systems of partial differential equations*, Numer. Linear Algebra Appl. **18** (2011), no. 1, 1–40, DOI 10.1002/nla.716. MR2769031

[48] P. Monk, *Finite Element Methods for Maxwell's Equations*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2003, DOI 10.1093/acprof:oso/9780198508885.001.0001. MR2059447

[49] M. A. Murad and A. F. D. Loula, *Improved accuracy in finite element analysis of Biot's consolidation problem*, Comput. Methods Appl. Mech. Engrg. **95** (1992), no. 3, 359–382, DOI 10.1016/0045-7825(92)90193-N. MR1156589

[50] M. A. Murad and A. F. D. Loula, *On stability and convergence of finite element approximations of Biot's consolidation problem*, Internat. J. Numer. Methods Engrg. **37** (1994), no. 4, 645–667, DOI 10.1002/nme.1620370407. MR1257948

[51] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie, Éditeurs, Paris; Academia, Éditeurs, Prague, 1967.

[52] R. A. Nicolaides, *Existence, uniqueness and approximation for generalized saddle point problems*, SIAM J. Numer. Anal. **19** (1982), no. 2, 349–357, DOI 10.1137/0719021. MR650055

[53] R. Oyarzúa and R. Ruiz-Baier, *Locking-free finite element methods for poroelasticity*, SIAM J. Numer. Anal. **54** (2016), no. 5, 2951–2973, DOI 10.1137/15M1050082. MR3552204

[54] P. J. Phillips and M. F. Wheeler, *A coupling of mixed and continuous Galerkin finite element methods for poroelasticity. I. The continuous in time case*, Comput. Geosci. **11** (2007), no. 2, 131–144, DOI 10.1007/s10596-007-9045-y. MR2327964

[55] P. J. Phillips and M. F. Wheeler, *A coupling of mixed and continuous Galerkin finite element methods for poroelasticity. II. The discrete-in-time case*, Comput. Geosci. **11** (2007), no. 2, 145–158, DOI 10.1007/s10596-007-9044-z. MR2327966

[56] W. Schilders, *Introduction to Model Order Reduction*, Model order reduction: theory, research aspects and applications, Math. Ind., vol. 13, Springer, Berlin, 2008, pp. 3–32, DOI 10.1007/978-3-540-78841-6_1. MR2497742
[57] R. Temam, *Navier-Stokes Equations*, 3rd ed., Studies in Mathematics and its Applications, vol. 2, North-Holland Publishing Co., Amsterdam, 1984. Theory and numerical analysis; With an appendix by F. Thomasset. MR769654
[58] K. Terzaghi, *Erdbaumechanik auf bodenphysikalischer Grundlage*, F. Deuticke, 1925.
[59] A. Toselli and O. Widlund, *Domain Decomposition Methods—Algorithms and Theory*, Springer Series in Computational Mathematics, vol. 34, Springer-Verlag, Berlin, 2005, DOI 10.1007/b137868. MR2104179
[60] B. Tully and Y. Ventikos, *Cerebral water transport using multiple-network poroelastic theory: application to normal pressure hydrocephalus*, J. Fluid Mech. **667** (2011), 188–215, DOI 10.1017/S0022112010004428. MR2754490
[61] P. S. Vassilevski and R. D. Lazarov, *Preconditioning mixed finite element saddle-point elliptic problems*, Numer. Linear Algebra Appl. **3** (1996), no. 1, 1–20, DOI 10.1002/(SICI)1099-1506(199601/02)3:1⟨1::AID-NLA67⟩3.3.CO;2-5. MR1373366
[62] S.-Y. Yi, *Convergence analysis of a new mixed finite element method for Biot's consolidation model*, Numer. Methods Partial Differential Equations **30** (2014), no. 4, 1189–1210, DOI 10.1002/num.21865. MR3200272
[63] W. Zulehner, *Nonstandard norms and robust estimates for saddle point problems*, SIAM J. Matrix Anal. Appl. **32** (2011), no. 2, 536–560, DOI 10.1137/100814767. MR2817503

Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802
Email address: huq11@psu.edu

Faculty of Mathematics, University of Duisburg-Essen, Thea-Leymann-Strasse 9, Essen 45127, Germany
Email address: johannes.kraus@uni-due.de

Faculty of Mathematics, University of Duisburg-Essen, Thea-Leymann-Strasse 9, Essen 45127, Germany
Email address: maria.lymbery@uni-due.de

Faculty of Mathematics, University of Duisburg-Essen, Thea-Leymann-Strasse 9, Essen 45127, Germany
Email address: fadi.philo@uni-due.de

ROBUST APPROXIMATION OF GENERALIZED BIOT-BRINKMAN PROBLEMS

Journal of Scientific Computing (2022) 93:77
https://doi.org/10.1007/s10915-022-02029-w

Robust Approximation of Generalized Biot-Brinkman Problems

Qingguo Hong[1] · Johannes Kraus[2] · Miroslav Kuchta[3] · Maria Lymbery[2] · Kent-André Mardal[3,4] · Marie E. Rognes[3]

Received: 12 February 2022 / Revised: 19 September 2022 / Accepted: 10 October 2022 /
Published online: 8 November 2022

Abstract

The generalized Biot-Brinkman equations describe the displacement, pressures and fluxes in an elastic medium permeated by multiple viscous fluid networks and can be used to study complex poromechanical interactions in geophysics, biophysics and other engineering sciences. These equations extend on the Biot and multiple-network poroelasticity equations on the one hand and Brinkman flow models on the other hand, and as such embody a range of singular perturbation problems in realistic parameter regimes. In this paper, we introduce, theoretically analyze and numerically investigate a class of three-field finite element formulations of the generalized Biot-Brinkman equations. By introducing appropriate norms, we demonstrate that the proposed finite element discretization, as well as an associated preconditioning strategy, is robust with respect to the relevant parameter regimes. The theoretical analysis is complemented by numerical examples.

Qingguo Hong
huq11@psu.edu

Johannes Kraus
johannes.kraus@uni-due.de

Miroslav Kuchta
miroslav@simula.no

Maria Lymbery
maria.lymbery@uni-due.de

Kent-André Mardal
kent-and@simula.no

Marie E. Rognes
meg@simula.no

1 Department of Mathematics, Pennsylvania State University, State College, USA

2 Faculty of Mathematics, University of Duisburg-Essen, Essen, Germany

3 Department of Numerical Analysis and Scientific Computing, Simula Research Laboratory, Oslo, Norway

4 Department of Mathematics, University of Oslo, Oslo, Norway

Keyword Poromechanics, Finite element method, Preconditioning, Biot equations, Brinkman approximation, Multiple-network poroelasticity

1 Introduction

The study of the mechanical response of fluid-filled porous media – *poromechanics* – is essential in geophysics, biophysics and civil engineering. Through a series of seminal works dating from 1941 and onwards [7, 8], Biot introduced governing equations for the dynamic behavior of a linearly elastic solid matrix permeated by a viscous fluid with flow through the pore network described by Darcy's law [17, 47]. Double-porosity models, extending upon Biot's single fluid network to the case of two interacting networks, were used to describe the motion of liquids in fissured rocks as early as in the 1960s [5, 32, 48]. Later, multiple-network poroelasticity equations emerged in the context of reservoir modelling [4] to describe elastic media permeated by multiple networks characterised by different porosities, permeabilities and/or interactions. Since the early 2000s, poromechanics has been applied to model the heart [13, 38] as well as the brain and central nervous system [16, 20, 44–46].

Several recent papers [6, 12, 14, 31] have in biomedical applications such as the perfusion of the heart and glymphatic system of the brain, in addition to multiple networks, also accounted for viscous effects of the fluid. At its core, the effect of viscosity can be accounted for by replacing the Darcy approximation in the poroelasticity model by a Brinkman approximation [11, 40]. We here introduce multiple-network poroelasticity models incorporating viscosity under the term *generalized Biot-Brinkman equations*. In a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ comprising n fluid networks, the generalized Biot-Brinkman equations read as follows: find the displacement $\boldsymbol{u} = \boldsymbol{u}(x,t)$, fluid fluxes $\boldsymbol{v}_i = \boldsymbol{v}_i(x,t)$ and corresponding (negative) fluid pressures $p_i = p_i(x,t)$, for $i = 1, \ldots, n$ satisfying

$$-\operatorname{div}(\boldsymbol{\sigma}(\boldsymbol{u}) - \boldsymbol{\alpha} \cdot \boldsymbol{p}\boldsymbol{I}) = \boldsymbol{f}, \tag{1.1a}$$

$$-\nu_i \operatorname{div} \boldsymbol{\varepsilon}(\boldsymbol{v}_i) + \boldsymbol{v}_i + K_i \nabla p_i = \boldsymbol{r}_i, \tag{1.1b}$$

$$-c_i \dot{p}_i - \bar{\beta}_i p_i - \alpha_i \operatorname{div} \dot{\boldsymbol{u}} - \operatorname{div} \boldsymbol{v}_i + \boldsymbol{\beta}_i \cdot \boldsymbol{p} = g_i, \tag{1.1c}$$

over $\Omega \times (0, T)$ for $T > 0$, and where (1.1b) and (1.1c) hold for $i = 1, \ldots, n$. In (1.1a), we have introduced the vector notation $\boldsymbol{p} = (p_1, \ldots, p_n)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, where α_i is the Biot-Willis coefficient associated with network i. The elastic stress and strain tensors are:

$$\boldsymbol{\sigma}(\boldsymbol{u}) = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \operatorname{div}(\boldsymbol{u})\boldsymbol{I}, \quad \boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T), \tag{1.2}$$

respectively, and with Lamé parameters μ and λ. Moreover, for each fluid network i, ν_i denotes the fluid viscosity and K_i is its hydraulic conductance tensor. Furthermore, (1.1c) is an equivalent formulation of the standard multiple-network poroelasticity mass balance equations [4, 24, 36] with transfer coefficients β_{ij}, denoting $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{in})$ and $\bar{\beta}_i = \sum_j \beta_{ij}$, when the fluid transfer into network i is given by

$$\textstyle\sum_{j=1, j\neq i}^{n} \beta_{ij}(p_i - p_j).$$

The constants c_i in (1.1c) denote the constrained specific storage coefficients, see e.g. [43] and the references therein. Finally, the prescribed right hand side $\boldsymbol{f}$ denotes body forces, while g_i denotes a fluid source and $\boldsymbol{r}_i$ represents an external flux, both of the two latter in each network i. In the case $n = 1$ and $\nu = 0$, (1.1) reduces to the Biot equations.

The generalized Biot-Brinkman problem (1.1) defines a challenging system of PDEs to solve numerically. One reason for this is the large number of material parameters, several of which give rise to singular perturbation problems such as in the extreme cases of (near) incompressibility ($\lambda \to \infty$) and impermeability ($K_i \to 0$). Specifically, $\lambda \gg \mu$ is associated with numerical locking; if (1.1) is scaled by $1/\lambda$, the elastic term of the equation reads $\operatorname{div} \frac{2\mu}{\lambda}\boldsymbol{\varepsilon}(\boldsymbol{u}) + \nabla \operatorname{div} \boldsymbol{u} = f$ which transforms from an H^1 problem to an $H(\operatorname{div})$ problem as λ tends to infinity. Similar singular perturbation problems arise, now for the flux variable $\boldsymbol{v}_i$, as ν_i tends to zero. Furthermore, certain parameter ranges of the storage coefficients and permeabilities ($K_i \ll c_i$) give rise to singular perturbation problems in the Darcy sub-system, see e.g. [37] and references therein. Finally, we mention that large transfer coefficients β_{ij} and/or small Biot-Willis coefficients α_i can lead to strong coupling of the different subsystems and prevent direct exploitation of each subsystem's properties.

In the case of vanishing viscosities ($\nu_i = 0, \forall i$) the system (refeq:mBB:t)reduces to the multiple-network poroelasticity (MPET) equations. Robust and conservative numerical approximations of the MPET equations have been studied in the context of (near) incompressibility [36] as well as other material parameters [24, 26, 27]. Parameter-independent preconditioning and splitting schemes as well as a-posteriori error analysis and adaptivity have also been identified for the MPET equations [18, 25, 27, 39]. However, the generalized Biot-Brinkman system has received little attention from the numerical community. Therefore, the purpose of this paper is to identify and analyze stable finite element approximation schemes and preconditioning techniques for the time-discrete generalized Biot-Brinkman systems, with particular focus on parameter robustness.

This paper is organized as follows. After introducing notation, context and preliminaries in Sect. 2, we prove that the time-discrete generalized Biot-Brinkman system is well-posed in appropriate function spaces in Sect. 3. We introduce a fully discrete generalized Biot-Brinkman problem in Sect. 4 and prove that the discrete approximations satisfy a near optimal a-priori error estimate in appropriate norms independently of material parameters. We also propose a natural preconditioner. The theoretical analysis is complemented by numerical experiments in Sect. 5.

2 Preliminaries and Notation

In this section of preliminaries, we give assumptions on the material parameters, present a rescaling of a time-discrete generalized Biot-Brinkman system and introduce parameter-weighted norms and function spaces.

2.1 Material Parameters

We assume that the elastic Lamé coefficients satisfy the standard conditions $\mu > 0$ and $d\lambda + 2\mu > 0$. The transfer coefficients are such that $\beta_{ij} = \beta_{ji} \geq 0$ for $i \neq j$ while $\beta_{ii} = 0$, and the specific storage coefficients $c_i \geq 0$ for $i = 1, \ldots, n$. The Biot-Willis coefficients are bounded between zero and one by construction: $0 < \alpha_i \leq 1$. We also assume that the hydraulic conductances $K_i > 0$ for $i = 1, \ldots, n$. Further, our focus will be on the case $\nu_i > 0$. For spatially-varying material parameters, we assume that each of the above conditions holds point-wise and that each parameter field is uniformly bounded from above and below.

2.2 Time Discretization, Rescaling and Structure

Taking an implicit Euler time-discretization of (refeq:mBB:t)with uniform timestep τ, multiplying (1.1c) by τ, rearranging terms and removing the time-dependence from the notation, we obtain the following problem structure to be solved over Ω at each time step: find the unknown displacement $\boldsymbol{u} = \boldsymbol{u}(x)$, fluid fluxes $\boldsymbol{v}_i = \boldsymbol{v}_i(x)$ and corresponding (negative) fluid pressures $p_i = p_i(x)$, for $i = 1, \ldots, n$ satisfying

$$\begin{aligned} -\operatorname{div}(\boldsymbol{\sigma}(\boldsymbol{u}) - \boldsymbol{\alpha} \cdot \boldsymbol{p}\boldsymbol{I}) &= \boldsymbol{f}, \\ -\nu_i \operatorname{div} \boldsymbol{\varepsilon}(\boldsymbol{v}_i) + \boldsymbol{v}_i + K_i \nabla p_i &= \boldsymbol{r}_i, \\ -\left(c_i + \tau\bar{\beta}_i\right) p_i - \alpha_i \operatorname{div} \boldsymbol{u} - \tau \operatorname{div} \boldsymbol{v}_i + \tau\boldsymbol{\beta}_i \cdot \boldsymbol{p} &= \tau g_i. \end{aligned}$$

Multiplying by τK_i^{-1} in the second equation(s) for the sake of symmetry gives

$$-\operatorname{div}(\boldsymbol{\sigma}(\boldsymbol{u}) - \boldsymbol{\alpha} \cdot \boldsymbol{p}\boldsymbol{I}) = \boldsymbol{f}, \tag{2.2a}$$

$$-\nu_i \tau K_i^{-1} \operatorname{div} \boldsymbol{\varepsilon}(\boldsymbol{v}_i) + \tau K_i^{-1} \boldsymbol{v}_i + \tau \nabla p_i = \tau K_i^{-1} \boldsymbol{r}_i, \tag{2.2b}$$

$$-\left(c_i + \tau\bar{\beta}_i\right) p_i - \alpha_i \operatorname{div} \boldsymbol{u} - \tau \operatorname{div} \boldsymbol{v}_i + \tau\boldsymbol{\beta}_i \cdot \boldsymbol{p} = \tau g_i. \tag{2.2c}$$

For the sake of readability, we define

$$s_i = c_i + \tau\bar{\beta}_i, \quad \gamma_i = \tau\nu_i K_i^{-1}, \tag{2.3}$$

recalling that $\bar{\beta}_i = \sum_j \beta_{ij}$ and $\beta_{ii} = 0$, and set

$$R^{-1} = \max\{(1+\nu_1)\tau K_1^{-1}, \ldots, (1+\nu_n)\tau K_n^{-1}\}. \tag{2.4}$$

Using this notation, we introduce four $n \times n$ parameter matrices

$$\Lambda_1 = -\tau \begin{pmatrix} 0 & \beta_{12} & \ldots & \beta_{1n} \\ \beta_{21} & 0 & \ldots & \beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n1} & \beta_{n2} & \ldots & 0 \end{pmatrix}, \tag{2.5}$$

and

$$\Lambda_2 = \operatorname{diag}(s_1, s_2, \ldots, s_n), \quad \Lambda_3 = \tau^2 R I, \quad \Lambda_4 = \frac{1}{2\mu + \lambda} \boldsymbol{\alpha}\boldsymbol{\alpha}^T, \tag{2.6}$$

before defining

$$\Lambda = \sum_{i=1}^{4} \Lambda_i. \tag{2.7}$$

In the case $n = 1$, dropping the subscripts i, j for readability and with the newly introduced parameter notation, the operator structure of the rescaled system (2.2) is

$$\begin{pmatrix} -\operatorname{div}\boldsymbol{\sigma} & \boldsymbol{0} & \alpha\nabla \\ \boldsymbol{0} & -\gamma \operatorname{div}\boldsymbol{\varepsilon} + \tau K^{-1}\boldsymbol{I} & \tau\nabla \\ -\alpha \operatorname{div} & -\tau \operatorname{div} & -(\Lambda_1 + \Lambda_2) \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{r} \\ \boldsymbol{g} \end{pmatrix}, \tag{2.8}$$

for $(\Lambda_1 + \Lambda_2) = c\boldsymbol{I}$, and $\boldsymbol{p} = p$ in the $n = 1$ case. The same structure holds for $n > 2$ when denoting $\boldsymbol{v}^T = (\boldsymbol{v}_1^T, \boldsymbol{v}_2^T, \ldots, \boldsymbol{v}_n^T)$, $(\operatorname{Div}\boldsymbol{v})^T = (\operatorname{div}\boldsymbol{v}_1, \ldots, \operatorname{div}\boldsymbol{v}_n)$.

By the assumption of symmetric transfer, i.e. $\beta_{ij} = \beta_{ji}$, Λ_1 and Λ are symmetric. Moreover, as $\Lambda_1 + \Lambda_2$ is weakly diagonally dominant and thus symmetric positive semi-definite, Λ_3 is symmetric positive definite, and Λ_4 is symmetric positive semi-definite, it follows that Λ is symmetric positive definite.

2.3 Domain and Boundary Conditions

Assume that Ω is open and bounded in $\mathbb{R}^d$, $d = 2, 3$ with Lipschitz boundary $\partial\Omega$. We consider the following idealized boundary conditions for the theoretical analysis of the time-discrete generalized Biot-Brinkman system (2.8) over Ω. We assume that the displacement is prescribed (and equal to zero for simplicity) on the entire boundary $\partial\Omega$. Furthermore for each of the flux momentum equations we assume datum on the normal flux $\boldsymbol{v}_i \cdot \boldsymbol{n}$ *and* the tangential part of the traction associated with the viscous term $\boldsymbol{\varepsilon}(\boldsymbol{v}_i) \cdot \boldsymbol{n}$. Combined, we thus set

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{x}) &= \boldsymbol{0} \quad \boldsymbol{x} \in \partial\Omega, \\ \boldsymbol{v}_i \cdot \boldsymbol{n}(\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{n} \times (\boldsymbol{\varepsilon}(\boldsymbol{v}_i) \cdot \boldsymbol{n})\,(\boldsymbol{x}) &= \boldsymbol{0} \quad \boldsymbol{x} \in \partial\Omega, \end{aligned} \tag{2.9}$$

for $i = 1, \ldots, n$.

2.4 Function Spaces and Norms

We use standard notation for the Sobolev spaces $L^2(\Omega)$, $H^1(\Omega)$ and $H(\mathrm{div}, \Omega)$, and denote the $L^2(\Omega)$-inner product and norm by $(\cdot, \cdot)$ and $\|\cdot\|$, respectively. We let $L^2_0(\Omega)$ denote the space of L^2 functions with zero mean. For a Banach space U, its dual space is denoted U' and the duality pairing between U and U' by $\langle \cdot, \cdot \rangle_{U' \times U}$.

For the displacement, flux and pressure spaces, we define

$$\boldsymbol{U} = \{\boldsymbol{u} \in H^1(\Omega)^d : \boldsymbol{u} = \boldsymbol{0} \text{ on } \partial\Omega\}, \tag{2.10a}$$

$$\boldsymbol{V}_i = \{\boldsymbol{v}_i \in H^1(\Omega)^d : \boldsymbol{v}_i \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega\}, \tag{2.10b}$$

$$P_i = L^2_0(\Omega), \tag{2.10c}$$

for $i = 1, \ldots, n$, and subsequently define

$$\boldsymbol{V} = V_1 \times \cdots \times V_n, \quad \boldsymbol{P} = P_1 \times \cdots \times P_n. \tag{2.11}$$

We also equip these spaces with the following parameter-weighted inner products

$$(\boldsymbol{u}, \boldsymbol{w})_{\boldsymbol{U}} = (2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{w})) + \lambda(\mathrm{div}\,\boldsymbol{u}, \mathrm{div}\,\boldsymbol{w}), \tag{2.12a}$$

$$(\boldsymbol{v}, \boldsymbol{z})_{\boldsymbol{V}} = \sum_{i=1}^{n} (\gamma_i \boldsymbol{\varepsilon}(\boldsymbol{v}_i), \boldsymbol{\varepsilon}(\boldsymbol{z}_i)) + (\tau K_i^{-1} \boldsymbol{v}_i, \boldsymbol{z}_i) + (\Lambda^{-1}\tau^2 \mathrm{Div}\boldsymbol{v}, \mathrm{Div}\boldsymbol{z}), \tag{2.12b}$$

$$(\boldsymbol{p}, \boldsymbol{q})_{\boldsymbol{P}} = (\Lambda \boldsymbol{p}, \boldsymbol{q}), \tag{2.12c}$$

and denote the induced norms by $\|\cdot\|_{\boldsymbol{U}}$, $\|\cdot\|_{\boldsymbol{V}}$, and $\|\cdot\|_{\boldsymbol{P}}$, respectively. These are indeed inner products and norms by the assumptions on the material parameters given and in particular the symmetric positive-definiteness of Λ.

3 Well-Posedness of the Biot-Brinkman System

3.1 Abstract Form and Related Results

System (2.8) is a special case of the abstract saddle-point problem

$$\begin{pmatrix} A_1 & 0 & B_1^T \\ 0 & A_2 & B_2^T \\ B_1 & B_2 & -A_3 \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{p} \end{pmatrix}, \tag{3.1}$$

where $A_1 : \boldsymbol{U} \to \boldsymbol{U}'$, $A_2 : \boldsymbol{V} \to \boldsymbol{V}'$, and $A_3 : \boldsymbol{P} \to \boldsymbol{P}'$ are symmetric and positive (semi-)definite, and $B_1 : \boldsymbol{U} \to \boldsymbol{P}'$, $B_2 : \boldsymbol{V} \to \boldsymbol{P}'$ are linear operators. In terms of bilinear forms, we can write (3.1) as

$$a_1(\boldsymbol{u}, \boldsymbol{w}) + b_1(\boldsymbol{w}, \boldsymbol{p}) = (\boldsymbol{f}, \boldsymbol{w}), \tag{3.2a}$$
$$a_2(\boldsymbol{v}, \boldsymbol{z}) + b_2(\boldsymbol{z}, \boldsymbol{p}) = (\boldsymbol{r}, \boldsymbol{z}), \tag{3.2b}$$
$$b_1(\boldsymbol{u}, \boldsymbol{q}) + b_2(\boldsymbol{v}, \boldsymbol{q}) - a_3(\boldsymbol{p}, \boldsymbol{q}) = (\boldsymbol{g}, \boldsymbol{q}). \tag{3.2c}$$

This abstract form was studied in the context of twofold saddle point problems and equivalence of inf-sup stability conditions by Howell and Walkington [30] for the case where $A_3 = A_2 = 0$.

3.2 Three-Field Variational Formulation of the Biot-Brinkman System

We consider the following variational formulation of the Biot-Brinkman system (2.8) with the boundary conditions given by (2.9): given $\boldsymbol{f}, \boldsymbol{r}, \boldsymbol{g}$, find $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$ such that (3.2) holds with

$$a_1(\boldsymbol{u}, \boldsymbol{w}) = (\sigma(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{w})), \tag{3.3a}$$
$$a_2(\boldsymbol{v}, \boldsymbol{z}) = \textstyle\sum_{i=1}^n (\gamma_i \boldsymbol{\varepsilon}(\boldsymbol{v}_i), \boldsymbol{\varepsilon}(\boldsymbol{z}_i)) + (\tau K_i^{-1} \boldsymbol{v}_i, \boldsymbol{z}_i), \tag{3.3b}$$
$$a_3(\boldsymbol{p}, \boldsymbol{q}) = \textstyle\sum_{i=1}^n (s_i p_i, q_i) - \sum_{i,j=1}^n (\tau \beta_{ij} p_j, q_i), \tag{3.3c}$$
$$b_1(\boldsymbol{w}, \boldsymbol{p}) = -\textstyle\sum_{i=1}^n (\operatorname{div} \boldsymbol{w}, \alpha_i p_i) \equiv -(\operatorname{div} \boldsymbol{w}, \boldsymbol{\alpha} \cdot \boldsymbol{p}), \tag{3.3d}$$
$$b_2(\boldsymbol{v}, \boldsymbol{q}) = -\textstyle\sum_{i=1}^n (\tau \operatorname{div} \boldsymbol{v}_i, \boldsymbol{q}_i), \tag{3.3e}$$

for all $\boldsymbol{w} \in \boldsymbol{U}$, $\boldsymbol{z} \in \boldsymbol{V}$, and $\boldsymbol{q} \in \boldsymbol{P}$. Equivalently, $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$ solves

$$\mathcal{A}((\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}), (\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{q})) = ((\boldsymbol{f}, \boldsymbol{r}, \boldsymbol{g}), (\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{q})), \tag{3.4}$$

for all $(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{q}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$ where

$$\begin{aligned} \mathcal{A}((\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}), (\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{q})) = {} & a_1(\boldsymbol{u}, \boldsymbol{w}) + a_2(\boldsymbol{v}, \boldsymbol{z}) + b_1(\boldsymbol{w}, \boldsymbol{p}) + b_1(\boldsymbol{u}, \boldsymbol{q}) \\ & + b_2(\boldsymbol{z}, \boldsymbol{p}) + b_2(\boldsymbol{w}, \boldsymbol{q}) - a_3(\boldsymbol{p}, \boldsymbol{q}). \end{aligned} \tag{3.5}$$

We refer to (3.2)–(3.3), or also (3.4), as a three-field formulation of the Biot-Brinkman system, with three-field referring to the three groups of fields (displacement, fluxes and pressures).

3.3 Stability Properties

In this section we prove the main theoretical result of this paper, that is, the uniform well-posedness of problem (3.2)–(3.3) under the norms induced by (2.12), as stated in Theorem 3.5.

The proof utilizes the abstract framework for the stability analysis of perturbed saddle-point problems that has recently been presented in [26]. It is performed in two steps. In the first step, we recast the system (3.2)–(3.3) into the following two-by-two (single) perturbed saddle-point problem

$$\begin{aligned}\mathcal{A}((\boldsymbol{u},\boldsymbol{v},\boldsymbol{p}),(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q})) &= \mathcal{A}((\bar{\boldsymbol{u}},\boldsymbol{p}),(\bar{\boldsymbol{w}},\boldsymbol{q}))\\ &= a(\bar{\boldsymbol{u}},\bar{\boldsymbol{w}}) + b(\bar{\boldsymbol{w}},\boldsymbol{p}) + b(\bar{\boldsymbol{u}},\boldsymbol{q}) - c(\boldsymbol{p},\boldsymbol{q}),\end{aligned} \tag{3.6}$$

where $\bar{\boldsymbol{u}} = (\boldsymbol{u},\boldsymbol{v})$, $\bar{\boldsymbol{w}} = (\boldsymbol{w},\boldsymbol{z})$ and

$$\begin{aligned}a(\bar{\boldsymbol{u}},\bar{\boldsymbol{w}}) &= a_1(\boldsymbol{u},\boldsymbol{w}) + a_2(\boldsymbol{v},\boldsymbol{z}),\\ b(\bar{\boldsymbol{w}},\boldsymbol{p}) &= b_1(\boldsymbol{w},\boldsymbol{p}) + b_2(\boldsymbol{z},\boldsymbol{p}),\\ c(\boldsymbol{p},\boldsymbol{q}) &= a_3(\boldsymbol{p},\boldsymbol{q}),\end{aligned}$$

with $a_1(\cdot,\cdot)$, $a_2(\cdot,\cdot)$, $a_3(\cdot,\cdot)$, $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ as defined in (3.3). Then, according to Theorem 5 in [26], for properly chosen seminorms $|\cdot|_{\boldsymbol{Q}}$ and $|\cdot|_{\bar{\boldsymbol{V}}}$, which are specified in Theorem 3.2 below, the uniform well-posedness of this problem is guaranteed under the fitted (full) norms

$$\|\boldsymbol{q}\|_{\boldsymbol{Q}}^2 = |\boldsymbol{q}|_{\boldsymbol{Q}}^2 + c(\boldsymbol{q},\boldsymbol{q}) =: \langle \bar{\boldsymbol{Q}}\boldsymbol{q},\boldsymbol{q}\rangle_{\boldsymbol{Q}'\times\boldsymbol{Q}}, \tag{3.7}$$

$$\|\bar{\boldsymbol{w}}\|_{\bar{\boldsymbol{V}}}^2 = |\bar{\boldsymbol{w}}|_{\bar{\boldsymbol{V}}}^2 + \langle B\bar{\boldsymbol{w}},\bar{\boldsymbol{Q}}^{-1}B\bar{\boldsymbol{w}}\rangle_{\boldsymbol{Q}'\times\boldsymbol{Q}}, \tag{3.8}$$

if the following two conditions are satisfied for positive constants c_a and c_b which are independent of all model parameters:

$$a(\bar{\boldsymbol{v}},\bar{\boldsymbol{v}}) \geq c_a|\bar{\boldsymbol{v}}|_{\bar{\boldsymbol{V}}}^2 \qquad \forall \bar{\boldsymbol{v}} \in \bar{\boldsymbol{V}}, \tag{3.9}$$

$$\sup_{\bar{\boldsymbol{v}}\in\bar{\boldsymbol{V}}} \frac{b(\bar{\boldsymbol{v}},\boldsymbol{q})}{\|\bar{\boldsymbol{v}}\|_{\bar{\boldsymbol{V}}}} \geq c_b|\boldsymbol{q}|_{\boldsymbol{Q}} \qquad \forall \boldsymbol{q} \in \boldsymbol{Q}. \tag{3.10}$$

This means that under the conditions (3.9) and (3.10) the bilinear form in (3.6) satisfies the estimates

$$|\mathcal{A}((\boldsymbol{u},\boldsymbol{v},\boldsymbol{p}),(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q}))| \leq C_b\|(\boldsymbol{u},\boldsymbol{v},\boldsymbol{p})\|_{\bar{\boldsymbol{X}}}\|(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q})\|_{\bar{\boldsymbol{X}}}, \tag{3.11}$$

and

$$\inf_{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{p})\in\boldsymbol{X}} \sup_{(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q})\in\boldsymbol{X}} \frac{\mathcal{A}((\boldsymbol{u},\boldsymbol{v},\boldsymbol{p}),(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q}))}{\|(\boldsymbol{u},\boldsymbol{v},\boldsymbol{p})\|_{\bar{\boldsymbol{X}}}\|(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q})\|_{\bar{\boldsymbol{X}}}} \geq \omega, \tag{3.12}$$

for the combined norm $\|(\cdot,\cdot,\cdot)\|_{\bar{\boldsymbol{X}}}$ defined by

$$\|(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q})\|_{\bar{\boldsymbol{X}}}^2 := \|\boldsymbol{q}\|_{\boldsymbol{Q}}^2 + \|\bar{\boldsymbol{w}}\|_{\bar{\boldsymbol{V}}}^2, \tag{3.13}$$

on the space $\boldsymbol{X} = \boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}$ with constants C_b and ω that do not depend on any of the model parameters.

Before we turn to the proof of estimates (3.11) and (3.12) in Theorem 3.2 below, we recall appropriate inf-sup conditions for the spaces $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{P}$ in Lemma 3.1.

Lemma 3.1 *The following conditions hold with constants $\beta_d > 0$ and $\beta_s > 0$:*

$$\inf_{q\in P_i}\sup_{\boldsymbol{v}\in \boldsymbol{V}_i}\frac{(\mathrm{div}\boldsymbol{v},q)}{\|\boldsymbol{v}\|_1\|q\|}\geq\beta_d,\quad i=1,\ldots,n, \tag{3.14}$$

$$\inf_{(q_1,\ldots,q_n)\in P_1\times\cdots\times P_n}\sup_{\boldsymbol{u}\in\boldsymbol{U}}\frac{\left(\mathrm{div}\boldsymbol{u},\sum_{i=1}^n q_i\right)}{\|\boldsymbol{u}\|_1\left\|\sum_{i=1}^n q_i\right\|}\geq\beta_s. \tag{3.15}$$

Proof See [9, 10].

Theorem 3.2 *Consider problem (3.2)–(3.3) on the space $\boldsymbol{X}=\boldsymbol{U}\times\boldsymbol{V}\times\boldsymbol{P}=\bar{\boldsymbol{V}}\times\boldsymbol{Q}$ and define the combined norm $\|\cdot\|_{\bar{\boldsymbol{X}}}$ via (3.13) where the fitted norms $\|\cdot\|_{\boldsymbol{Q}}$ and $\|\cdot\|_{\bar{\boldsymbol{V}}}$ are defined by (3.7)–(3.8) with seminorms*

$$|\boldsymbol{q}|_{\boldsymbol{Q}}^2=((\Lambda_3+\Lambda_4)\boldsymbol{q},\boldsymbol{q}), \tag{3.16}$$

$$|\bar{\boldsymbol{w}}|_{\bar{\boldsymbol{V}}}^2=a(\bar{\boldsymbol{w}},\bar{\boldsymbol{w}}). \tag{3.17}$$

Then, the continuity and stability estimates (3.11) and (3.12) hold with positive constants C_b and ω that are independent of all model parameters.

Proof To prove statement (3.11), one uses the Cauchy-Schwarz inequality and the definition of the norms.

In order to prove (3.12) we verify the conditions of Theorem 5 in [26], i.e., conditions (3.9) and (3.10). Noting that $|\bar{\boldsymbol{w}}|_{\bar{\boldsymbol{V}}}^2=a(\bar{\boldsymbol{w}},\bar{\boldsymbol{w}})$, we find that condition (3.9) trivially holds with $c_a=1$ so it remains to show (3.10). The bilinear form b is induced by the operator $B:\bar{\boldsymbol{V}}\to\boldsymbol{Q}'$ that is given by

$$B=\begin{pmatrix}-\alpha_1\,\mathrm{div} & -\tau\,\mathrm{div} & 0 & 0 & \ldots & 0\\ -\alpha_2\,\mathrm{div} & 0 & -\tau\,\mathrm{div} & 0 & \ldots & 0\\ -\alpha_3\,\mathrm{div} & 0 & 0 & -\tau\,\mathrm{div} & \ldots & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ -\alpha_n\,\mathrm{div} & 0 & 0 & 0 & \ldots & -\tau\,\mathrm{div}\end{pmatrix}.$$

Thanks to Lemma 3.1, for a given $(\bar{\boldsymbol{u}},\boldsymbol{p})$ we can choose test functions $\bar{\boldsymbol{w}}=(\boldsymbol{w},\boldsymbol{z})$ such that

$$\begin{aligned}-\mathrm{div}\,\boldsymbol{w}&=\frac{1}{2\mu+\lambda}\sum_{i=1}^n\alpha_i p_i,\quad \|\boldsymbol{w}\|_1\leq\beta_s^{-1}\frac{1}{2\mu+\lambda}\|\sum_{i=1}^n\alpha_i p_i\|,\\ -\mathrm{div}\,\boldsymbol{z}_i&=\tau R p_i,\quad \|\boldsymbol{z}_i\|_1\leq\beta_s^{-1}\tau R\|p_i\|,\quad i=1,\ldots,n.\end{aligned}$$

With these choices we find that

$$\begin{aligned}b(\bar{\boldsymbol{w}},\boldsymbol{p})&=-(\mathrm{div}\,\boldsymbol{w},\sum_{i=1}^n\alpha_i p_i)-\sum_{i=1}^n(\tau\,\mathrm{div}\,\boldsymbol{z}_i,p_i)\\ &=\frac{1}{2\mu+\lambda}\left(\sum_{i=1}^n\alpha_i p_i,\sum_{i=1}^n\alpha_i p_i\right)+\sum_{i=1}^n(\tau^2 R p_i,p_i)\\ &=(\Lambda_4\boldsymbol{p},\boldsymbol{p})+(\Lambda_3\boldsymbol{p},\boldsymbol{p})=|\boldsymbol{p}|_{\boldsymbol{Q}}^2.\end{aligned}$$

In view of (3.8) and noting that $\langle B\bar{\boldsymbol{w}}, \bar{Q}^{-1}B\bar{\boldsymbol{w}}\rangle_{Q'\times Q} = (\Lambda^{-1}B\bar{\boldsymbol{w}}, \bar{\boldsymbol{w}})$, we obtain

$$
\begin{aligned}
\|\bar{\boldsymbol{w}}\|_{\bar{V}}^2 &= 2\mu(\boldsymbol{\varepsilon}(\boldsymbol{w}), \boldsymbol{\varepsilon}(\boldsymbol{w})) + \lambda(\operatorname{div}\boldsymbol{w}, \operatorname{div}\boldsymbol{w}) + \sum_{i=1}^{n}\gamma_i(\boldsymbol{\varepsilon}(\boldsymbol{z}_i), \boldsymbol{\varepsilon}(\boldsymbol{z}_i)) \\
&\quad + \sum_{i=1}^{n}(\tau K_i^{-1}\boldsymbol{z}_i, \boldsymbol{z}_i) + (\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) \\
&\le \beta_s^{-2}(2\mu+\lambda)\left(\frac{1}{2\mu+\lambda}\right)^2 \|\sum_{i=1}^{n}\alpha_i p_i\|^2 + \sum_{i=1}^{n}\gamma_i\beta_s^{-2}\tau^2R^2\|p_i\|^2 \\
&\quad + \sum_{i=1}^{n}\tau K_i^{-1}\beta_s^{-2}\tau^2R^2\|p_i\|^2 + (\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) \\
&\le \beta_s^{-2}\frac{1}{2\mu+\lambda}\|\sum_{i=1}^{n}\alpha_i p_i\|^2 + \beta_s^{-2}\sum_{i=1}^{n}(\gamma_i + \tau K_i^{-1})\tau^2R^2\|p_i\|^2 + (\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) \\
&\le \beta_s^{-2}\frac{1}{2\mu+\lambda}\|\sum_{i=1}^{n}\alpha_i p_i\|^2 + \beta_s^{-2}\sum_{i=1}^{n}\tau^2R\|p_i\|^2 + (\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) \\
&\le \beta_s^{-2}\left((\Lambda_4\boldsymbol{p}, \boldsymbol{p}) + (\Lambda_3\boldsymbol{p}, \boldsymbol{p})\right) + ((\Lambda_3+\Lambda_4)^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) \\
&\le (\beta_s^{-2}+1)|\boldsymbol{p}|_{\boldsymbol{Q}}^2,
\end{aligned}
$$

where we have also used $(\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) \le ((\Lambda_3+\Lambda_4)^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}})$ and $B\bar{\boldsymbol{w}} = (\Lambda_3+\Lambda_4)\boldsymbol{p}$. Finally, (3.10) follows from

$$
\sup_{\bar{\boldsymbol{v}}\in\bar{\boldsymbol{V}}}\frac{b(\bar{\boldsymbol{v}}, \boldsymbol{q})}{\|\bar{\boldsymbol{v}}\|_{\bar{V}}} \ge \frac{b(\bar{\boldsymbol{w}}, \boldsymbol{q})}{\|\bar{\boldsymbol{w}}\|_{\bar{V}}} \ge \frac{1}{\sqrt{\beta_s^{-2}+1}}\frac{|\boldsymbol{q}|_{\boldsymbol{Q}}^2}{|\boldsymbol{q}|_{\boldsymbol{Q}}} = c_b|\boldsymbol{q}|_{\boldsymbol{Q}}, \quad \forall \boldsymbol{q}\in\boldsymbol{Q}.
$$

We have now established the well-posedness of the Biot-Brinkman problem under the specific combined norm $\|\cdot\|_{\bar{X}}$ of the form (3.13), specified through (3.16) and (3.17). Next, we show that this combined norm is equivalent to the norm $\|\cdot\|_{X}$ defined by

$$
\|(\boldsymbol{w}, z, \boldsymbol{q})\|_{X}^2 := \|\boldsymbol{w}\|_{U}^2 + \|z\|_{V}^2 + \|\boldsymbol{q}\|_{P}^2. \tag{3.18}
$$

The following Lemma is useful in establishing this norm equivalence, cf. [25, Lemma 2.1] where the statement has been proven for $\boldsymbol{\alpha} = (1, 1, \ldots, 1)^T$.

Lemma 3.3 *For any $a > 0$ and $b > 0$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$, we have that*

$$
(aI_{n\times n} + b\boldsymbol{\alpha}\boldsymbol{\alpha}^T)^{-1} = a^{-1}I - a^{-1}(ab^{-1} + \boldsymbol{\alpha}^T\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T, \tag{3.19}
$$

and

$$
\boldsymbol{\alpha}^T(aI_{n\times n} + b\boldsymbol{\alpha}\boldsymbol{\alpha}^T)^{-1}\boldsymbol{\alpha} = \frac{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}{ab^{-1} + \boldsymbol{\alpha}^T\boldsymbol{\alpha}}b^{-1} \le b^{-1}. \tag{3.20}
$$

Proof The proof follows the lines of the proof of Lemma 2.1 in [25].

Now we can establish the following norm equivalence result.

Lemma 3.4 *The norm (3.18) defined in terms of (2.12) is equivalent to the combined norm (3.13) based on (3.16) and (3.17).*

Proof First, we note that

$$B\bar{\boldsymbol{w}} = \begin{pmatrix} -\alpha_1 \mathrm{div}\boldsymbol{w} - \tau\,\mathrm{div}\,z_1 \\ -\alpha_2 \mathrm{div}\boldsymbol{w} - \tau\,\mathrm{div}\,z_2 \\ \vdots \\ -\alpha_n \mathrm{div}\boldsymbol{w} - \tau\,\mathrm{div}\,z_n \end{pmatrix} = -\mathrm{div}\boldsymbol{w} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} + \tau \begin{pmatrix} -\mathrm{div}\,z_1 \\ -\mathrm{div}\,z_2 \\ \vdots \\ -\mathrm{div}\,z_n \end{pmatrix}$$
$$\equiv -\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w} - \tau\mathrm{Div}z.$$

Then for any $1 > \epsilon > 0$, by Cauchy's inequality, we obtain

$$\begin{aligned}(\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) &= \left(\Lambda^{-1}(\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w} + \tau\mathrm{Div}z), (\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w} + \tau\mathrm{Div}z)\right) \\ &= (\Lambda^{-1}\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}, \boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}) + 2(\Lambda^{-1}\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}, \tau\mathrm{Div}z) + (\Lambda^{-1}\tau\mathrm{Div}z, \tau\mathrm{Div}z) \\ &\geq -(\epsilon^{-1} - 1)(\Lambda^{-1}\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}, \boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}) + (1-\epsilon)(\Lambda^{-1}\tau\mathrm{Div}z, \tau\mathrm{Div}z) \\ &\geq -(\epsilon^{-1} - 1)((\Lambda_3 + \Lambda_4)^{-1}\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}, \boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}) + (1-\epsilon)(\Lambda^{-1}\tau\mathrm{Div}z, \tau\mathrm{Div}z).\end{aligned}$$

By Lemma 3.3, with $a = \tau^2 R$, $b = \frac{1}{2\mu+\lambda}$, we have

$$\begin{aligned}(\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) &\geq -(\epsilon^{-1} - 1)((\Lambda_3 + \Lambda_4)^{-1}\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}, \boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}) + (1-\epsilon)(\Lambda^{-1}\tau\mathrm{Div}z, \tau\mathrm{Div}z) \\ &= -(\epsilon^{-1} - 1)(\boldsymbol{\alpha}^T(\Lambda_3 + \Lambda_4)^{-1}\boldsymbol{\alpha}\mathrm{div}\boldsymbol{w}, \mathrm{div}\boldsymbol{w}) + (1-\epsilon)(\Lambda^{-1}\tau\mathrm{Div}z, \tau\mathrm{Div}z) \\ &\geq -(\epsilon^{-1} - 1)(2\mu + \lambda)(\mathrm{div}\,\boldsymbol{w}, \mathrm{div}\,\boldsymbol{w}) + (1-\epsilon)(\Lambda^{-1}\tau\mathrm{Div}z, \tau\mathrm{Div}z).\end{aligned}$$

Therefore, we get

$$\begin{aligned}\|\bar{\boldsymbol{w}}\|_{\bar{V}}^2 &= 2\mu(\boldsymbol{\varepsilon}(\boldsymbol{w}), \boldsymbol{\varepsilon}(\boldsymbol{w})) + \lambda(\mathrm{div}\,\boldsymbol{w}, \mathrm{div}\,\boldsymbol{w})x + \sum_{i=1}^{n}\gamma_i(\boldsymbol{\varepsilon}(z_i), \boldsymbol{\varepsilon}(z_i)) \\ &\quad + \sum_{i=1}^{n}(\tau K_i^{-1}z_i, z_i) + (\Lambda^{-1}B\bar{\boldsymbol{w}}, B\bar{\boldsymbol{w}}) \\ &\geq 2\mu(\boldsymbol{\varepsilon}(\boldsymbol{w}), \boldsymbol{\varepsilon}(\boldsymbol{w})) + \lambda(\mathrm{div}\,\boldsymbol{w}, \mathrm{div}\,\boldsymbol{w}) - (\epsilon^{-1} - 1)(2\mu + \lambda)(\mathrm{div}\,\boldsymbol{w}, \mathrm{div}\,\boldsymbol{w}) \\ &\quad + \sum_{i=1}^{n}\gamma_i(\boldsymbol{\varepsilon}(z_i), \boldsymbol{\varepsilon}(z_i)) + \sum_{i=1}^{n}(\tau K_i^{-1}z_i, z_i) + (1-\epsilon)(\Lambda^{-1}\tau\mathrm{Div}z, \tau\mathrm{Div}z).\end{aligned}$$

Now, for $\epsilon = \frac{2}{3}$, we obtain

$$\begin{aligned}\|\bar{\boldsymbol{w}}\|_{\bar{V}}^2 &\geq 2\mu(\boldsymbol{\varepsilon}(\boldsymbol{w}), \boldsymbol{\varepsilon}(\boldsymbol{w})) + \lambda(\mathrm{div}\,\boldsymbol{w}, \mathrm{div}\,\boldsymbol{w}) - \frac{1}{2}(2\mu + \lambda)(\mathrm{div}\,\boldsymbol{w}, \mathrm{div}\,\boldsymbol{w}) \\ &\quad + \sum_{i=1}^{n}\gamma_i(\boldsymbol{\varepsilon}(z_i), \boldsymbol{\varepsilon}(z_i)) + \sum_{i=1}^{n}(\tau K_i^{-1}z_i, z_i) + \frac{1}{3}(\Lambda^{-1}\tau^2\mathrm{Div}z, \mathrm{Div}z) \\ &\geq \frac{1}{2}\left(2\mu(\boldsymbol{\varepsilon}(\boldsymbol{w}), \boldsymbol{\varepsilon}(\boldsymbol{w})) + \lambda(\mathrm{div}\,\boldsymbol{w}, \mathrm{div}\,\boldsymbol{w})\right) \\ &\quad + \frac{1}{3}\left(\sum_{i=1}^{n}\gamma_i(\boldsymbol{\varepsilon}(z_i), \boldsymbol{\varepsilon}(z_i)) + \sum_{i=1}^{n}(\tau K_i^{-1}z_i, z_i) + (\Lambda^{-1}\tau^2\mathrm{Div}z, \mathrm{Div}z)\right),\end{aligned}$$

namely $\|\boldsymbol{w}\|_{\boldsymbol{U}}^2 + \|z\|_{\boldsymbol{V}}^2 \lesssim \|\bar{\boldsymbol{w}}\|_{\bar{\boldsymbol{V}}}^2$. On the other hand, it is obvious that

$$\|\bar{\boldsymbol{w}}\|_{\bar{\boldsymbol{V}}}^2 \lesssim \|\boldsymbol{w}\|_{\boldsymbol{U}}^2 + \|z\|_{\boldsymbol{V}}^2.$$

Together, this gives $\|\bar{\boldsymbol{w}}\|_{\bar{\boldsymbol{V}}}^2 \cong \|\boldsymbol{w}\|_{\boldsymbol{U}}^2 + \|\boldsymbol{z}\|_{\boldsymbol{V}}^2$.

In view of Theorem 3.2 and Lemma 3.4, we conclude that the Biot-Brinkman problem is also well-posed under the norm (3.18) defined in terms of (2.12). We summarize our results in the following theorem.

Theorem 3.5 (i) *There exists a positive constant C_b independent of the parameters λ, K_i^{-1}, s_i, β_{ij}, $i, j \in \{1, \ldots, n\}$, the network scale n and the time step τ such that the inequality*

$$|\mathcal{A}((\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}), (\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{q}))| \leq C_b(\|\boldsymbol{u}|_{\boldsymbol{U}} + \|\boldsymbol{v}\|_{\boldsymbol{V}} + \|\boldsymbol{p}\|_{\boldsymbol{P}})(\|\boldsymbol{w}\|_{\boldsymbol{U}} + \|\boldsymbol{z}\|_{\boldsymbol{V}} + \|\boldsymbol{q}\|_{\boldsymbol{P}})$$

holds true for any $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$, $(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{q}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$.

(ii) *There is a constant $\omega > 0$ independent of the parameters $\lambda, K_i^{-1}, s_i, \beta_{ij}, i, j \in \{1, \ldots, n\}$, the number of networks n and the time step τ such that*

$$\inf_{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{p})\in\boldsymbol{X}} \sup_{(\boldsymbol{w},\boldsymbol{z},\boldsymbol{q})\in\boldsymbol{X}} \frac{\mathcal{A}((\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}), (\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{q}))}{(\|\boldsymbol{u}\|_{\boldsymbol{U}} + \|\boldsymbol{v}\|_{\boldsymbol{V}} + \|\boldsymbol{p}\|_{\boldsymbol{P}})(\|\boldsymbol{w}\|_{\boldsymbol{U}} + \|\boldsymbol{z}\|_{\boldsymbol{V}} + \|\boldsymbol{q}\|_{\boldsymbol{P}})} \geq \omega,$$

where $\boldsymbol{X} := \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$.

(iii) *The MPET system (3.4) has a unique solution $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}) \in \boldsymbol{U} \times \boldsymbol{V} \times \boldsymbol{P}$ and the following stability estimate holds:*

$$\|\boldsymbol{u}\|_{\boldsymbol{U}} + \|\boldsymbol{v}\|_{\boldsymbol{V}} + \|\boldsymbol{p}\|_{\boldsymbol{P}} \leq C_1(\|\boldsymbol{f}\|_{\boldsymbol{U}'} + \|\boldsymbol{g}\|_{\boldsymbol{P}'}),$$

where C_1 is a positive constant independent of the parameters $\lambda, K_i^{-1}, s_i, \beta_{ij}, i, j \in \{1, \ldots, n\}$, the network scale n and the time step τ, and $\|\boldsymbol{f}\|_{\boldsymbol{U}'} = \sup_{\boldsymbol{w}\in\boldsymbol{U}} \frac{(\boldsymbol{f},\boldsymbol{w})}{\|\boldsymbol{w}\|_{\boldsymbol{U}}}$, $\|\boldsymbol{g}\|_{\boldsymbol{P}'} = \sup_{\boldsymbol{q}\in\boldsymbol{P}} \frac{(\boldsymbol{g},\boldsymbol{q})}{\|\boldsymbol{q}\|_{\boldsymbol{P}}} = \|\Lambda^{-\frac{1}{2}}\boldsymbol{g}\|$.

4 Discrete Generalized Biot-Brinkman Problems

Stable and parameter-robust discretizations for the multiple network poroelasticity equations have been proposed based on a classical three-field formulation using a discontinuous Galerkin (DG) [1, 29] formulation of the momentum equation resulting in strong mass conservation, see [24], or based on a total pressure formulation in the setting of conforming methods in [36]. These discrete models have been developed as generalizations of the corresponding Biot models, see [23] in case of conservative discretizations and [35] in case of the total pressure scheme. A hybridized version of the method in [23] has recently been presented in [33]. For other conforming parameter-robust discretizations of the Biot model see also [15, 42] and [34], where the latter method is based on a total pressure formulation introducing the flux as a fourth field, which then also results in mass conservation. In this paper we extend the approach from [23, 24] to obtain mass-conservative discretizations for the generalized Biot-Brinkman system (3.2)–(3.3), which generalizes the MPET system.

4.1 Notation

Consider a shape-regular triangulation $\mathcal{T}_h$ of the domain Ω into triangles/tetrahedrons, where the subscript h indicates the mesh-size. Following the standard notation, we first denote the set of all interior edges/faces and the set of all boundary edges/faces of $\mathcal{T}_h$ by $\mathcal{E}_h^I$ and $\mathcal{E}_h^B$

respectively, their union by $\mathcal{E}_h$ and then we define the broken Sobolev spaces

$$H^s(\mathcal{T}_h) = \{\phi \in L^2(\Omega), \text{ such that } \phi|_T \in H^s(T) \text{ for all } T \in \mathcal{T}_h\},$$

for $s \geq 1$.

Next we introduce the notion of jumps $[\cdot]$ and averages $\{\cdot\}$ as follows. For any $q \in H^1(\mathcal{T}_h)$, $\boldsymbol{v} \in H^1(\mathcal{T}_h)^d$ and $\boldsymbol{\tau} \in H^1(\mathcal{T}_h)^{d\times d}$ and any $e \in \mathcal{E}_h^I$ the jumps are given as

$$[q] = q|_{\partial T_1 \cap e} - q|_{\partial T_2 \cap e}, \quad [\boldsymbol{v}] = \boldsymbol{v}|_{\partial T_1 \cap e} - \boldsymbol{v}|_{\partial T_2 \cap e},$$

and the averages as

$$\{\boldsymbol{v}\} = \frac{1}{2}(\boldsymbol{v}|_{\partial T_1 \cap e} \cdot \boldsymbol{n}_1 - \boldsymbol{v}|_{\partial T_2 \cap e} \cdot \boldsymbol{n}_2), \quad \{\boldsymbol{\tau}\} = \frac{1}{2}(\boldsymbol{\tau}|_{\partial T_1 \cap e}\boldsymbol{n}_1 - \boldsymbol{\tau}|_{\partial T_2 \cap e}\boldsymbol{n}_2),$$

while for $e \in \mathcal{E}_h^B$,

$$[q] = q|_e, \quad [\boldsymbol{v}] = \boldsymbol{v}|_e, \quad \{\boldsymbol{v}\} = \boldsymbol{v}|_e \cdot \boldsymbol{n}, \quad \{\boldsymbol{\tau}\} = \boldsymbol{\tau}|_e \boldsymbol{n}.$$

Here T_1 and T_2 are any two elements from the triangulation that share an edge or face e while $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ denote the corresponding unit normal vectors to e pointing to the exterior of T_1 and T_2, respectively.

4.2 Mixed finite element spaces and discrete formulation

We consider the following finite element spaces to approximate the displacement, fluxes and pressures:

$$\begin{aligned}
\boldsymbol{U}_h &= \{\boldsymbol{u} \in H(\mathrm{div}, \Omega) : \boldsymbol{u}|_T \in \boldsymbol{U}(T),\ T \in \mathcal{T}_h;\ \boldsymbol{u} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega\},\\
\boldsymbol{V}_{i,h} &= \{\boldsymbol{v} \in H(\mathrm{div}, \Omega) : \boldsymbol{v}|_T \in \boldsymbol{V}_i(T),\ T \in \mathcal{T}_h;\ \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega\},\quad i = 1, \ldots, n,\\
P_{i,h} &= \left\{p \in L^2(\Omega) : p|_T \in P_i(T),\ T \in \mathcal{T}_h;\ \int_\Omega p dx = 0\right\},\quad i = 1, \ldots, n.
\end{aligned}$$

The discretizations that we consider here, define the local spaces $\boldsymbol{U}(T)/\boldsymbol{V}_i(T)/P_i(T)$ via the triplets of spaces $\mathrm{BDM}_l(T)/\mathrm{BDM}_l(T)/\mathrm{P}_{l-1}(T)$, or $\mathrm{RT}_l(T)/\mathrm{RT}_l(T)/\mathrm{P}_{l-1}(T)$, or $\mathrm{BDM}_l(T)/\mathrm{RT}_l(T)/\mathrm{P}_{l-1}(T)$, or $\mathrm{RT}_l(T)/\mathrm{BDM}_l(T)/\mathrm{P}_{l-1}(T)$ for $l \geq 1$. Note that for each of these choices, the condition $\mathrm{div}\,\boldsymbol{U}(T) = \mathrm{div}\,\boldsymbol{V}_i(T) = P_i(T)$ is fulfilled. We remark that the tangential part of the displacement boundary condition (2.9) is enforced by a Nitsche method, see e.g. [21].

From a computational point of view, it is preferable to choose $\boldsymbol{U}(T)/\boldsymbol{V}_i(T)/P_i(T)$ $= \mathrm{BDM}_l(T)/\mathrm{BDM}_l(T)/\mathrm{P}_{l-1}(T)$ since the l-th order BDM element achieves the same convergence order using less unknowns than l-th order RT element when approximating the Laplacian operator. Furthermore the orthogonality constraint for the pressures in $P_{i,h}$ is realized in the implementation by introducing (scalar) Lagrange multipliers.

Remark 4.1 Working with rectangular or hexahedral meshes, one can also use Raviart-Thomas or Brezzi-Douglas-Fortin-Marini elements on rectangles and cubes. In the former case, the local space $\boldsymbol{U}(T) = \boldsymbol{V}_i(T) = \mathrm{RT}_l(T)$ for displacements and fluxes fits the local space $P_i(T) = \mathrm{Q}_l(T)$ for pressures, in the latter case $\boldsymbol{U}(T) = \boldsymbol{V}_i(T) = \mathrm{BDFM}_{l+1}(T)$ fits $P_i(T) = \mathrm{P}_l(T)$, in the sense that $\mathrm{div}\,\boldsymbol{U}(T) = \mathrm{div}\,\boldsymbol{V}_i(T) = P_i(T)$ is satisfied again, ensuring strong (pointwise) mass conservation. Here T denotes a rectangle in two and a cube in three space dimensions and $\mathrm{Q}_l(T)$ the space of polynomials on T which are of degree less than

or equal to l in each variable (when all other variables are fixed) whereas $\mathrm{P}_l(T)$ denotes the local polynomials of (total) degree at most l. For more details, see [9].

Let us denote $\boldsymbol{v}_h^T = (\boldsymbol{v}_{1,h}^T, \ldots, \boldsymbol{v}_{n,h}^T)$, $\boldsymbol{p}_h^T = (p_{1,h}, \ldots, p_{n,h})$, $\boldsymbol{z}_h^T = (\boldsymbol{z}_{1,h}^T, \ldots, \boldsymbol{z}_{n,h}^T)$, $\boldsymbol{q}_h^T = (q_{1,h}, \ldots, q_{n,h})$ and

$$\boldsymbol{V}_h = \boldsymbol{V}_{1,h} \times \ldots \times \boldsymbol{V}_{n,h}, \quad \boldsymbol{P}_h = P_{1,h} \times \ldots \times P_{n,h}, \quad \boldsymbol{X}_h = \boldsymbol{U}_h \times \boldsymbol{V}_h \times \boldsymbol{P}_h.$$

The discretization of the variational problem (3.2)–(3.3) now is given as follows: find $(\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h) \in \boldsymbol{X}_h$, such that for any $(\boldsymbol{w}_h, \boldsymbol{z}_h, \boldsymbol{q}_h) \in \boldsymbol{X}_h$ and $i = 1, \ldots, n$

$$a_h(\boldsymbol{u}_h, \boldsymbol{w}_h) + \lambda(\operatorname{div} \boldsymbol{u}_h, \operatorname{div} \boldsymbol{w}_h) + (\boldsymbol{\alpha} \cdot \boldsymbol{p}_h, \operatorname{div} \boldsymbol{w}_h) = (\boldsymbol{f}, \boldsymbol{w}_h), \tag{4.1a}$$

$$\gamma_i a_h(\boldsymbol{v}_{i,h}, \boldsymbol{z}_{i,h}) + (\tau K_i^{-1} \boldsymbol{v}_{i,h}, \boldsymbol{z}_{i,h}) + (p_{i,h}, \tau \operatorname{div} \boldsymbol{z}_{i,h}) = 0, \tag{4.1b}$$

$$(\operatorname{div} \boldsymbol{u}_h, \alpha_i q_{i,h}) + (\tau \operatorname{div} \boldsymbol{v}_{i,h}, q_{i.h}) - s_i(p_{i,h}, q_{i,h}) + \sum_{j=1}^{n} \tau \beta_{ij}(p_{j,h}, q_{i,h}) = (g_i, q_{i,h}), \tag{4.1c}$$

where

$$a_h(\boldsymbol{\phi}, \boldsymbol{\psi}) = \sum_{T \in \mathcal{T}_h} \int_T \boldsymbol{\varepsilon}(\boldsymbol{\phi}) : \boldsymbol{\varepsilon}(\boldsymbol{\psi}) \, \mathrm{d}x - \sum_{e \in \mathcal{E}_h} \int_e \{\boldsymbol{\varepsilon}(\boldsymbol{\phi})\} \cdot [\boldsymbol{\psi}_t] \, \mathrm{d}s - \sum_{e \in \mathcal{E}_h} \int_e \{\boldsymbol{\varepsilon}(\boldsymbol{\psi})\} \cdot [\boldsymbol{\phi}_t] \, \mathrm{d}s + \sum_{e \in \mathcal{E}_h} \int_e \eta h_e^{-1} [\boldsymbol{\phi}_t] \cdot [\boldsymbol{\psi}_t] \, \mathrm{d}s, \tag{4.2}$$

and η is a stabilization parameter independent of all other problem parameters, the network scale n and the mesh size h, h_e is the size of edge e. However, η will depend on shape regularity of the triangulation and polynomial order of the finite element space and will affect the condition number, see also Remark 5.1.

We note that the discrete variational problem (4.1) has been derived for the weak formulation (3.4) with homogeneous boundary conditions. For general rescaled boundary conditions with DG discretizations we refer the reader to e.g. [23].

4.3 Stability Properties

For any function $\boldsymbol{\phi} \in \boldsymbol{H}^2(\mathcal{T}_h) := H^2(\mathcal{T}_h)^d$, consider the following mesh dependent norms

$$\|\boldsymbol{\phi}\|_h^2 = \sum_{T \in \mathcal{T}_h} \|\boldsymbol{\varepsilon}(\boldsymbol{\phi})\|_T^2 + \sum_{e \in \mathcal{E}_h} h_e^{-1} \|[\boldsymbol{\phi}_t]\|_e^2,$$

$$\|\boldsymbol{\phi}\|_{1,h}^2 = \sum_{T \in \mathcal{T}_h} \|\nabla \boldsymbol{\phi}\|_T^2 + \sum_{e \in \mathcal{E}_h} h_e^{-1} \|[\boldsymbol{\phi}_t]\|_e^2,$$

and

$$\|\boldsymbol{\phi}\|_{DG}^2 = \sum_{T \in \mathcal{T}_h} \|\nabla \boldsymbol{\phi}\|_T^2 + \sum_{e \in \mathcal{E}_h} h_e^{-1} \|[\boldsymbol{\phi}_t]\|_e^2 + \sum_{T \in \mathcal{T}_h} h_T^2 |\boldsymbol{\phi}|_{2,T}^2. \tag{4.3}$$

Details about the well-posedness and approximation properties of the DG formulation of elasticity, Stokes and Brinkman-type systems can be found in [22, 28].

Now, for $\boldsymbol{u} \in H(\operatorname{div}, \Omega) \cap \boldsymbol{H}^2(\mathcal{T}_h)$, we define the norm

$$\|\boldsymbol{u}\|_{U_h}^2 = \|\boldsymbol{u}\|_{DG}^2 + \lambda \|\operatorname{div} \boldsymbol{u}\|^2, \tag{4.4}$$

and for $\boldsymbol{v} \in H(\mathrm{div}, \Omega) \cap \boldsymbol{H}^2(\mathcal{T}_h)$, we define the norm

$$\|\boldsymbol{v}\|_{\boldsymbol{V}_h}^2 = \sum_{i=1}^{n} \left(\gamma_i \|\boldsymbol{v}_i\|_{DG}^2 + (\tau K_i^{-1} \boldsymbol{v}_i, \boldsymbol{v}_i)\right) + (\Lambda^{-1} \mathrm{Div}\, \boldsymbol{v}, \mathrm{Div}\, \boldsymbol{v}). \tag{4.5}$$

The well-posedness and approximation properties of the DG formulation are detailed in [22, 28]. Here we briefly present some important results:

- $\|\cdot\|_{DG}$, $\|\cdot\|_h$, and $\|\cdot\|_{1,h}$ are equivalent on $\boldsymbol{U}_h$; that is

$$\|\boldsymbol{u}_h\|_{DG} \eqsim \|\boldsymbol{u}_h\|_h \eqsim \|\boldsymbol{u}_h\|_{1,h}, \text{ for all } \boldsymbol{u}_h \in \boldsymbol{U}_h.$$

- a_h from (4.2) is continuous and it holds true that

$$|a_h(\boldsymbol{u}, \boldsymbol{w})| \lesssim \|\boldsymbol{u}\|_{DG} \|\boldsymbol{w}\|_{DG}, \quad \text{for all} \quad \boldsymbol{u},\ \boldsymbol{w} \in \boldsymbol{H}^2(\mathcal{T}_h). \tag{4.6}$$

- The following inf-sup conditions are satisfied

$$\begin{aligned} &\inf_{(q_{1,h},\cdots,q_{n,h}) \in P_{1,h} \times \cdots \times P_{n,h}} \sup_{\boldsymbol{u}_h \in \boldsymbol{U}_h} \frac{(\mathrm{div}\, \boldsymbol{u}_h, \sum_{i=1}^{n} q_{i,h})}{\|\boldsymbol{u}_h\|_{1,h} \|\sum_{i=1}^{n} q_{i,h}\|} \geq \beta_{sd}, \\ &\inf_{q_{i,h} \in P_{i,h}} \sup_{\boldsymbol{v}_{i,h} \in \boldsymbol{V}_{i,h}} \frac{(\mathrm{div}\, \boldsymbol{v}_{i,h}, q_{i,h})}{\|\boldsymbol{v}_{i,h}\|_{1,h} \|q_{i,h}\|} \geq \beta_{dd}, \quad i = 1, \ldots, n. \end{aligned} \tag{4.7}$$

Using the definition of the matrices Λ_1 and Λ_2, next we define the bilinear form

$$\begin{aligned} &\mathcal{A}_h((\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h), (\boldsymbol{w}_h, \boldsymbol{z}_h, \boldsymbol{q}_h)) \\ &= a_h(\boldsymbol{u}_h, \boldsymbol{w}_h) + \lambda(\mathrm{div}\, \boldsymbol{u}_h, \mathrm{div}\, \boldsymbol{w}_h) \\ &\quad + \sum_{i=1}^{n} (\alpha_i p_{i,h}, \mathrm{div} \boldsymbol{w}_h) + \sum_{i=1}^{n} \gamma_i a_h(\boldsymbol{v}_{i,h}, \boldsymbol{z}_{i,h}) \\ &\quad + \sum_{i=1}^{n} (\tau K_i^{-1} \boldsymbol{v}_{i,h}, \boldsymbol{z}_{i,h}) + \tau(\boldsymbol{p}_h, \mathrm{Div}\, \boldsymbol{z}_h) \\ &\quad + \sum_{i=1}^{n} (\mathrm{div}\, \boldsymbol{u}_h, \alpha_i q_{i,h}) + \tau(\mathrm{Div}\, \boldsymbol{v}_h, \boldsymbol{q}_h) - ((\Lambda_1 + \Lambda_2) \boldsymbol{p}_h, \boldsymbol{q}_h), \end{aligned} \tag{4.8}$$

related to problem (4.1a)–(4.1c).

We equip $\boldsymbol{X}_h$ with the norm defined by $\|(\cdot, \cdot, \cdot)\|_{\boldsymbol{X}_h}^2 := \|\cdot\|_{\boldsymbol{U}_h}^2 + \|\cdot\|_{\boldsymbol{V}_h}^2 + \|\cdot\|_{\boldsymbol{P}}^2$. Similar to Theorem 3.5, the following uniform stability result holds:

Theorem 4.2 (i) *For any $\boldsymbol{u}_h, \boldsymbol{w}_h \in \boldsymbol{U}_h$; $\boldsymbol{v}_h, \boldsymbol{z}_h \in \boldsymbol{V}_h$; $\boldsymbol{p}_h, \boldsymbol{q}_h \in \boldsymbol{P}_h$ there exists a positive constant C_{bd} independent of all model parameters, the network scale n and the mesh size h such that the inequality*

$$|\mathcal{A}_h((\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h), (\boldsymbol{w}_h, \boldsymbol{z}_h, \boldsymbol{q}_h))| \leq C_{bd} \|(\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h)\|_{X_h} \|(\boldsymbol{w}_h, \boldsymbol{z}_h, \boldsymbol{q}_h)\|_{X_h}$$

holds true.

(ii) *There exists a constant $\omega_d > 0$ independent of all discretization and model parameters such that*

$$\inf_{(\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h) \in \boldsymbol{X}_h} \sup_{(\boldsymbol{w}_h, \boldsymbol{z}_h, \boldsymbol{q}_h) \in \boldsymbol{X}_h} \frac{\mathcal{A}_h((\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h), (\boldsymbol{w}_h, \boldsymbol{z}_h, \boldsymbol{q}_h))}{\|(\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h)\|_{X_h} \|(\boldsymbol{w}_h, \boldsymbol{z}_h, \boldsymbol{q}_h)\|_{X_h}} \geq \omega_d. \tag{4.9}$$

(iii) *Let* $(\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h) \in \boldsymbol{X}_h$ *solve* (4.1a)-(4.1c) *and*

$$\|\boldsymbol{f}\|_{\boldsymbol{U}_h'} = \sup_{\boldsymbol{w}_h \in \boldsymbol{U}_h} \frac{(\boldsymbol{f}, \boldsymbol{w}_h)}{\|\boldsymbol{w}_h\|_{\boldsymbol{U}_h}}, \quad \|\boldsymbol{g}\|_{\boldsymbol{P}'} = \sup_{\boldsymbol{q}_h \in \boldsymbol{P}_h} \frac{(\boldsymbol{g}, \boldsymbol{q}_h)}{\|\boldsymbol{q}_h\|_{\boldsymbol{P}}}.$$

Then the estimate

$$\|\boldsymbol{u}_h\|_{\boldsymbol{U}_h} + \|\boldsymbol{v}_h\|_{\boldsymbol{V}} + \|\boldsymbol{p}_h\|_{\boldsymbol{P}} \leq C_2(\|\boldsymbol{f}\|_{\boldsymbol{U}_h'} + \|\boldsymbol{g}\|_{\boldsymbol{P}'})$$

holds with a constant C_2 *independent of the network scale* n*, the mesh size* h*, the time step* τ *and the parameters* λ, K_i^{-1}, s_i, β_{ij}, $i, j \in \{1, \ldots, n\}$.

4.4 Error Estimates

This subsection summarizes the error estimates that follow from the stability results presented in Sect. 4.3.

Theorem 4.3 *Assume that* $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{p}) \in \boldsymbol{U} \cap \boldsymbol{H}^2(\mathcal{T}_h) \times \boldsymbol{V} \cap \boldsymbol{H}^2(\mathcal{T}_h) \times \boldsymbol{P}$ *is the unique solution of* (3.2)–(3.3)*, and let* $(\boldsymbol{u}_h, \boldsymbol{v}_h, \boldsymbol{p}_h)$ *be the solution of* (4.1)*. Then the error estimates*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{U}_h} + \|\boldsymbol{v} - \boldsymbol{v}_h\|_{\boldsymbol{V}_h} \lesssim \inf_{\boldsymbol{w}_h \in \boldsymbol{U}_h, \boldsymbol{z}_h \in \boldsymbol{V}_h} \Big(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{\boldsymbol{U}_h} + \|\boldsymbol{v} - \boldsymbol{z}_h\|_{\boldsymbol{V}_h} \Big), \tag{4.10}$$

and

$$\|\boldsymbol{p} - \boldsymbol{p}_h\|_{\boldsymbol{P}} \lesssim \inf_{\boldsymbol{w}_h \in \boldsymbol{U}_h, \boldsymbol{z}_h \in \boldsymbol{V}_h, \boldsymbol{q}_h \in \boldsymbol{P}_h} \Big(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{\boldsymbol{U}_h} + \|\boldsymbol{v} - \boldsymbol{z}_h\|_{\boldsymbol{V}_h} + \|\boldsymbol{p} - \boldsymbol{q}_h\|_{\boldsymbol{P}} \Big) \tag{4.11}$$

hold true, where the inequality constants are independent of the parameters λ, K_i^{-1}, s_i, β_{ij} *for* $i, j = 1, \ldots, n$*, the network scale* n*, the mesh size* h *and the time step* τ.

Proof The proof of this result is analogous to the proof of Theorem 5.2 in [23].

Remark 4.4 In particular, the above theorem shows that the proposed discretizations are locking-free. Note that estimate (4.10) controls the error in $\boldsymbol{u}$ plus the error in $\boldsymbol{v}$ by the sum of the errors of the corresponding best approximations whereas estimate (4.11) requires the best approximation errors of all three vector variables $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{p}$ to control the error in $\boldsymbol{p}$.

4.5 A Norm Equivalent Preconditioner

We consider the following block-diagonal operator

$$\mathcal{B} := \begin{bmatrix} \mathcal{B}_{\boldsymbol{u}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mathcal{B}_{\boldsymbol{v}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \mathcal{B}_{\boldsymbol{p}} \end{bmatrix}^{-1}, \tag{4.12}$$

where

$$\mathcal{B}_{\boldsymbol{u}} = -\operatorname{div}\boldsymbol{\varepsilon} - \lambda\nabla\operatorname{div},$$

$$\mathcal{B}_{\boldsymbol{v}} = \begin{bmatrix} -\gamma_1\operatorname{div}\boldsymbol{\varepsilon} + \tau K_1^{-1} I & 0 & \dots & 0 \\ 0 & -\gamma_2\operatorname{div}\boldsymbol{\varepsilon} + \tau K_2^{-1} I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -\gamma_n\operatorname{div}\boldsymbol{\varepsilon} + \tau K_n^{-1} I \end{bmatrix} - \begin{bmatrix} \tilde{\Lambda}_{11}\nabla\operatorname{div} & \tilde{\Lambda}_{12}\nabla\operatorname{div} & \dots & \tilde{\Lambda}_{1n}\nabla\operatorname{div} \\ \tilde{\Lambda}_{21}\nabla\operatorname{div} & \tilde{\Lambda}_{22}\nabla\operatorname{div} & \dots & \tilde{\Lambda}_{2n}\nabla\operatorname{div} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\Lambda}_{n1}\nabla\operatorname{div} & \tilde{\Lambda}_{n2}\nabla\operatorname{div} & \dots & \tilde{\Lambda}_{nn}\nabla\operatorname{div} \end{bmatrix},$$

and

$$\mathcal{B}_{\boldsymbol{p}} = \begin{bmatrix} \Lambda_{11} I & \Lambda_{12} I & \dots & \Lambda_{1n} I \\ \Lambda_{21} I & \Lambda_{22} I & \dots & \Lambda_{2n} I \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{n1} I & \Lambda_{n2} I & \dots & \Lambda_{nn} I \end{bmatrix}.$$

Here, Λ_{ij}, $\tilde{\Lambda}_{ij}$, $i, j = 1, \dots, n$ are the entries of Λ and Λ^{-1}, respectively.

As substantiated in [24], the stability results for the operator $\mathcal{A}$ in (3.5) imply that the operator $\mathcal{B}$ is a uniform norm-equivalent (canonical) block-diagonal preconditioner that is robust with respect to all model and discretization parameters. Note that $\mathcal{B}$ defines a canonical uniform block-diagonal preconditioner on the continuous as well as on the discrete level as long as discrete inf-sup conditions analogous to (3.14) and (3.15) are satisfied, cf. [24].

For discrete counterpart, denote by $\mathcal{A}_h$ the operator induced by the bilinear form (4.8), namely

$$\mathcal{A}_h := \begin{bmatrix} A_{h,\boldsymbol{u}} & \mathbf{0} & B_{h,\boldsymbol{u}}^T \\ \mathbf{0} & A_{h,\boldsymbol{v}} & B_{h,\boldsymbol{v}}^T \\ B_{h,\boldsymbol{u}} & B_{h,\boldsymbol{v}} & -A_{h,\boldsymbol{p}} \end{bmatrix}, \tag{4.13}$$

where

$$A_{h,\boldsymbol{u}} = -\operatorname{div}_h\boldsymbol{\varepsilon}_h - \lambda\nabla_h\operatorname{div}_h$$

$$A_{h,\boldsymbol{v}} = \begin{bmatrix} -\gamma_1\operatorname{div}_h\boldsymbol{\varepsilon}_h + \tau K_1^{-1} I_h & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & -\gamma_n\operatorname{div}_h\boldsymbol{\varepsilon}_h + \tau K_n^{-1} I_h \end{bmatrix},$$

$$B_{h,\boldsymbol{u}} = \begin{bmatrix} -\alpha_1\operatorname{div}_h \\ -\alpha_2\operatorname{div}_h \\ \vdots \\ -\alpha_n\operatorname{div}_h \end{bmatrix}, \quad B_{h,\boldsymbol{v}} = \begin{bmatrix} -\tau\operatorname{div}_h & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & -\tau\operatorname{div}_h \end{bmatrix},$$

and

$$A_{h,\boldsymbol{p}} = \begin{pmatrix} s_1 I_h & -\tau\beta_{12} I_h & \dots & -\tau\beta_{1n} I_h \\ -\tau\beta_{21} I_h & s_2 I_h & \dots & -\tau\beta_{2n} I_h \\ \vdots & \vdots & \ddots & \vdots \\ -\tau\beta_{n1} I_h & -\tau\beta_{n2} I_h & \dots & s_n I_h \end{pmatrix}.$$

And the corresponding block preconditioner for $\mathcal{A}_h$ is

$$\mathcal{B}_h := \begin{bmatrix} \mathcal{B}_{h,\boldsymbol{u}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{B}_{h,\boldsymbol{v}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{B}_{h,\boldsymbol{p}} \end{bmatrix}^{-1}, \tag{4.14}$$

where

$$\mathcal{B}_{h,\boldsymbol{u}} = -\operatorname{div}_h \boldsymbol{\varepsilon}_h - \lambda \nabla_h \operatorname{div}_h,$$

$$\mathcal{B}_{h,\boldsymbol{v}} = \begin{bmatrix} -\gamma_1 \operatorname{div}_h \boldsymbol{\varepsilon}_h + \tau K_1^{-1} I_h & 0 & \dots & 0 \\ 0 & -\gamma_2 \operatorname{div}_h \boldsymbol{\varepsilon}_h + \tau K_2^{-1} I_h & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -\gamma_n \operatorname{div}_h \boldsymbol{\varepsilon}_h + \tau K_n^{-1} I_h \end{bmatrix}$$

$$- \begin{bmatrix} \tilde{\Lambda}_{11} \nabla_h \operatorname{div}_h & \tilde{\Lambda}_{12} \nabla_h \operatorname{div}_h & \dots & \tilde{\Lambda}_{1n} \nabla_h \operatorname{div}_h \\ \tilde{\Lambda}_{21} \nabla_h \operatorname{div}_h & \tilde{\Lambda}_{22} \nabla_h \operatorname{div}_h & \dots & \tilde{\Lambda}_{2n} \nabla_h \operatorname{div}_h \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\Lambda}_{n1} \nabla_h \operatorname{div}_h & \tilde{\Lambda}_{n2} \nabla_h \operatorname{div}_h & \dots & \tilde{\Lambda}_{nn} \nabla_h \operatorname{div}_h \end{bmatrix},$$

and

$$\mathcal{B}_{h,\boldsymbol{p}} = \begin{bmatrix} \Lambda_{11} I_h & \Lambda_{12} I_h & \dots & \Lambda_{1n} I_h \\ \Lambda_{21} I_h & \Lambda_{22} I_h & \dots & \Lambda_{2n} I_h \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{n1} I_h & \Lambda_{n2} I_h & \dots & \Lambda_{nn} I_h \end{bmatrix}.$$

5 Numerical Experiments

In this section we present numerical experiments whose results corroborate stability properties of the finite element discretization of the generalized Biot-Brinkman model (see Sect. 4.4) and the preconditioner (4.12). We shall first demonstrate parameter robustness of the exact preconditioner through a sensitivity study of the conditioning of the preconditioned Biot-Brinkman system. Afterwards, scalable realization of the preconditioner in terms multilevel methods for the displacement and flux blocks is discussed. For simplicity, all the experiments concern the domain $\Omega = (0, 1)^2$. The implementation was carried in the Firedrake finite element framework [41], which provides easy access to geometric multigrid solvers via the PCPATCH library [19].

5.1 Error Estimates

We consider a single network, $n = 1$, case of the generalized Biot-Brinkman model (3.5), with parameters $\mu = 1, \tau = 10^{-1}, \alpha_1 = 10^{-3}$ and $c_1 = 10^{-2}$ fixed (arbitrarily) while K_1, ν_1

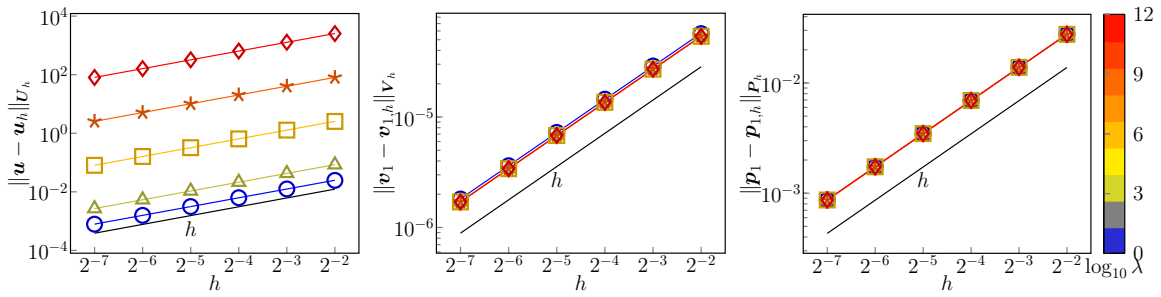


Fig. 1 Error approximation of the BDM_1-BDM_1-P_0 discretization of the single network Biot-Brinkman model. Parameters $\mu = 1$, $\tau = 10^{-1}$, $\alpha_1 = 10^{-3}$, $c_1 = 10^{-2}$, $\nu_1 = 1$ and $K_1 = 1$ are fixed. Line colors correspond to different values of λ (Color figure online)

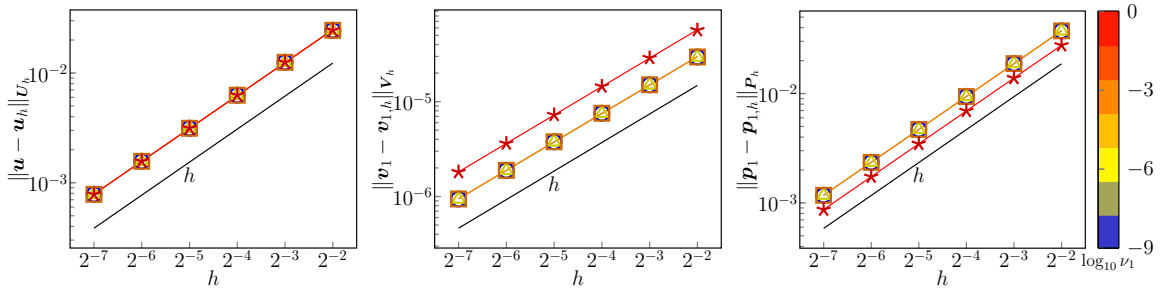


Fig. 2 Error approximation of the BDM_1-BDM_1-P_0 discretization of the single network Biot-Brinkman model. Parameters $\mu = 1$, $\tau = 10^{-1}$, $\alpha_1 = 10^{-3}$, $c_1 = 10^{-2}$, $K_1 = 1$ and $\lambda = 1$ are fixed. Line colors correspond to different values of ν_1 (Color figure online)

and λ shall be varied in order to test robustness of the error estimates established in Sect. 4.4. To this end, we solve (2.8) with the right hand side computed based on the exact solution

$$\boldsymbol{u} = \left(\frac{\partial \phi}{\partial y}, -\frac{\partial \phi}{\partial x}\right), \quad \boldsymbol{v}_1 = \nabla \phi_1, \quad p_1 = \sin \pi (x - y), \tag{5.1}$$

where

$$\phi = x^2(x-1)^2 y^2 (y-1)^2, \quad \phi_1 = x^4 (x-1)^4 y^4 (y-1)^4.$$

It can be seen that the manufactured solution satisfies the homogeneous conditions $\boldsymbol{u}|_{\partial\Omega} = \boldsymbol{0}$, $\boldsymbol{v}_1 \cdot \boldsymbol{n}|_{\partial\Omega} = 0$ for $\Omega = (0,1)^2$.

Using discretization by BDM_1 elements for $\boldsymbol{U}_h$, $\boldsymbol{V}_{1,h}$ and piece-wise constant elements for the pressure space $P_{1,h}$, Figs. 1 and 3 show the errors of the numerical approximations in the parameter-dependent norms (4.4), (4.5) and $\|\cdot\|_{\boldsymbol{P}}$ defined in (2.12c) when one of the parameters λ, K_1 and ν_1 is varied. In all the cases the expected linear convergence can be observed. In particular, the rate is independent of the parameter variations. We note that the error here is computed on a finer mesh than the finite element solution in order to prevent aliasing.

5.2 Robustness of Exact Preconditioner

We verify robustness of the canonical preconditioner (4.12) using a generalized Biot-Brinkman system with two networks. As the parameter space then counts 12 parameters in total we shall for simplicity fix material properties of one of the networks (below we choose the network $i = 1$) to unity in addition to setting $\mu = 1$, $\tau = 1$. This choice leaves

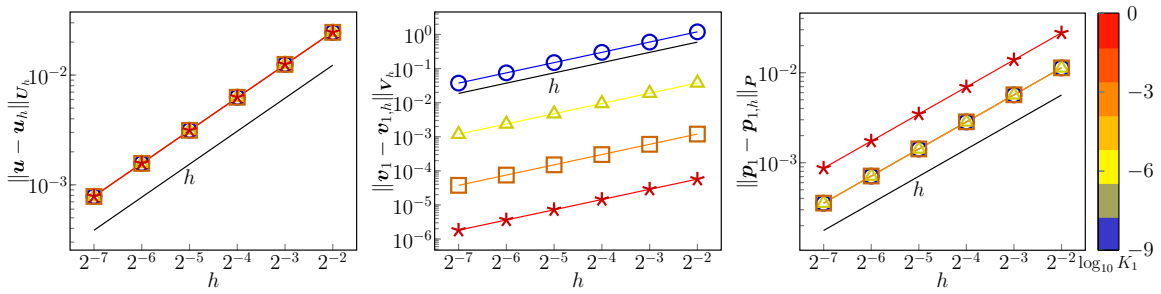


Fig. 3 Error approximation of the BDM_1-BDM_1-P_0 discretization of the single network Biot-Brinkman model. Parameters $\mu = 1$, $\tau = 10^{-1}$, $\alpha_1 = 10^{-3}$, $c_1 = 10^{-2}$, $\nu_1 = 1$ and $\lambda = 1$ are fixed. Line colors correspond to different values of K_1 (Color figure online)

parameters λ, c_2, α_2, ν_2, K_2 as well as the transfer coefficient $\beta := \beta_{12}$ to be varied. In the following experiments we let $1 \leq \lambda \leq 10^{12}$, $10^{-9} \leq \nu_2, K_2, \alpha_2 \leq 1$, $10^{-6} \leq \beta \leq 10^6$ and $c_2 \in \{0, 1\}$ in order to perform a systematic sensitivity study. We note that we do not vary directly the scaling parameters introduced in (3.5) but instead change the material parameters in (1.1).

For the above choice of parameters the two-network problem is considered on the domain $\Omega = (0, 1)^2$ with boundary conditions $\boldsymbol{u} = \boldsymbol{0}$ on the left and right sides and $(\boldsymbol{\sigma} + \boldsymbol{\alpha} \cdot \boldsymbol{pI}) \cdot \boldsymbol{n} = \boldsymbol{0}$ on the remaining part of the boundary; similarly, the Dirichlet conditions $\boldsymbol{v}_i \cdot \boldsymbol{n} = 0, i = 1, 2$ on the fluxes are prescribed only on the left and right sides.

Having constructed spaces $\boldsymbol{U}_h$, $\boldsymbol{V}_{1,h}$ $\boldsymbol{V}_{2,h}$ with BDM_1 elements and pressure spaces $P_{1,h}$ $P_{2,h}$ in terms of piece-wise constants our results are summarized in Figs. 4 and 6 where slices of the explored parameter space are shown. It can be seen that the condition numbers remain bounded. Concretely, given discrete operators $\mathcal{A}_h, \mathcal{B}_h$ that respectively discretize (3.5) and the preconditioner (4.12) the condition number is computed based on the generalized eigenvalue problem $\mathcal{A}_h x_k = \lambda_k \mathcal{B}_h^{-1} x_k$ as $\max_k |\lambda_k| / \min_k |\lambda_k|$. The higher condition numbers (of about 8.5) are typically attained when $c_2 = 0$, $\lambda = 1$ and $\beta \ll 1$. We remark that with $c_2 = 0$ and all parameters but β set to 1 the condition number of Λ ranges from 2.64 when $\beta = 10^{-6}$ to about 10^6 when $\beta = 10^6$.

5.3 Multigrid Preconditioning

Having seen that the exact preconditioner (4.12) yields parameter-robustness let us next discuss possible construction of a scalable approximation of the operator $\mathcal{B}$. Here, in order to approximate $\mathcal{B}_{\boldsymbol{u}}$ and $\mathcal{B}_{\boldsymbol{v}}$, we follow [2, 19, 22] and employ vertex-star relaxation schemes as part of geometric multigrid $F(2, 2)$-cycle for the elastic block and $W(2, 2)$-cycle for the flux block. Numerical experiments documenting robustness of the cycles for their respective blocks are reported in Appendix 1.

To test performance of the multigrid-based preconditioner $\mathcal{B}$ we consider the two-network system from Sect. 5.2 where we set $c_2 = 0$, $\alpha_2 = 1$, $\beta \in \{10^{-6}, 10^6\}$ while the remaining parameters are fixed to unity. We remark that for these parameter values the highest condition numbers are attained with the exact preconditioner, cf. Fig. 4. Furthermore, differing from the setup of the sensitivity study, we (strongly) enforce $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ and $\boldsymbol{v}_i \cdot \boldsymbol{n} = 0$, $i = 1, 2$, on

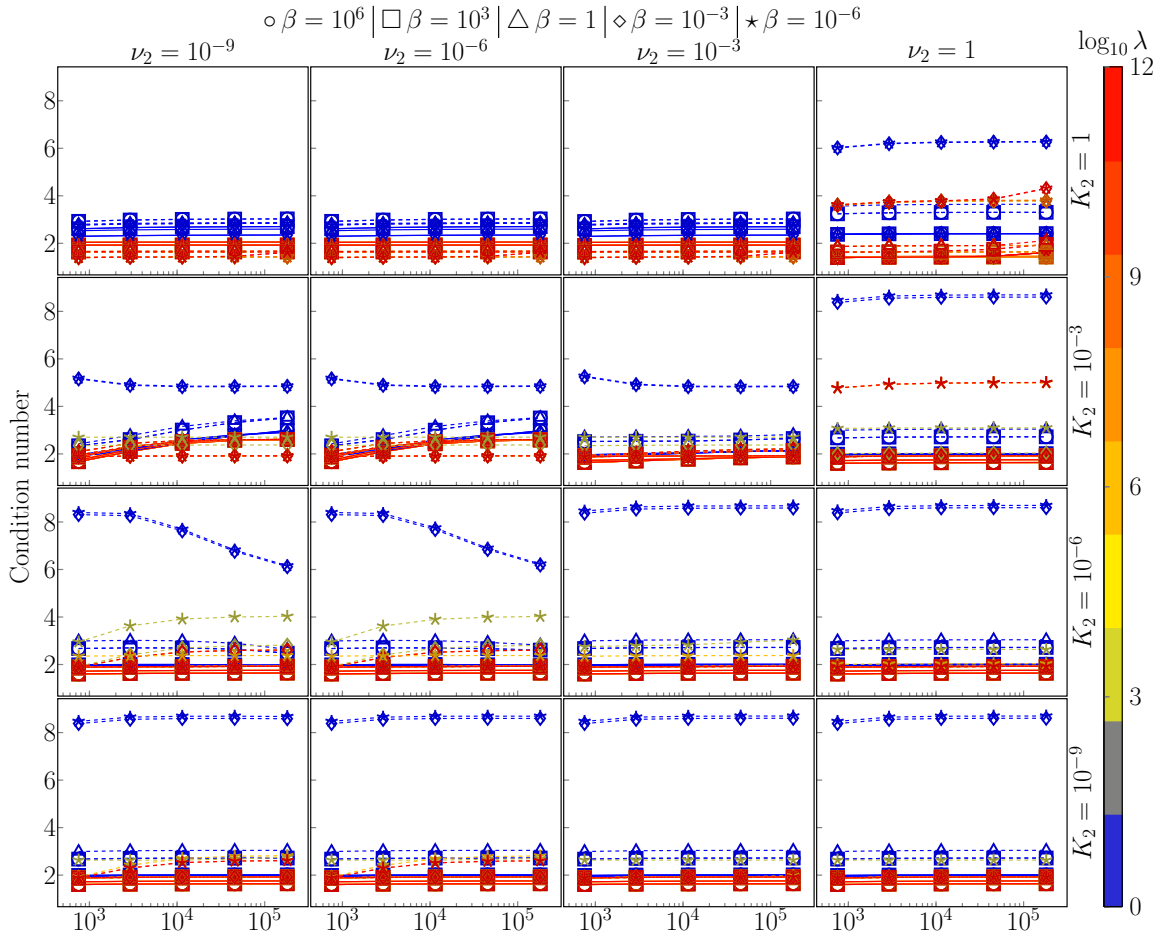


Fig. 4 Performance of Biot-Brinkman preconditioner (4.12) for $\alpha_2 = 1$ and varying parameters λ, ν_2, K_2, β (denoted by markers). Binary storage capacity is considered: $c_2 = 1$ (solid lines), $c_2 = 0$ (dashed lines). The remaining parameters are fixed at 1. Discretization by BDM_1-$(\mathrm{BDM}_1)^2$-$(\mathrm{P}_0)^2$ elements. Highest condition numbers correspond to $\beta \ll 1$ and $c_2 = 0$, $\lambda = 1$

the entire boundary.[1] As before, the finite element discretization is based on the BDM_1 and P_0 elements.

In Figs. 7 and 8 we report the dependence on the mesh size and parameter values of the iteration counts of the preconditioned MinRes solver where as the preconditioner both the exact Riesz map (4.12) and the multigrid-based approximation are used. More specifically, the multigrid cycles for the displacement and flux blocks use 3 grid levels applying the exact L^2-projection as the transfer operator. For both $\mathcal{B}_{\boldsymbol{u}}$ and $\mathcal{B}_{\boldsymbol{v}}$ the vertex-star relaxation uses damped Richardson smoother. Comparing the results we observe that the use of multigrid in (4.12) translates to a slight (about 1.5x) increase in the number of Krylov iterations compared to the exact preconditioner. However, the iterations appear bounded in the mesh size and the parameter variations.

We finally compare the cost of the exact and inexact Biot-Brinkman preconditioners for case $K_2 = 10^{-3}$, $\lambda = 1$, $\beta = 10^{-6}$ which required most iterations in the previous experiments, cf. Fig. 8. Our results are summarized in Table 1 and Fig. 9. We observe that

[1] The reason for not prescribing the complete displacement vector as a boundary condition are limitations in the PCPATCH framework which was used to implement the multigrid algorithm. In particular, the software currently lacks support for exterior facet integrals (see e.g. [3]) which are required with BDM elements to weakly enforce conditions on the tangential displacement by the Nitsche method.

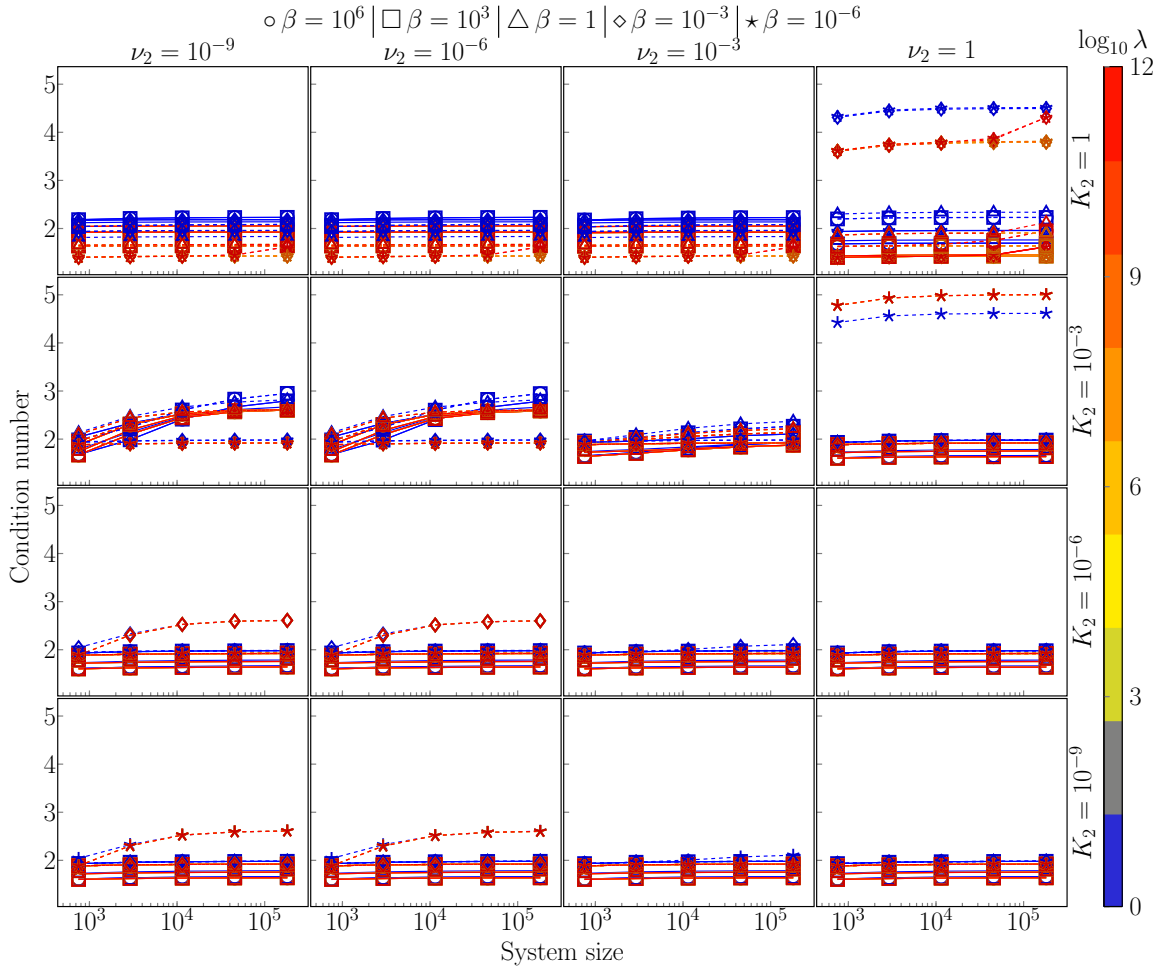


Fig. 5 Performance of Biot-Brinkman preconditioner (4.12) for $\alpha_2 = 10^{-4}$ and varying parameters λ, ν_2, K_2, β (denoted by markers). Binary storage capacity is considered: $c_2 = 1$ (solid lines), $c_2 = 0$ (dashed lines). The remaining parameters are fixed at 1. Discretization by BDM_1-$(\mathrm{BDM}_1)^2$-$(\mathrm{P}_0)^2$ elements

despite requiring more iterations for convergence the solution time[2] with the multigrid-based preconditioner is noticeably faster. We remark that for the sake of simple comparison the computations were done in serial using single-threaded execution. However, the latter setting is particularly unfavorable for the exact preconditioner $\mathcal{B}$ as modern LU solvers are known for their thread efficiency.

We note that the solver time and scaling properties are essentially determined by the method of computing action of blocks $\mathcal{B}_{\boldsymbol{u}}$ and $\mathcal{B}_{\boldsymbol{v}}$.

By using multigrid for the displacement and flux blocks the resulting solution algorithm appears to be order optimal, cf. Fig. 9. In particular, we observe that the computational time and memory usage of the solver scale linearly with the problem size.

Remark 5.1 Stabilization parameters enter in the discrete system operator, see (4.2), as well as in the preconditioner blocks $\mathcal{B}_{\boldsymbol{u}}$ and $\mathcal{B}_{\boldsymbol{v}}$ and have to be chosen large enough such that the related operators are positive definite, see e.g. [28].

In this study we have used the same value for all the penalty parameters, namely $\eta = 5$, and kept η fixed throughout all the experiments, in particular in the sensitity analysis. To illustrate the effect of the penalty parameter on performance of the Biot-Brinkman pre-

[2] The comparison is done in terms of the aggregate of the setup time of the linear system, the preconditioner and the run time of the Krylov solver.

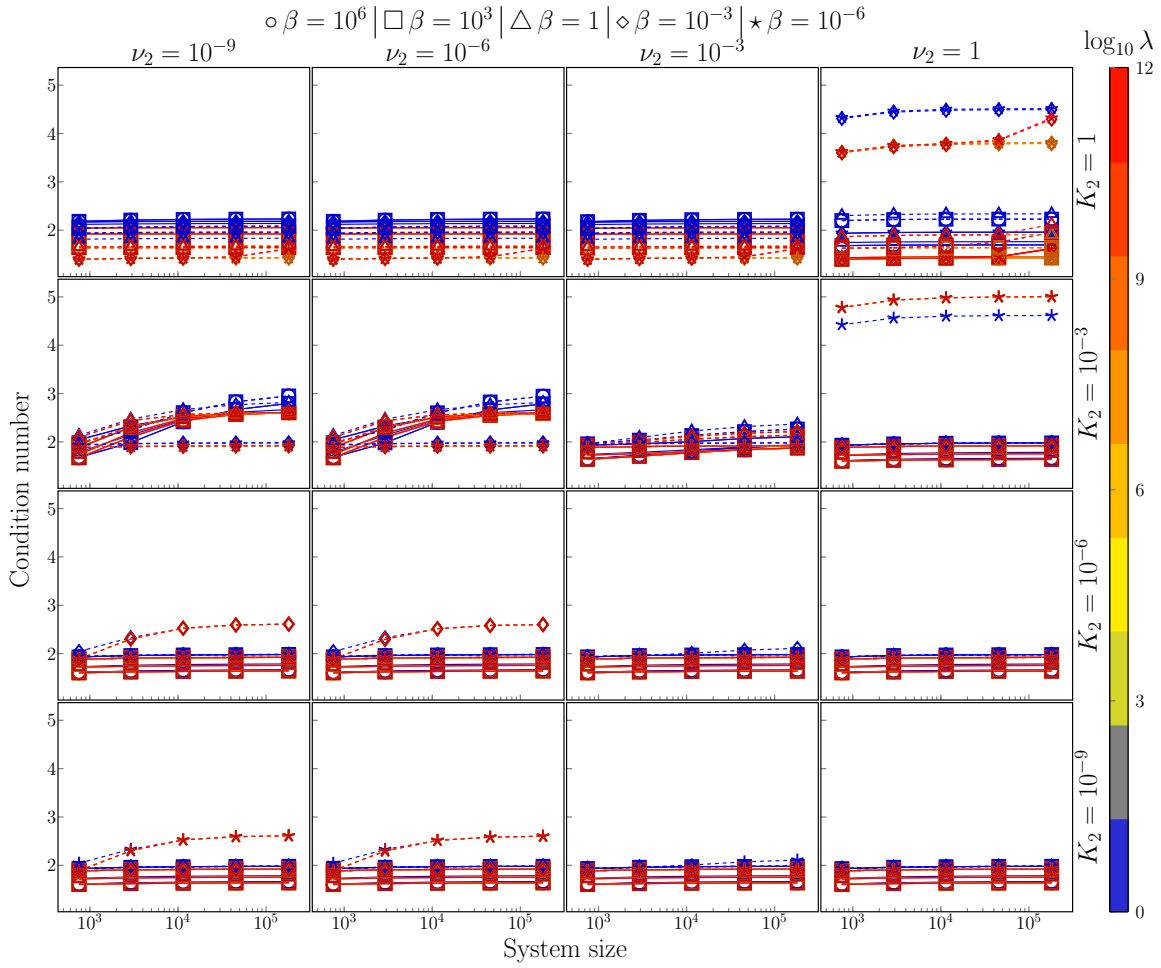


Fig. 6 Performance of Biot-Brinkman preconditioner (4.12) for $\alpha_2 = 10^{-8}$ and varying parameters λ, ν_2, K_2, β (denoted by markers). Binary storage capacity is considered: $c_2 = 1$ (solid lines), $c_2 = 0$ (dashed lines). The remaining parameters are fixed at 1. Discretization by BDM_1-$(\mathrm{BDM}_1)^2$-$(\mathrm{P}_0)^2$ elements

conditioner (4.12) we consider the experimental setup from Sect. 5.2. For simplicity all physical parameters shall be fixed at 1 while the Lamé parameter λ is varied together with $\eta = \{2, 5, 10, 20, 50, 100\}$. We remark that with $\eta = 1$ the MinRes solver failed to converge.

Considering the results shown in Fig. 10 it can be seen that all the values of η lead to iteration counts bounded in mesh size for both the exact preconditioner and the inexact one using geometric multigrid. However, the choice of η plays a role in the speed of convergence. In particular, larger values of the penalty seem to lead to larger iteration counts as can be seen from the performance of the multrigrid-based preconditioner (especially for $\lambda = 10^3, 10^6$). This effect is less pronounced with the exact preconditioner. We remark that the mulgrid preconditioners use the identical smoothing scheme (in particular the relaxation parameter of the Richardson smoother is fixed) for all the parameter values.

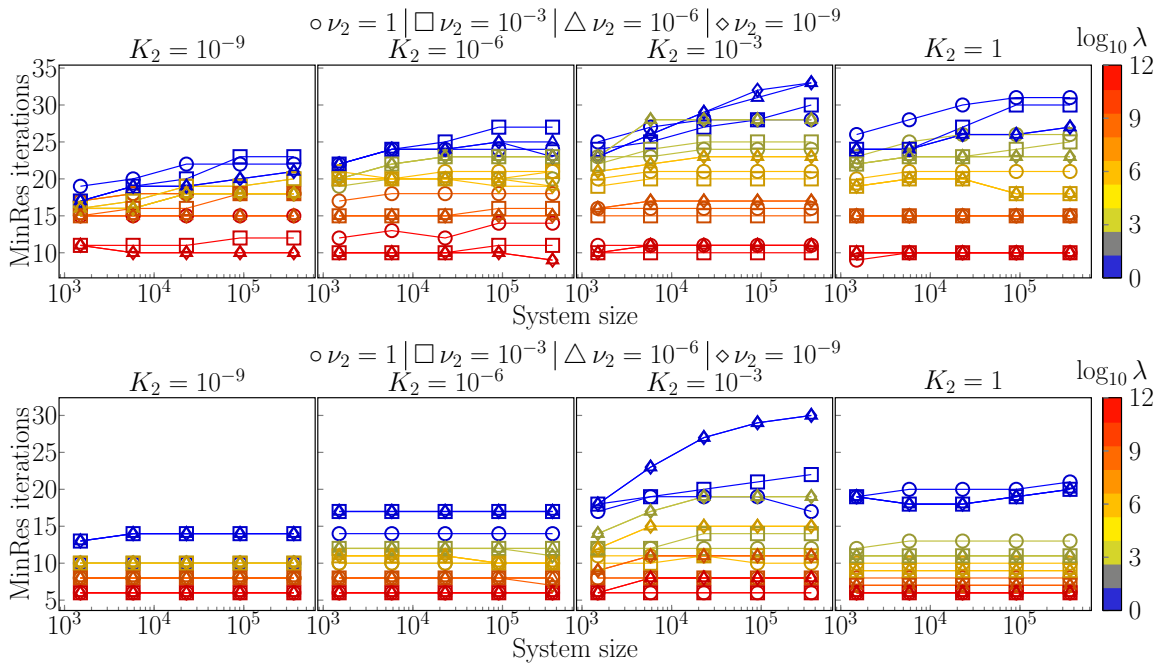


Fig. 7 Number of preconditioned MinRes iterations for 2-network Biot-Brinkman system with preconditioner (4.12). (Top) The displacement and flux blocks are realized by geometric multigrid while $\mathcal{B}_{\boldsymbol{p}}$ is computed by LU. (Bottom) Exact (LU-inverted) preconditioner is used. Transfer coefficient $\beta = 10^6$, while $c_2 = 0$, $\alpha_2 = 1$ and the remaining problem parameters are set to 1

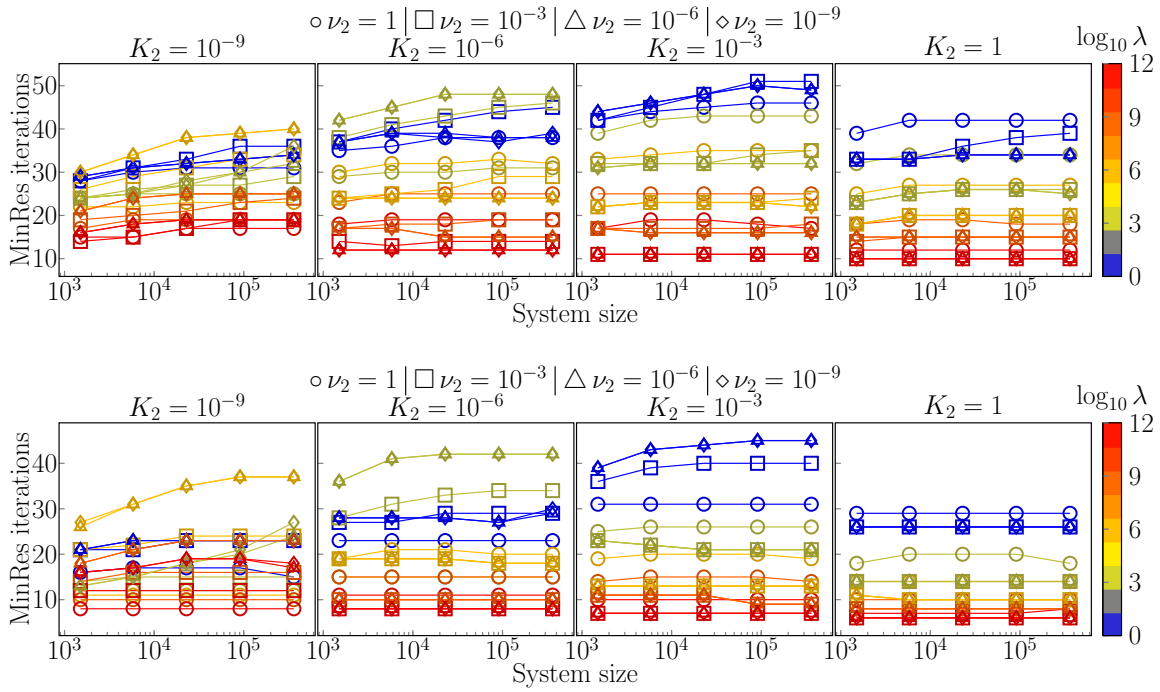


Fig. 8 Number of preconditioned MinRes iterations for 2-network Biot-Brinkman system with preconditioner (4.12). (Top) The displacement and flux blocks are realized by geometric multigrid while $\mathcal{B}_{\boldsymbol{p}}$ is computed by LU. (Bottom) Exact preconditioner is used. Transfer coefficient $\beta = 10^{-6}$, while $c_2 = 0$, $\alpha_2 = 1$ and the remaining problem parameters are set to 1

Table 1 Performance of exact (LU) and approximate multigrid-based (MG) preconditioners for the two-network generalized Biot-Brinkman model

	MinRes iterations LU				MinRes iterations MG			
ν_2/h	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-3}	2^{-4}	2^{-5}	2^{-6}
10^{-9}	43	44	45	45	46	48	50	49
10^{-6}	43	44	45	45	46	48	50	49
10^{-3}	39	40	40	40	45	48	51	51
1	31	31	31	31	44	45	46	46
	Solve time LU [s]				Solve time MG [s]			
ν_2/h	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-3}	2^{-4}	2^{-5}	2^{-6}
10^{-9}	2.50	4.64	23.18	181.00	4.47	7.05	18.22	64.29
10^{-6}	2.51	4.65	23.12	180.36	4.59	7.15	18.21	64.47
10^{-3}	2.50	4.64	23.06	180.34	4.57	7.05	18.24	65.45
1	2.51	4.57	22.74	178.84	4.45	6.94	17.63	62.83

Parameter ν_2 is varied while $c_2 = 0$, $K_2 = 10^{-3}$, $\beta = 10^{-6}$ and the remaining parameters are set to 1. Number of unknowns in the systems ranges from 6×10^3 to 362×10^3. Solve time aggregates setup time of the linear system, the preconditioner and the run time of the Krylov solver. Computations were done in serial with threading disabled by setting `OMP_NUM_THREADS=1`

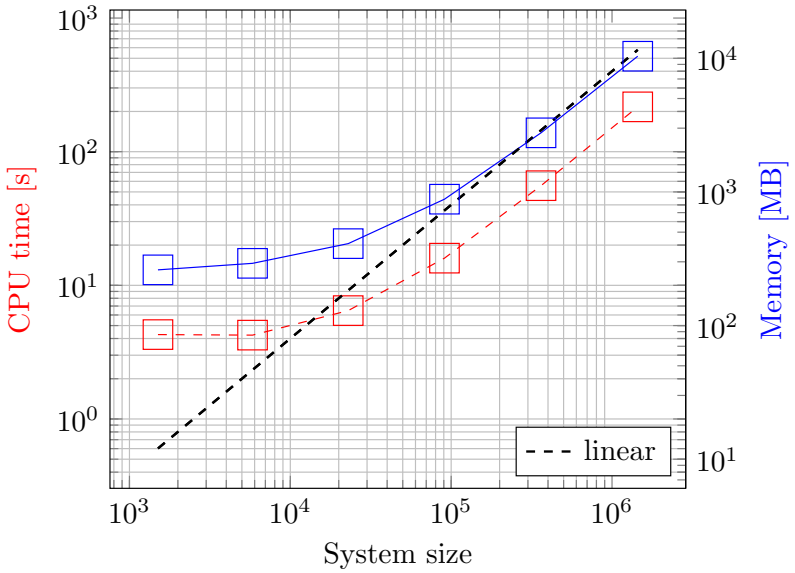


Fig. 9 Scaling of the two-network generalized Biot-Brinkman solver with preconditioner (4.12) using geometric multigrid for the displacement and flux blocks while $\mathcal{B}_p$ is computed by LU. Two dimensional setup from Table 1 is considered with $\beta = 10^{-6}$ and $\nu_1 = 10^{-3}$. Computations were done in serial with threading disabled by setting `OMP_NUM_THREADS=1`

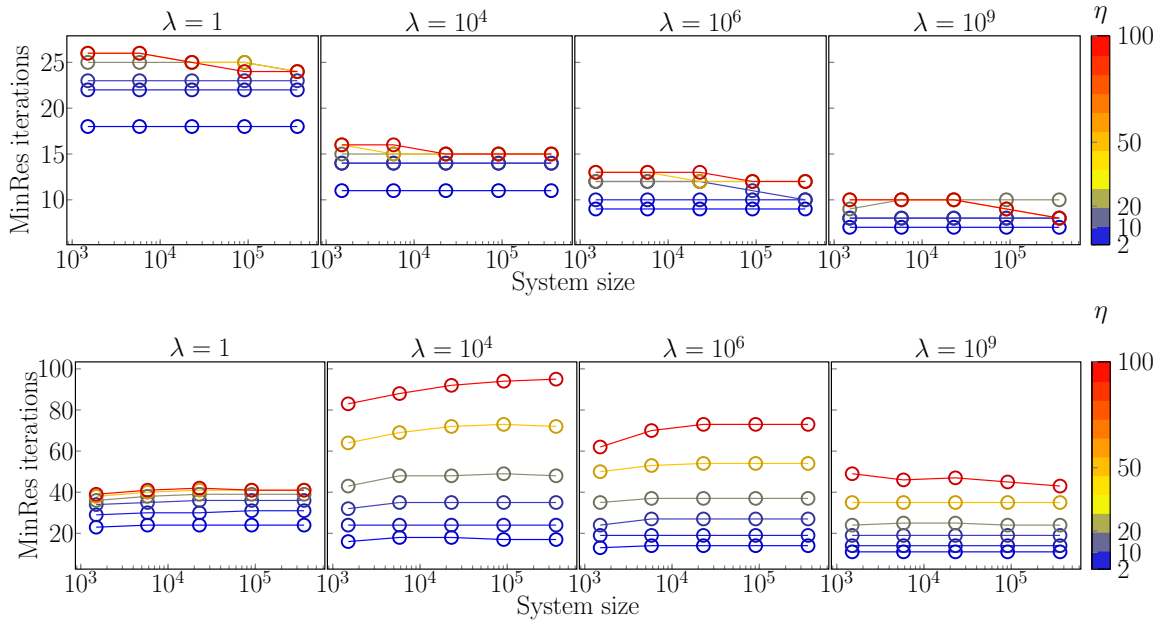


Fig. 10 Effect of stabilization parameter $\eta > 1$, see (4.2), in terms of number of preconditioned MinRes iterations with the Biot-Brinkman preconditioner (4.12). (Top) Exact (LU-inverted) preconditioner is considered. (Bottom) The displacement and flux blocks use geometric multigrid while $\mathcal{B}_p$ is computed by LU. All the physical parameters except for the Lamé parameter λ are kept constant at value 1

Acknowledgements J. Kraus and M. Lymbery acknowledge the support of this work by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the project "Physics-oriented solvers for multicompartmental poromechanics" under grant number 456235063. M. Kuchta acknowledges support from the Research Council of Norway (RCN) grant No 303362. K. A. Mardal acknowledges support from the Research Council of Norway, grant 300305 and 301013. M. E. Rognes has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement 714892.

Data Availability The datasets generated during and/or analyzed during the current study are available in the GitHub repository, https://github.com/MiroK/biot-brinkman-paper.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Appendix A. Components of Multigrid Preconditioner

In this section we report numerical experiments demonstrating robustness of geometric multigrid preconditioners for blocks $\mathcal{B}_u$ and $\mathcal{B}_v$ of the Biot-Brinkman preconditioner (4.12). Adapting the unit square geometry and the setup of boundary conditions from Sect. 5.3 we investigate performance of the preconditioners by considering boundedness of the (preconditioned) conjugate gradient (CG) iterations. In the following, the initial vector is set to 0 and the convergence of the CG solver is determined by reduction of the preconditioned residual norm by a factor 10^8. Finally, both systems are discretized by BDM_1 elements.

Table 2 confirms robustness of the $F(2, 2)$-cycle for the displacement block of (4.12). In particular, the iterations can be seen to be bounded in mesh size and the Lamé parameter λ.

Table 2 Number of preconditioned conjugate gradient iterations for approximating the displacement block $\mathcal{B}_{\boldsymbol{u}}$ of the Biot-Brinkman preconditioner

λ	$\log_2 h$					
	-3	-4	-5	-6	-7	-8
1	10	10	9	9	9	9
10^3	14	14	13	13	12	12
10^6	14	14	13	13	13	12
10^9	14	14	14	13	13	13
10^{12}	14	15	14	14	15	16

Geometric multigrid preconditioner uses $F(2, 2)$-cycle with 3 levels and a vertex-star (damped Richardson) smoother. In all experiments $\mu = 1$

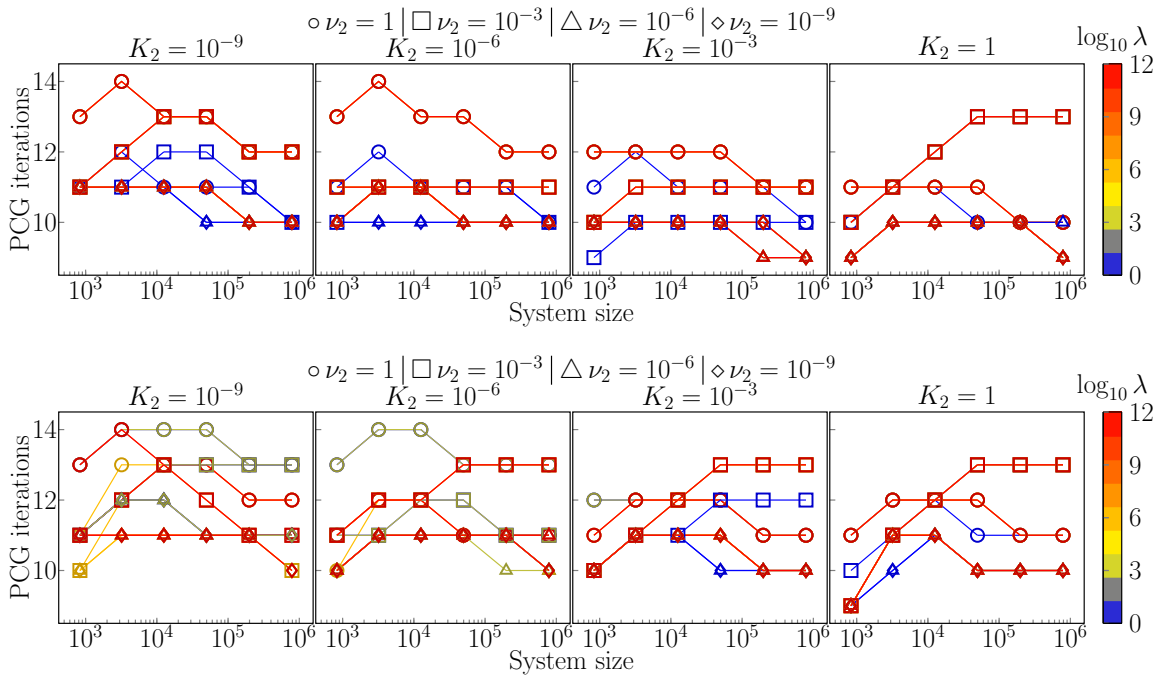


Fig. 11 Number of preconditioned conjugate gradient iterations for approximating the flux block $\mathcal{B}_{\boldsymbol{v}}$ of the Biot-Brinkman preconditioner. The preconditioner uses $W(2, 2)$-cycle of geometric multigrid with vertex-star (damped Richardson) smoother and 3 grid levels. (Top) Transfer coefficient $\beta = 10^6$, (bottom) $\beta = 10^{-6}$. Values of K_2, ν_2 (encoded by markers) and λ (encoded by line color) are varied. In both setups $c_2 = 0, \alpha_2 = 1$ and the remaining problem parameters are set to 1

For the flux block $\mathcal{B}_{\boldsymbol{v}}$ we limit the investigations to the two-network case and set $c_2 = 0$, $\alpha_2 = 1$ as these parameter values yielded the stiffest problems (in terms of their condition numbers) in the robustness study of Sect. 5.2. Performance of the geometric multigrid preconditioner using a $W(2, 2)$-cycle with vertex-star smoother is then summarized in Fig. 11. We observe that the number of CG iterations is bounded in the mesh size and variations in K_2, ν_2 and the exchange coefficient β. We remark that for some parameter configurations the observed dependence of the iteration counts is not monotone in mesh size. In particular, the number of preconditioned CG iterations on a finer mesh can be smaller than on a coarse one. However, in these cases the difference is 1 or 2 iterations with the former being the typical value.

References

1. Arnold, D.N., Brezzi, F., Cockburn, B., Marini, L.D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**, 1749–1779 (2002)
2. Arnold, D.N., Falk, R.S., Winther, R.: Multigrid in H(div) and H(curl). Numer. Math. **85**, 197–217 (2000)
3. Aznaran, F., Kirby, R., Farrell, P.: Transformations for Piola-mapped elements, arXiv preprint arXiv:2110.13224, (2021)
4. Bai, M., Elsworth, D., Roegiers, J.-C.: Multiporosity/multipermeability approach to the simulation of naturally fractured reservoirs. Water Resour. Res. **29**, 1621–1633 (1993)
5. Barenblatt, G., Zheltov, G., Kochina, I.: Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata]. J. Appl. Math. Mech. **24**(5), 1286–1303 (1960)
6. Barnafi, N., Zunino, P., Dedè, L., Quarteroni, A.: Mathematical analysis and numerical approximation of a general linearized poro-hyperelastic model. Comput. Math. Appl. **91**, 202–228 (2021)
7. Biot, M.: General theory of three-dimensional consolidation. J. Appl. Phys. **12**, 155–164 (1941)
8. Biot, M.: Theory of elasticity and consolidation for a porous anisotropic solid. J. Appl. Phys. **26**, 182–185 (1955)
9. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications, vol. 44. Springer, Heidelberg (2013)
10. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129–151
11. Brinkman, H.C.: A calculation of the viscous force exerted by a flowing fluid on a dense swarm of particles. Flow Turbul. Combust. **1**, 27–34 (1949)
12. Burtschell, B., Moireau, P., Chapelle, D.: Numerical analysis for an energy-stable total discretization of a poromechanics model with inf-sup stability. Acta Math. Appl. Sin. **35**, 28–53 (2019)
13. Chabiniok, R., Wang, V.Y., Hadjicharalambous, M., Asner, L., Lee, J., Sermesant, M., Kuhl, E., Young, A.A., Moireau, P., Nash, M.P., et al.: Multiphysics and multiscale modelling, data-model fusion and integration of organ physiology in the clinic: ventricular cardiac mechanics. Interface focus **6**, 20150083 (2016)
14. Chapelle, D., Moireau, P.: General coupling of porous flows and hyperelastic formulations–from thermodynamics principles to energy balance and compatible time schemes. Europ. J. Mech. B Fluids **46**, 82–96 (2014)
15. Chen, S., Hong, Q., Xu, J., Yang, K.: Robust block preconditioners for poroelasticity. Comput. Methods Appl. Mech. Eng. **369**, 113229 (2020)
16. Chou, D., Vardakis, J., Guo, L., Tully, B., Ventikos, Y.: A fully dynamic multi-compartmental poroelastic system: application to aqueductal stenosis. J. Biomech. **49**, 2306–2312 (2016)
17. Darcy, H.: Les fontaines publiques de la ville de Dijon: exposition et application..., Victor Dalmont, 1856
18. Eliseussen, E., Rognes, M.E., Thompson, T.B.: A-posteriori error estimation and adaptivity for multiple-network poroelasticity, arXiv preprint arXiv:2111.13456, (2021)
19. Farrell, P.E., Knepley, M.G., Mitchell, L., Wechsung, F.: PCPATCH: software for the topological construction of multigrid relaxation methods. ACM Trans. Math. Softw. **47**(3), 1–22 (2021)
20. Guo, L., Vardakis, J., Lassila, T., Mitolo, M., Ravikumar, N., Chou, D., Lange, M., Sarrami-Foroushani, A., Tully, B., Taylor, Z., Varma, S., Venneri, A., Frangi, A., Ventikos, Y.: Subject-specific multi-poroelastic model for exploring the risk factors associated with the early stages of Alzheimer's disease. Interface Focus **8**, 20170019 (2018)
21. Hansbo, P., Larson, M.: Discontinuous Galerkin and the Crouzeix-Raviart element application to elasticity. ESAIM Math. Modell. Numer. Anal. **37**(1), 63–72 (2003)
22. Hong, Q., Kraus, J.: Uniformly stable discontinuous Galerkin discretization and robust iterative solution methods for the Brinkman problem. SIAM J. Numer. Anal. **54**, 2750–2774 (2016)
23. Hong, Q., Kraus, J.: Parameter-robust stability of classical three-field formulation of Biot's consolidation model. Electron. Trans. Numer. Anal. **48**, 202–226 (2018)
24. Hong, Q., Kraus, J., Lymbery, M., Philo, F.: Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models. Numer. Linear Algebra Appl. **26**, e2242 (2019)
25. Hong, Q., Kraus, J., Lymbery, M., Philo, F.: Parameter-robust Uzawa-type iterative methods for double saddle point problems arising in Biot's consolidation and multiple-network poroelasticity models. Math. Models Methods Appl. Sci. **30**, 2523–2555 (2020)
26. Hong, Q., Kraus, J., Lymbery, M., Philo, F.: A new framework for the stability analysis of perturbed saddle-point problems and applications, arXiv:2103.09357v3 [math.NA], (2021)

27. Hong, Q., Kraus, J., Lymbery, M., Wheeler, M.F.: Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems. Multiscale Model. Simulat. **18**, 916–941 (2020)
28. Hong, Q., Kraus, J., Xu, J., Zikatanov, L.: A robust multigrid method for discontinuous Galerkin discretizations of Stokes and linear elasticity equations. Numer. Math. **132**, 23–49 (2016)
29. Hong, Q., Wang, F., Wu, S., Xu, J.: A unified study of continuous and discontinuous Galerkin methods. SCI. CHINA Math. **62**, 1–32 (2019)
30. Howell, J.S., Walkington, N.J.: Inf-sup conditions for twofold saddle point problems. Numer. Math. **118**, 663 (2011)
31. Kedarasetti, R., Drew, P.J., Costanzo, F.: Arterial vasodilation drives convective fluid flow in the brain: a poroelastic model, bioRxiv, (2021)
32. Khaled, M., Beskos, D., Aifantis, E.: On the theory of consolidation with double porosity. 3. a finite-element formulation, (1984), pp. 101–123
33. Kraus, J., Lederer, P.L., Lymbery, M., Schöberl, J.: Uniformly well-posed hybridized discontinuous Galerkin/hybrid mixed discretizations for Biot's consolidation model. Comput. Methods Appl. Mech. Engrg. **384**, 113991 (2021)
34. Kumar, S., Oyarzúa, R., Ruiz-Baier, R., Sandilya, R.: Conservative discontinuous finite volume and mixed schemes for a new four-field formulation in poroelasticity. ESAIM Math. Model. Numer. Anal. **54**, 273–299 (2020)
35. Lee, J., Mardal, K.-A., Winther, R.: Parameter-robust discretization and preconditioning of Biot's consolidation model. SIAM J. Sci. Comput. **39**, A1–A24 (2017)
36. Lee, J.J., Piersanti, E., Mardal, K.-A., Rognes, M.E.: A mixed finite element method for nearly incompressible multiple-network poroelasticity. SIAM J. Sci. Comput. **41**, A722–A747 (2019)
37. Mardal, K.-A., Rognes, M.E., Thompson, T.B.: Accurate discretization of poroelasticity without Darcy stability. BIT Numer. Math. **61**(3), 941–976 (2021)
38. Nash, M.P., Hunter, P.J.: Computational mechanics of the heart. J. Elast. Phys. Sci. solids **61**, 113–141 (2000)
39. Piersanti, E., Lee, J.J., Thompson, T., Mardal, K.-A., Rognes, M.E.: Parameter robust preconditioning by congruence for multiple-network poroelasticity. SIAM J. Sci. Comput. **43**, B984–B1007 (2021)
40. Rajagopal, K.R.: On a hierarchy of approximate models for flows of incompressible fluids through porous solids. Math. Models Methods Appl. Sci. **17**, 215–252 (2007)
41. Rathgeber, F., Ham, D.A., Mitchell, L., Lange, M., Luporini, F., Mcrae, A.T.T., Bercea, G.-T., Markall, G.R., Kelly, P.H.J.: Firedrake: automating the finite element method by composing abstractions. ACM Trans. Math. Softw. **43**(3), 1–27 (2016)
42. Rodrigo, C., Hu, X., Ohm, P., Adler, J.H., Gaspar, F.J., Zikatanov, L.: New stabilized discretizations for poroelasticity and the Stokes' equations. Comput. Methods Appl. Mech. Eng. **341**, 467–484 (2018)
43. Showalter, R.: Poroelastic filtration coupled to Stokes flow, Lecture Notes in Pure and Appl. Math. **242**, 229–241 (2010)
44. Støverud, K.H., Alnæs, M., Langtangen, H.P., Haughton, V., Mardal, K.-A.: Poro-elastic modeling of Syringomyelia-a systematic study of the effects of pia mater, central canal, median fissure, white and gray matter on pressure wave propagation and fluid movement within the cervical spinal cord. Comput. Methods Biomech. Biomed. Eng. **19**, 686–698 (2016)
45. Tully, B., Ventikos, Y.: Cerebral water transport using multiple-network poroelastic theory: application to normal pressure hydrocephalus. J. Fluid Mech. **667**, 188–215 (2011)
46. Vardakis, J., Chou, D., Tully, B., Hung, C., Lee, T., Tsui, P., Ventikos, Y.: Investigating cerebral oedema using poroelasticity. Med. Eng. Phys. **38**, 48–57 (2016)
47. Whitaker, S.: Flow in porous media I: a theoretical derivation of Darcy's law. Transp. Porous Med. **1**, 3–25 (1986)
48. Wilson, R., Aifantis, E.: On the theory of consolidation with double porosity. Int. J. Eng. Sci. **20**, 1009–1035 (1982)

C^1-CONFORMING VARIATIONAL DISCRETIZATION OF THE BIHARMONIC WAVE EQUATION

Computers and Mathematics with Applications 119 (2022) 208–219

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

www.elsevier.com/locate/camwa

C^1-conforming variational discretization of the biharmonic wave equation

Markus Bause [a], Maria Lymbery [b,*], Kevin Osthues [b]

[a] Helmut Schmidt University, Faculty of Mechanical Engineering, Holstenhofweg 85, 22043 Hamburg, Germany
[b] Faculty of Mathematics, University Duisburg-Essen, 45127 Essen, Germany

ARTICLE INFO

Keywords:
Biharmonic wave equation
Combined Galerkin and collocation technique
Optimal order error estimates
Bogner–Fox–Schmit element

ABSTRACT

Biharmonic wave equations are of importance to various applications including thin plate analyses. The innovation of this work comes through the numerical approximation of their solutions by a C^1-conforming in space and time finite element approach. Therein, the smoothness properties of solutions to the continuous evolution problem are embodied. Time discretization is based on a combined Galerkin and collocation technique. For space discretization the Bogner–Fox–Schmit element is applied. Optimal order error estimates are proven. The convergence and performance properties are illustrated by numerical experiments with complex wave profiles in homogeneous and heterogeneous media, illustrating that the approach offers high potential also for sophisticated multi-physics and/or multi-scale systems.

1. Introduction

In this work we propose and analyze a space-time finite element approximation by C^1-conforming in space and time discrete functions of the initial-boundary value problem for the biharmonic wave equation,

$$\partial_{tt}u(\boldsymbol{x},t)+\Delta^2 u(\boldsymbol{x},t)=f(\boldsymbol{x},t), \qquad \text{in } \Omega\times(0,T], \tag{1a}$$

$$u(\boldsymbol{x},0)=u_0(\boldsymbol{x}), \qquad \text{in } \Omega, \tag{1b}$$

$$\partial_t u(\boldsymbol{x},0)=u_1(\boldsymbol{x}), \qquad \text{in } \Omega, \tag{1c}$$

$$u(\boldsymbol{x},t)=0, \qquad \text{on } \partial\Omega\times(0,T], \tag{1d}$$

$$\partial_{\boldsymbol{n}} u(\boldsymbol{x},t)=0, \qquad \text{on } \partial\Omega\times(0,T], \tag{1e}$$

for a bounded domain $\Omega\subset\mathbb{R}^2$. This model is encountered in the modeling of various physical phenomena, such as plate bending and thin plate elasticity. The dynamic theory of thin Kirchhoff–Love plates investigates the propagation of waves in the plates as well as standing waves and vibration modes. Moreover, system (1) can be studied as a prototype model for more sophisticated Kirchhoff-type equations, such as the Euler–Bernoulli equation describing the deflection of viscoelastic plates.

The finite element discretization of fourth order differential operators in space has been subject to intensive research in the literature. The Bogner–Fox–Schmit (BFS) element [21] is a classical C^1-conforming thin plate element obtained by taking the tensor products of cubic Hermite splines. The discrete solutions are continuously differentiable on tensor product (rectangular) elements, which can be a serious drawback since it limits the applicability of the resulting finite element method. However, for geometries allowing tensor product discretization it is considered to be one of the most efficient elements for plate analysis, cf. [55, p. 153]. It is also a reasonably low order element for plates which is very simple to implement, in contrast with triangular elements which either use higher order polynomials, such as the Argyris element [11], or macro element techniques, such as the Clough–Tocher element [29]. Due to the appreciable advantages of the BFS element and our target to propose a C^1-conforming in space and time finite element approach for (1), the BFS element is applied here.

We note that the finite element approximation of the biharmonic operator continues to be an active field of research, for recent contributions, e.g. [25,26,30]. In particular, discretization methods that support polyhedral meshes (the mesh cells can be polyhedra or have a simple shape but contain hanging nodes) and hinge on the primal formulation of the biharmonic equation leading to a symmetric positive definite system matrix are currently focused. These methods can be classified into three groups, depending on the dimension of the smallest geometric object to which discrete unknowns are attached. This criterion influences the stencil of the method. Furthermore, it has an impact on the level of conformity that can be achieved for the discrete solution. The methods in the first group were developed for the case where $\Omega\subset\mathbb{R}^2$. They attach discrete unknowns to the mesh vertices, edges, and cells and can achieve C^1-conformity. Salient examples are the C^1-conforming virtual element methods (VEM) from [24,28,7,27,10] and the C^0-conforming

* Corresponding author.
E-mail addresses: bause@hsu-hh.de (M. Bause), maria.lymbery@uni-due.de (M. Lymbery), kevin.osthues@uni-due.de (K. Osthues).

https://doi.org/10.1016/j.camwa.2022.06.005
Received 12 July 2021; Received in revised form 16 February 2022; Accepted 8 June 2022

VEM from [53]. In [7] the proposed VEM for the Cahn–Hilliard equation has been also shown to be conforming in H^2 while using a very simple set of degrees of freedom. Another example is the nonconforming VEM from [23,54,9]. The methods in the second group attach discrete unknowns only to the mesh faces and cells for $\Omega \subset \mathbb{R}^d$, with $d \geq 2$. They admit static condensation, and provide a nonconforming approximation to the solution. Moreover, the nonconforming VEM in [9] has been demonstrated to be of arbitrary order of accuracy for biharmonic problems. The two salient examples are the weak Galerkin methods from [46,52,51] and the hybrid high-order method from [22]. Finally, the methods in the third group attach discrete unknowns only to the mesh cells and belong to the class of interior penalty discontinuous Galerkin methods. These are also nonconforming methods; cf. [45,47,34]. Important examples of nonconforming finite elements on simplicial meshes are the Morley element [44,48] and the Hsieh–Clough–Tocher element (cf., e.g., [54, Chap. 6]). The spatial discretization of wave problems by discontinuous Galerkin methods has been focused further in the literature, cf., e.g., [35,8].

Among the most attractive methods for time discretization of second-order differential equations in time are the so-called *continuous Galerkin or Galerkin–Petrov* (cf., e.g., [12,33,40]) and the *discontinuous Galerkin* (cf., e.g., [37,40]) schemes. For lowest order elements, these methods can be identified with certain well-known difference schemes, e.g. with the classical trapezoidal Newmark scheme (cf., e.g., [49,50,36]), the backward Euler scheme and the Crank–Nicolson scheme. Strong relations and equivalences between variational time discretizations, collocation methods and Runge–Kutta schemes have been observed. In the literature, the relations are exploited in the formulation and analysis of the schemes. For this we refer to, e.g., [2,3]. Recently, variational time discretizations that lead to discrete solutions of higher order regularity in time [6,16] have been devised for the second-order hyperbolic wave equations and analyzed carefully. In particular, optimal order error estimates are proved in [6,16]. In [16], a C^1-conforming in time family of space-time finite element approximation that is based on a post-processing of the continuous in time Galerkin approximation is introduced. Concepts that are developed in [31] for first-order hyperbolic problems are transferred to the wave equation written as a first order system in time. In [16], a family of Galerkin–collocation approximation schemes with C^1- and C^2-regular in time discrete solutions are proposed and investigated by an optimal order error analysis and computational experiments. The conceptual basis of the families of approximations to the wave equation is the establishment of a connection between the Galerkin method for the time discretization and the classical collocation methods, with the perspective of achieving the accuracy of the former with reduced computational costs provided by the latter in terms of less complex algebraic systems. Further numerical studies for the wave equation can be found in [14,5]. For the application of the Galerkin–collocation to mathematical models of fluid flow and systems of ordinary differential equations we refer to [4,19,20]. In the numerical experiments, the Galerkin–collocation schemes have proved their superiority over lower-order and standard difference schemes. In particular, energy conservation is ensured which is an essential feature for discretization schemes to second-order hyperbolic problems since the physics of solutions to the continuous problem are preserved.

As a logical consequence, for the biharmonic wave problem (1) it appears to be promising to combine the Galerkin–collocation time discretization with the BFS finite element discretization of the spatial variables to a C^1-conforming approximation in space and time. This conceptionally new idea is applied here and yields an innovation and superiority over existing discretization schemes. The latter is shown by computational comparative studies in Sec. 6. The higher regularity of the discrete solutions can be of special advantage for future applications in multi-physics, for instance, if terms coupling subproblems depend on temporal or spatial derivatives of the solution to the respective other subsystem. Further, such space-time finite element approaches may lead to progress in the development of multi-scale approaches in space and time. Finally, such families of schemes allow the application of space-time adaptive methods based on the dual residual approach for goal-oriented error control. For the latter one, we refer to [13,15,41] for parabolic problems. In this work, we present the combined Galerkin–collocation and BFS finite element approximation of (1). Key ingredients of the construction of the Galerkin–collocation approach are the application of a special quadrature formula, proposed in [38], and the definition of a related interpolation operator for the right-hand side term of the variational equation. Both of them use derivatives of the given function. The Galerkin–collocation scheme relies in an essential way on the perfectly matching set of the polynomial spaces (trial and test space), quadrature formula, and interpolation operator. Then, a numerical error analysis is performed, optimal order error estimates are proved. Here, we restrict ourselves to presenting and stressing the differences to the wave equation for the Laplacian considered in [6]. Finally, a numerical study of the proposed discretization scheme is presented in order to illustrate the analyses.

This paper is organized as follows. In Section 2, we introduce our notation and formulate problem (1) as a first-order system in time. In Section 3, the Galerkin–collocation method is considered for time discretization. Some beneficial results for the error analysis are summarized in Section 4. In Section 5, we prove error estimates for the introduced Galerkin–collocation method for the plate vibration problem (1). Finally, in Section 6 we present numerical experiments confirming the error estimates and perform a comparative study by evaluating the gain in accuracy of the proposed full C^1-conforming approximation scheme over only continuous in time approximations.

2. Preliminaries and notation

2.1. Evolution form

Throughout this paper, standard notation is used for Sobolev and Bochner spaces. By B we denote a Banach space. We use $(\cdot,\cdot)$ for the $L^2(\Omega)$ inner product inducing the norm

$$\|\cdot\| = \|\cdot\|_{L^2(\Omega)}$$

and $\langle\cdot,\cdot\rangle$ for the duality pairing between a Hilbert space and its dual space. For the Sobolev norms we adopt the notation

$$\|\cdot\|_m = \|\cdot\|_{H^m(\Omega)} \quad \text{for } m \in \mathbb{N},\ m \geq 1$$

and further define the spaces

$$H = L^2(\Omega) \quad \text{and} \quad V = H_0^2(\Omega).$$

Let V' be the dual space of V. We introduce the operator $A : V \to V'$ which for any given $u \in V$ is uniquely defined by

$$\langle Au, v\rangle = (\Delta u, \Delta v) \qquad \forall v \in V$$

and also the operator $\mathcal{L} : V \times H \to H \times V'$ given by

$$\mathcal{L} = \begin{pmatrix} 0 & -I \\ A & 0 \end{pmatrix}.$$

Here I is the identity operator that acts on H. For the error analysis, we define the energy norm on $H_0^2(\Omega) \times L^2(\Omega)$ by $|||(w_0, w_1)|||^2 = \|\Delta w_0\|^2 + \|w_1\|^2$.

In order to formulate problem (1) in an evolutionary form we further introduce the space

$$X := L^2(0,T;V) \times L^2(0,T;H)$$

and set

$$u^0 = u \qquad \text{and} \qquad u^1 = \partial_t u.$$

With this notation then problem (1) can be equivalently stated as: Find $U = (u^0, u^1) \in X$ satisfying

$$\partial_t U + \mathcal{L}U = F \qquad \text{in} \quad (0,T) \tag{2a}$$

$$U(0) = U_0 \tag{2b}$$

where $f \in L^2(0,T;H)$ is given, $F = (0,f)$ and $U_0 = (u_0, u_1)$.

The existence and uniqueness of solutions to (2) is a classical result, cf. [42, p. 273, Thm. 1.1], and [43, p. 275, Thm. 8.2]. Further, we have the following regularity result $H^2(\Omega) \subset\subset C(\overline{\Omega})$ for $\Omega \subset \mathbb{R}^2$, cf. [1].

2.2. Time discretization

Our aim is to replace the time interval with a discrete time mesh and subsequently to iteratively compute the solution of (1) in the time nodes. For this reason, we split $I = (0,T]$ into $N \in \mathbb{N}$ time subintervals

$$I_n = \left(t_{n-1}, t_n\right], \qquad n = 1, \dots, N,$$

where $0 = t_0 < t_1 < \cdots < t_N = T$ and introduce the time step parameter $\tau = \max_{n=1,\dots,N} \tau_n$, where $\tau_n = t_n - t_{n-1}$. The set $\mathcal{M}_\tau := \{I_1, \dots, I_N\}$ of time intervals represents the time mesh. For simplicity, we use $I_0 = \{t_0\}$.

We denote the space of all B-valued polynomials in time of order $k \in \mathbb{N}_0$ over a given interval I_n by

$$\mathbb{P}_k(I_n; B) = \left\{ w_\tau : I_n \to B : w_\tau(t) = \sum_{j=0}^{k} W^j t^j,\ \forall t \in I_n,\ W^j \in B\ \forall j \right\}.$$

Moreover, for an integer $k \in \mathbb{N}$ we introduce the space of globally continuous functions in time, $X_\tau^k(B)$, and the space of global L^2-functions in time, $Y_\tau^k(B)$, as follows

$$X_\tau^k(B) := \left\{ w_\tau \in C(\bar{I}; B) : w_\tau\big|_{I_n} \in \mathbb{P}_k(I_n; B)\ \forall I_n \in \mathcal{M}_\tau \right\},$$

$$Y_\tau^k(B) := \left\{ w_\tau \in L^2(I; B) : w_\tau\big|_{I_n} \in \mathbb{P}_k(I_n; B)\ \forall I_n \in \mathcal{M}_\tau \right\}.$$

We designate

$$\partial_t^s w(t_n^+) := \lim_{t \to t_n^{+0}} \partial_t^s w(t) \quad \text{and} \quad \partial_t^s w(t_n^-) := \lim_{t \to t_n^{-0}} \partial_t^s w(t)$$

to be the one-sided limits of the s-th derivative of a piecewise sufficiently smooth with respect to the time mesh $\mathcal{M}_\tau$ function $w : I \to B$ where $s \in \mathbb{N}_0$.

3. Discretizations of space and time

3.1. Space discretization

Let $\mathcal{T}_h$ be a shape-regular mesh of the spatial domain Ω with $h > 0$ denoting the mesh size and let

$$V_h = \left\{ v_h \in C^1(\Omega) : v_h\big|_T \in \mathbb{Q}_3(T), v_h\big|_{\partial\Omega} = 0,\ \partial_{\boldsymbol{n}} v_h\big|_{\partial\Omega} = 0\ \forall T \in \mathcal{T}_h \right\}$$

be the finite element space built on the mesh using the Bogner–Fox–Schmit element. Here $\mathbb{Q}_3(T)$ denotes the set of all polynomials with maximum degree 3 in each variable.

We denote the L^2-orthogonal projection onto V_h by P_h, i.e.,

$$(P_h w, v_h) = (w, v_h) \qquad \forall v_h \in V_h$$

and define the elliptic operator $R_h : V \to V_h$ via

$$(\Delta R_h w, \Delta v_h) = (\Delta w, \Delta v_h) \qquad \forall v_h \in V_h. \tag{3}$$

For $w \in H^s \cap H_0^2$ we have the estimates

$$\|w - R_h w\|_m \le C h^{4-m} \|w\|_4, \qquad 0 \le m \le 3 \tag{4}$$

and

$$\|\Delta(w - R_h w)\|_m \le C h^{2-m} \|w\|_4, \qquad 0 \le m \le 1$$

which follow directly from the interpolation error estimates given in [25] along with Cea's lemma and the Aubin–Nitsche trick.

Further, we introduce the L^2-projection $\mathcal{P}_h : H \times H \to V_h \times V_h$ and the elliptic projection $\mathcal{R}_h : V \times V \to V_h \times V_h$ both of which are onto the product space $V_h \times V_h$ and also the discrete operator $A_h : V \to V_h$ for which it holds

$$\left(A_h w, v_h\right) = \left(\Delta w, \Delta v_h\right) \qquad \forall v_h \in V_h. \tag{5}$$

Therefore, if $w \in V \cap H^4(\Omega)$, we have

$$\left(A_h w, v_h\right) = \left(\Delta w, \Delta v_h\right) = \langle A w, v_h \rangle \qquad \forall v_h \in V_h$$

or

$$A_h w = P_h A w \qquad \text{for } w \in V \cap H^4(\Omega).$$

Moreover, for the operator $\mathcal{L}_h : V \times H \to V_h \times V_h$ defined as

$$\mathcal{L}_h = \begin{pmatrix} 0 & -P_h \\ A_h & 0 \end{pmatrix} \tag{6}$$

the following relation holds

$$\begin{aligned}\left(\mathcal{L}_h W, \Phi_h\right) &= \left(-w^1, \varphi_h^0\right) + \left(\Delta w^0, \Delta \varphi_h^1\right) = \left(-w^1, \varphi_h^0\right) + \langle A w^0, \varphi_h^1 \rangle \\ &= \langle \mathcal{L} W, \Phi_h \rangle\end{aligned}$$

for $W = (w^0, w^1) \in (V \cap H^4(\Omega)) \times H$ and for all $\Phi_h = (\varphi_h^0, \varphi_h^1) \in V_h \times V_h$ which demonstrates the consistency of $\mathcal{L}_h$ on $(V \cap H^4(\Omega)) \times H$, i.e.,

$$\mathcal{L}_h W = \mathcal{P}_h \mathcal{L} W. \tag{7}$$

Finally, an appropriate approximation in $V_h \times V_h$ of the initial value $U_0 \in V \times H$ is denoted by $U_{0,h}$.

3.2. Numerical integration

The following makes use of the Hermite-type, Gauss and Gauss-Lobatto quadrature formulas which for a sufficiently regular function g on the interval $\bar{I}_n = [t_{n-1}, t_n]$ read as

$$Q_n^H(g) = \left(\frac{\tau_n}{2}\right)^2 \hat{w}_L^H \partial_t g(t_{n-1}^+) + \frac{\tau_n}{2} \sum_{s=1}^{k-1} \hat{w}_s^H g(t_{n,s}^H) + \left(\frac{\tau_n}{2}\right)^2 \hat{w}_R^H \partial_t g(t_n^-), \tag{8a}$$

$$Q_n^G(g) = \frac{\tau_n}{2} \sum_{s=1}^{k-1} \hat{w}_s^G g(t_{n,s}^G), \tag{8b}$$

$$Q_n^{GL}(g) = \frac{\tau_n}{2} \sum_{s=1}^{k} \hat{w}_s^{GL} g(t_{n,s}^{GL}), \tag{8c}$$

respectively. Here, $t_{n,s}^H$, $t_{n,s}^G$ and $t_{n,s}^{GL}$ are the corresponding quadrature points on the interval while $\{\hat{w}_L^H, \hat{w}_R^H, \hat{w}_s^H\}$, $\hat{w}_s^G$ and $\hat{w}_s^{GL}$ denote the corresponding weights.

We also consider the global Hermite interpolation $I_\tau^H : C^1(\bar{I}; B) \to X_\tau^k(B)$ defined as

$$I_\tau^H w\big|_{I_n} := I_n^H(w\big|_{I_n}) \tag{9}$$

for all $n = 1, \dots, N$ where $I_n^H : C^1(\bar{I}_n; B) \to \mathbb{P}_k(\bar{I}_n; B)$ denotes the local Hermite interpolation operator with respect to point values and first derivatives on the interval $\bar{I}_n$.

3.3. Space-time discretizations

In this subsection we introduce the discretization of the biharmonic wave problem (1) by a space-time finite element approach utilizing a Galerkin–collocation approximation (cf. [6]) of the time variable along

with BFS element for the approximation in space. The time discretization combines Galerkin and collocation techniques. Moreover, for comparative studies and in order to analyze the impact of the discrete solution's higher regularity in time on the accuracy of the numerical results, the standard continuous in time Galerkin–Petrov approach (cf., e.g., [33,16]) is presented here briefly. Within the latter family of schemes, the Crank–Nicolson method is recovered for piecewise linear approximations.

3.3.1. The Galerkin–collocation method cGP-C$^1(k)$

The variational time discretization for the plate vibration problem (1) is derived following the idea in [18,6] and reads as follows:

Problem 1. Let $U_{\tau,h}(t_{n-1}^-)$ for $n>1$ and $U_{\tau,h}(t_0^-)=U_{0,h}$ for $n=1$ be given. Find $U_{\tau,h}\big|_{I_n} \in (\mathbb{P}_k(I_n;V_h))^2$ satisfying

$$U_{\tau,h}(t_{n-1}^+) = U_{\tau,h}(t_{n-1}^-), \tag{10a}$$

$$\partial_t U_{\tau,h}(t_{n-1}^+) = -\mathcal{L}_h U_{\tau,h}(t_{n-1}^+) + \mathcal{P}_h F(t_{n-1}^+), \tag{10b}$$

$$\partial_t U_{\tau,h}(t_n^-) = -\mathcal{L}_h U_{\tau,h}(t_n^-) + \mathcal{P}_h F(t_n^-), \tag{10c}$$

$$Q_n^H\big(\big(\partial_t U_{\tau,h}, V_{\tau,h}\big) + \big(\mathcal{L}_h U_{\tau,h}, V_{\tau,h}\big)\big) = Q_n^H\big(\big(F, V_{\tau,h}\big)\big) \tag{10d}$$

for all $V_{\tau,h} \in (\mathbb{P}_{k-3}(I_n;V_h))^2$.

The existence and uniqueness of a solution to Problem 1 has been discussed in [6], see also [18].

From the definition of the scheme it also follows that $U_{\tau,h} \in (C^1(\bar{I};V_h))^2$ and (10b) can be written as

$$\partial_t U_{\tau,h}(t_{n-1}^+) = \partial_t U_{\tau,h}(t_{n-1}^-),$$

where $\partial_t U_{\tau,h}(t_0^-) = -\mathcal{L}_h U_{0,h} + \mathcal{P}_h F(0)$ and, therefore, Problem 1 can be equivalently written as:

Problem 2. Let $k \geq 3$ be fixed and be given the values $\Big(u^0_{\tau,h}\big|_{I_{n-1}}(t_{n-1}), u^1_{\tau,h}\big|_{I_{n-1}}(t_{n-1})\Big) \in V_h^2$ for $1 < n \leq N$ and $\Big(u^0_{\tau,h}\big|_{I_0}(t_0), u^1_{\tau,h}\big|_{I_0}(t_0)\Big) = (u_{0,h}, u_{1,h})$ for $n=1$. Then the Galerkin–collocation for $\Big(u^0_{\tau,h}\big|_{I_n}, u^1_{\tau,h}\big|_{I_n}\Big) \in \mathbb{P}_k(I_n;V_h)^2$ is defined as

$$\partial_t^s u^0_{\tau,h}\big|_{I_n}(t_{n-1}) = \partial_t^s u^0_{\tau,h}\big|_{I_{n-1}}(t_{n-1}), \quad s\in\{0,1\}, \tag{11a}$$

$$\partial_t^s u^1_{\tau,h}\big|_{I_n}(t_{n-1}) = \partial_t^s u^1_{\tau,h}\big|_{I_{n-1}}(t_{n-1}), \quad s\in\{0,1\}, \tag{11b}$$

$$\partial_t u^0_{\tau,h}\big|_{I_n}(t_n) - u^1_{\tau,h}\big|_{I_n}(t_n) = 0, \tag{11c}$$

$$\partial_t u^1_{\tau,h}\big|_{I_n}(t_n) + A_h u^0_{\tau,h}\big|_{I_n}(t_n) = f(t_n), \tag{11d}$$

and

$$\int_{I_n}\int_{\Omega} \partial_t u^0_{\tau,h}\,\varphi^0_{\tau,h} - u^1_{\tau,h}\,\varphi^0_{\tau,h}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t = 0, \tag{11e}$$

$$\int_{I_n}\int_{\Omega} \partial_t u^1_{\tau,h}\,\varphi^1_{\tau,h} + A_h u^0_{\tau,h}\,\varphi^1_{\tau,h}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t = \int_{I_n}\int_{\Omega} f\,\varphi^1_{\tau,h}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t, \tag{11f}$$

for all $\Big(\varphi^0_{\tau,h}, \varphi^1_{\tau,h}\Big) \in (\mathbb{P}_{k-3}(I_n;V_h))^2$.

The discrete initial values $(u_{0,h}, u_{1,h}) \in V_h^2$ are determined from the interpolation of the functions (u_0, u_1). We use the interpolant $u_{1,h}$ for the value $\partial_t u^0_{\tau,h}(0)$ from (11a). In order to obtain an appropriate value for $\partial_t u^1_{\tau,h}(0)$ in (11b) we consider (1a) at time $t=0$ and interpolate the function

$$\partial_t u^1(\boldsymbol{x},0) = f(\boldsymbol{x},0) - \Delta^2 u(\boldsymbol{x},0).$$

The collocation conditions ensure a reduction in the size of the test space which results in a smaller linear system of equations.

Next, we summarize a result for scheme (10) used in the error analysis, cf. [6].

Proposition 3. *Consider the solution $U_{\tau,h} \in (X_\tau^k(V_h))^2$ of Problem 1. It holds that*

$$B_n^{GL}(U_{\tau,h}, V_{\tau,h}) = Q_n^{GL}\big(\big(I_\tau^H F, V_{\tau,h}\big)\big)$$

for all $V_{\tau,h} \in (\mathbb{P}_{k-2}(I_n;V_h))^2$ and for $n=1,\dots,N$, where

$$B_n^{GL}(U_{\tau,h}, V_{\tau,h}) = Q_n^{GL}\big(\big(\partial_t U_{\tau,h}, V_{\tau,h}\big) + \big(\mathcal{L}_h U_{\tau,h}, V_{\tau,h}\big)\big).$$

3.3.2. The cGP(k)-method

The second time discretization method for the dynamic plate vibration problem considered is the Crank–Nicolson method [32]. The differential equation is solved iteratively by determining the solution at the time interval points t_n. To derive the Crank–Nicolson method, we use $\Big(u^0_{\tau,h}\big|_{I_n}, u^1_{\tau,h}\big|_{I_n}\Big) \in \mathbb{P}_k(\bar{I}_n;V_h)^2$ and $\Big(\varphi^0_{\tau,h}\big|_{I_n}, \varphi^1_{\tau,h}\big|_{I_n}\Big) \in \mathbb{P}_{k-1}(\bar{I}_n;V_h)^2$ as ansatz [6]. Since the solution space differs from the test space, this is referred to as a continuous Galerkin–Petrov method, or cGP(k) for short. The discrete solution functions are globally continuous and use piecewise polynomials of degree k for the time discretization. The cGP(1)-method corresponds to the Crank–Nicolson method.

In contrast to the Galerkin–collocation from the previous subsection, the Crank–Nicolson method only provides a solution that is continuous in time, but not a continuously differentiable solution.

4. Error analysis

The error analysis in this section makes heavily use of the results from [39] for semi-linear second order hyperbolic wave equations which can be carried over to the plate vibration problem when $f=f(u)$. This adaptation, however, requires some non-trivial steps, which are exposed in detail in the appendix of [17].

Here, we build on these results and, prior to presenting the main theoretical findings, introduce some useful definitions. Let $B \subset H$ and $l \in \mathbb{N}$. The local L^2-projections $\Pi_n^l : L^2(I_n;B) \to \mathbb{P}_l(I_n;B)$ are defined by

$$\int_{I_n} \big(\Pi_n^l w, q\big)\,\mathrm{d}t = \int_{I_n} (w,q)\,\mathrm{d}t \qquad \forall q \in \mathbb{P}_l(I_n;B).$$

We consider the Hermite interpolant in time $I_\tau^{k+1} : C^1(\bar{I};B) \to C^1(\bar{I};B) \cap X_\tau^{k+1}(B)$ studied in [16,31]. For this operator it is fulfilled that

$$I_\tau^{k+1}u(t_n) = u(t_n), \quad \partial_t I_\tau^{k+1}u(t_n) = \partial_t u(t_n), \quad n=0,\dots,N,$$

and

$$I_\tau^{k+1}u(t_{n,\mu}^{GL}) = u(t_{n,\mu}^{GL}), \quad n=1,\dots,N,\ \mu=2,\dots,k-1$$

and for a smooth function u, the following error estimates hold true on each interval I_n

$$\|\partial_t u - \partial_t I_\tau^{k+1}u\|_{C^0(\bar{I}_n;B)} \leq C\tau_n^{k+1}\|u\|_{C^{k+2}(\bar{I}_n;B)},$$

$$\|\partial_t^2 u - \partial_t^2 I_\tau^{k+1}u\|_{C^0(\bar{I}_n;B)} \leq C\tau_n^{k}\|u\|_{C^{k+2}(\bar{I}_n;B)}.$$

Moreover, we define the operator $R_\tau^k u\big|_{I_n} \in \mathbb{P}_k(I_n;B)$ for $n=1,\dots,N$ via the $(k+1)$ conditions

$$R_\tau^k u\big|_{I_n}(t_{n-1}) = I_\tau^{k+1}u(t_{n-1}),$$

$$\partial_t R_\tau^{k+1} u\big|_{I_n}(t_{n,s}) = \partial_t I_\tau^{k+1}u(t_{n,s}), \quad s=0,\dots,k$$

and we set $R_\tau^k u(0) := u(0)$.

Here, we briefly summarize some of the properties of R_τ^k that are important for the analysis. Their proofs can be found in [16,31].

Lemma 4. *Let $k \geq 3$. For $n = 1, \dots, N$ and $u \in C^{k+1}(\overline{I}_n; B)$ the estimate*

$$\|u - R_\tau^k u\|_{C^0(\overline{I}_n;B)} \leq C\tau_n^{k+1}\|u\|_{C^{k+1}(\overline{I}_n;B)}$$

holds.

A direct consequence of Lemma 4 is given in the following Corollary.

Corollary 5. *For $n = 1, \dots, N$ and $u \in C^{k+1}(\overline{I}_n; B)$ the estimate*

$$\|\partial_t u - \partial_t R_\tau^k u\|_{C^0(\overline{I}_n;B)} \leq C\tau_n^k \|u\|_{C^{k+1}(\overline{I}_n;B)} \tag{12}$$

is fulfilled.

Next we consider the global Hermite interpolation operator defined in (9).

Lemma 6. *For $I_\tau^H : C^1(\bar{I}; H) \to X_\tau^k(H)$ the following estimates*

$$\|u - I_\tau^H u\|_{C^0(\bar{I}_n;B)} \leq C\tau_n^{k+1}\|u\|_{C^{k+1}(\bar{I}_n;B)}$$
$$\|\partial_t u - \partial_t I_\tau^H u\|_{C^0(\bar{I}_n;B)} \leq C\tau_n^{k}\|u\|_{C^{k+1}(\bar{I}_n;B)}$$

hold true for all $n = 1, \dots, N$ and all $u \in C^{k+1}(\bar{I}_n; B)$.

Another important result for our analysis that has been proved in [16] is presented as follows.

Lemma 7. *Let us consider the Gauss quadrature formula (8b). For all polynomials $p \in \mathbb{P}_{k-1}(I_n; B)$ and all $n = 1, \dots, N$ it is fulfilled that*

$$\Pi_\tau^{k-2} p(t_{n,s}^G) = p(t_{n,s}^G), \quad s = 1, \dots, k-1.$$

Finally, a useful norm bound, see [16,31], is presented.

Lemma 8. *For any $u \in \mathbb{P}_k(I_n; H)$ the following inequality*

$$\int_{I_n} \|u\|^2 \,\mathrm{d}t \leq C\tau_n \|u(t_{n-1})\|^2 + \tau_n^2 \int_{I_n} \|\partial_t u\|^2 \,\mathrm{d}t,$$

is fulfilled.

5. Error estimates

Our ultimate goal in this section is to prove estimates for the error

$$E(t) = U(t) - U_{\tau,h}(t) = (e^0(t), e^1(t)),$$

where the Galerkin–collocation approximation $U_{\tau,h}$ is the solution of Problem 2 and $U(t) = (u^0(t), u^1(t))$. To achieve this, we start with estimations for $\partial_t E$ and afterwards we estimate the error E. Note that E is continuously differentiable in time if $U \in (C^1(\overline{I}; V))^2$ holds for the exact solution.

5.1. Error estimates for $\partial_t U_{\tau,h}$

First, we derive estimates for $\partial_t U_{\tau,h}$, which will be used later.

Theorem 9. *Let $U_{\tau,h} \in (X_\tau^k(V_h))^2$ be the discrete solution of Problem 1. Then, the time derivative $\partial_t U_{\tau,h} \in (X_\tau^{k-1}(V_h))^2$ solves for $n = 1, \dots, N$*

$$B_n^{GL}(\partial_t U_{\tau,h}, V_{\tau,h}) = Q_n^{GL}\big(\big(\partial_t I_\tau^H F, V_{\tau,h}\big)\big) = \int_{I_n} \big(\partial_t I_\tau^H F, V_{\tau,h}\big) \,\mathrm{d}t$$

for all $V_{\tau,h} \in (\mathbb{P}_{k-2}(I_n; V_h))^2$.

Proof. The proof differs from proof [6, Theorem 5.1] only in the definition of the operator $\mathcal{L}_h$.

Lemma 10. *Let $U_{0,h} = (R_h u_0, R_h u_1)$. Then the identity*

$$\partial_t U_{\tau,h}(0) = \begin{pmatrix} R_h & 0 \\ 0 & P_h \end{pmatrix} \partial_t U(0),$$

holds.

Proof. From $U_{\tau,h}(0) = U_{0,h} = (R_h u_0, R_h u_1)$ together with (10a) and (10b) for $n = 1$ it follows

$$\begin{aligned}\partial_t U_{\tau,h}(0) &= -\mathcal{L}_h U_{\tau,h}(0) + \mathcal{P}_h F(0) \\ &= -\begin{pmatrix} 0 & -P_h \\ A_h & 0 \end{pmatrix}\begin{pmatrix} R_h u_0 \\ R_h u_1 \end{pmatrix} + \begin{pmatrix} 0 \\ P_h f(0) \end{pmatrix} \\ &= \begin{pmatrix} P_h R_h u_1 \\ -A_h R_h u_0 + P_h f(0) \end{pmatrix}.\end{aligned}$$

Using the definition of the operator R_h, (3), we obtain

$$\big(A_h R_h u_0, v_h\big) = \big(\Delta R_h u_0, \Delta v_h\big) = \big(\Delta u_0, \Delta v_h\big) = \langle A u_0, v_h\rangle = \big(P_h A u_0, v_h\big)$$

for all $v_h \in V_h$. Thus, $A_h R_h u_0 = P_h A u_0$. In addition, we have

$$\big(P_h R_h u_1, v_h\big) = \big(R_h u_1, v_h\big)$$

for all $v_h \in V_h$, from which we infer $P_h R_h u_1 = R_h u_1$. Thus, we have

$$\partial_t U_{\tau,h}(0) = \begin{pmatrix} R_h u_1 \\ -P_h A u_0 + P_h f(0) \end{pmatrix}$$

Calculating the right-hand side gives

$$\begin{aligned}\begin{pmatrix} R_h & 0 \\ 0 & P_h \end{pmatrix} \partial_t U(0) &= \begin{pmatrix} R_h & 0 \\ 0 & P_h \end{pmatrix}(-\mathcal{L}U(0) + F(0)) \\ &= \begin{pmatrix} R_h & 0 \\ 0 & P_h \end{pmatrix}\begin{pmatrix} u_1 \\ -Au_0 + f(0) \end{pmatrix} \\ &= \begin{pmatrix} R_h u_1 \\ -P_h A u_0 + P_h f(0) \end{pmatrix}.\end{aligned}$$

This proves the statement.

Theorem 11. *Let $\hat{u}$ be the solution of (1) with data $(\hat{f}, \hat{u}_0, \hat{u}_1)$ instead of (f, u_0, u_1). Furthermore, let $l \in \mathbb{N}$ and $\hat{f}_\tau$ be an approximation of $\hat{f}$ satisfying*

$$\|\hat{f} - \hat{f}_\tau\|_{C(\overline{I}_n;H)} \leq C_{\hat{f}} \tau_n^{l+1}, \quad n = 1, \dots, N,$$

where $C_{\hat{f}}$ is independent of n, N and τ_n. Let $\hat{U}_{\tau,h} = (\hat{u}_{\tau,h}^0, \hat{u}_{\tau,h}^1) \in (X_\tau^l(V_h))^2$ denote the solution of the local perturbed cGP(l)-cG(r) problem

$$\int_{I_n} \big(\partial_t \hat{U}_{\tau,h}, V_{\tau,h}\big) + \big(\mathcal{L}_h \hat{U}_{\tau,h}, V_{\tau,h}\big) \,\mathrm{d}t = \int_{I_n} \big(\hat{F}_\tau, V_{\tau,h}\big) \,\mathrm{d}t$$

for all test function $V_{\tau,h} = (v_{\tau,h}^0, v_{\tau,h}^1) \in (\mathbb{P}_{l-1}(I_n; V_h))^2$ with $\hat{F}_\tau = (0, \hat{f}_\tau)$ and initial value $\hat{U}_{\tau,h}(t_{n-1}^+) = \hat{U}_{\tau,h}(t_{n-1}^-)$ for $n > 1$ and $\hat{U}_{\tau,h}(t_0) = \hat{U}_{0,h} = (R_h \hat{u}_0, P_h \hat{u}_1)$. Then a sufficiently smooth exact solution $\hat{u}$ satisfies

$$\|\hat{u}(t) - \hat{u}_{\tau,h}^0(t)\| + \|\partial_t \hat{u}(t) - \hat{u}_{\tau,h}^1(t)\| \leq C(\tau^{l+1} C_t(\hat{u}) + h^4 C_x(\hat{u})), \tag{13}$$
$$\|\Delta(\hat{u}(t) - \hat{u}_{\tau,h}^0(t))\| \leq C(\tau^{l+1} C_t(\hat{u}) + h^2 C_x(\hat{u})) \tag{14}$$

for all $t \in \overline{I}$ where $C_t(\hat{u})$ and $C_x(\hat{u})$ depend on various temporal and spatial derivatives of $\hat{u}$.

Proof. The proof follows the lines of the proof of [16, Theorem 5.5].

The main result in this section is

Theorem 12. *Let the exact solution $U = (u^0, u^1) := (u, \partial_t u)$ be sufficiently smooth and $U_{0,h} := (R_h u_0, R_h u_1)$, then the following estimates hold true for all $t \in \bar{I}$*

$$\|\partial_t U(t) - \partial_t U_{\tau,h}(t)\| \le C(\tau^k C_t(\partial_t u) + h^4 C_x(\partial_t u)) \le C(\tau^k + h^4), \tag{15}$$

$$\|\Delta(\partial_t u^0(t) - \partial_t u^0_{\tau,h}(t))\| \le C(\tau^k C_t(\partial_t u) + h^2 C_x(\partial_t u)) \le C(\tau^k + h^2), \tag{16}$$

where the quantities $C_t(\partial_t u)$ and $C_x(\partial_t u)$ depend on various temporal and spatial derivatives of $\partial_t u$.

Proof. This proof differs from proof [6, Theorem 5.5] in the definition of the operators A, $\mathcal{L}_h$ and R_h as well as the finite element space V_h. Analogous to this proof in [6], estimates (15) and (16) follow from estimates (13) and (14).

5.2. Error estimates for $U_{\tau,h}$

In this subsection we wish to estimate $U_{\tau,h}$. Therefore, we use the following splitting

$$E(t) = \Theta(t) + E_{\tau,h}(t) \quad \text{with}$$

$$\Theta(t) = U(t) - \mathcal{R}_h R_\tau^k U(t) \quad \text{and} \quad E_{\tau,h}(t) = \mathcal{R}_h R_\tau^k U(t) - U_{\tau,h}(t),$$

where $E_{\tau,h}(t) = (e^0_{\tau,h}(t), e^1_{\tau,h}(t))$.

Lemma 13 *(Estimation of the interpolation error). Let $m \in \{0, 1\}$. Then, the error estimates*

$$\|\Theta(t)\|_m \le C(h^{4-m} + \tau_n^{k+1}), \quad t \in \overline{I}_n, \tag{17}$$

$$\|\partial_t \Theta(t)\|_m \le C(h^{4-m} + \tau_n^{k}), \quad t \in \overline{I}_n, \tag{18}$$

are valid for $n = 1, \dots, N$ where $\|\cdot\|_0 = \|\cdot\|$.

Proof. First, we note that for the elliptic projection R_h defined in (3) it holds that $\|R_h u\| \le C\|\Delta R_h u\| \le C\|\Delta u\|$. Then, the proof follows the same lines as the proof of [16, Lemma 5.7].

Using Lemma 4 and the approximation properties of the elliptic projection R_h we obtain

$$\begin{aligned}\|\Theta(t)\|_m &= \|U(t) - \mathcal{R}_h R_\tau^k U(t)\|_m \\ &\le \|U(t) - \mathcal{R}_h U(t)\|_m + \|\mathcal{R}_h(U(t) - R_\tau^k U(t))\|_m \\ &\le C h^{4-m} \|U\|_{C^0(\overline{I}; H^4(\Omega))} + \tau_n^{k+1} \|U\|_{C^{k+1}(\overline{I}; H^1(\Omega))}\end{aligned}$$

which proves (17). Applying estimate (12) and the fact that ∂_t and R_h commute yields

$$\begin{aligned}\|\partial_t \Theta(t)\|_m &\le \|\partial_t U(t) - \mathcal{R}_h \partial_t U(t)\|_m + \|\mathcal{R}_h(\partial_t U(t) - \partial_t R_\tau^{k+1} U(t))\|_m \\ &\le C h^{4-m} \|\partial_t U\|_{C^0(\overline{I}; H^4(\Omega))} + C \tau_n^{k+1} \|U\|_{C^{k+2}(\overline{I}; H^1(\Omega))}\end{aligned}$$

which shows (18).

Lemma 14 *(Consistency error). Let $U \in C^1(\overline{I}; V) \times C^1(\overline{I}; H)$. Then, we have*

$$B_n^{GL}(E, V_{\tau,h}) = Q_n^{GL}(\left(I_\tau^{GL} F - I_\tau^H F, V_{\tau,h}\right)) = Q_n^{GL}(\left(F - I_\tau^H F, V_\tau, h\right))$$

for all $V_{\tau,h} \in (Y^{k-2}_{\tau,h})^2$ and $n = 1, \dots, N$.

Proof. Because of Proposition 3 and the consistency property (7), the proof differs from the proof [6, Lemma 5.7] only in the definition of the operators $\mathcal{L}$ and $\mathcal{L}_h$.

For the proof of the following result we refer to [16, Lemma 5.9].

Lemma 15. *Let $p \in \mathbb{P}_k(I_n)$ be an arbitrary polynomial of degree less than or equal to k. Then, the identity*

$$\partial_t p(t^G_{n,\mu}) = \partial_t I_\tau^{GL} p(t^G_{n,\mu})$$

is satisfied for all Gauss points $t^G_{n,\mu} \in I_n$, $\mu = 1, \dots, k-1$.

Lemma 16 *(Stability). The following identity*

$$\begin{aligned}&B_n^{GL}((e^0_{\tau,h}, e^1_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})) \\ &= \frac{1}{2}(\|\Delta e^0_{\tau,h}(t_n)\|^2 - \|\Delta e^0_{\tau,h}(t_{n-1})\|^2 + \|e^1_{\tau,h}(t_n)\|^2 - \|e^1_{\tau,h}(t_{n-1})\|^2)\end{aligned}$$

holds true for all $n = 1, \dots, N$.

Proof. First, we note that $\left((e^0_{\tau,h}, e^1_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})\right) \in \mathbb{P}_{2k-2}(I_n; \mathbb{R})$. Using the Gauss-Lobatto interpolation operator, we obtain $I_\tau^{GL} e^1_{\tau,h} \in \mathbb{P}_{k-1}(I_n; V_h)$ and $A_h I_\tau^{GL} e^0_{\tau,h} \in \mathbb{P}_{k-1}(I_n; V_h)$. Then, we have

$$\begin{aligned}&B_n^{GL}((e^0_{\tau,h}, e^1_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})) \\ &= Q_n^{GL}(\left((\partial_t e^0_{\tau,h}, \partial_t e^1_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})\right)) \\ &\quad + Q_n^{GL}(\left((-P_h I_\tau^{GL} e^1_{\tau,h}, A_h I_\tau^{GL} e^0_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})\right)) \\ &= \int_{I_n} \left((\partial_t e^0_{\tau,h}, \partial_t e^1_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})\right) \mathrm{d}t \\ &\quad + \int_{I_n} \left((-P_h I_\tau^{GL} e^1_{\tau,h}, A_h I_\tau^{GL} e^0_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})\right) \mathrm{d}t \\ &= T_1 + T_2.\end{aligned} \tag{19}$$

Using Lemma 7 together with the exactness of the $(k-1)$-point Gauss quadrature formula for polynomials on $\mathbb{P}_{2k-3}(I_n; \mathbb{R})$ and subsequent application of Lemma 15, we obtain

$$\begin{aligned}T_1 &= \int_{I_n} \left((\Pi_n^{k-2} \partial_t e^0_{\tau,h}, \Pi_n^{k-2} \partial_t e^1_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})\right) \mathrm{d}t \\ &= Q_n^G(\left((\partial_t e^0_{\tau,h}, \partial_t e^1_{\tau,h}), (A_h I_\tau^{GL} e^0_{\tau,h}, I_\tau^{GL} e^1_{\tau,h})\right)) \\ &= Q_n^G(\left((\partial_t I_\tau^{GL} e^0_{\tau,h}, \partial_t I_\tau^{GL} e^1_{\tau,h}), (A_h I_\tau^{GL} e^0_{\tau,h}, I_\tau^{GL} e^1_{\tau,h})\right)) \\ &= \frac{\tau_n}{2} \sum_{\mu=1}^{k-1} \hat{w}_\mu^G \left(\partial_t I_\tau^{GL} e^0_{\tau,h}(t^G_{n,\mu}), A_h I_\tau^{GL} e^0_{\tau,h}(t^G_{n,\mu})\right) \\ &\quad + \frac{\tau_n}{2} \sum_{\mu=1}^{k-1} \hat{w}_\mu^G \left(\partial_t I_\tau^{GL} e^1_{\tau,h}(t^G_{n,\mu}), I_\tau^{GL} e^1_{\tau,h}(t^G_{n,\mu})\right) \\ &= \frac{\tau_n}{2} \sum_{\mu=1}^{k-1} \hat{w}_\mu^G \frac{1}{2} \mathrm{d}_t \|A_h^{1/2} I_\tau^{GL} e^0_{\tau,h}(t^G_{n,\mu})\|^2 + \frac{\tau_n}{2} \sum_{\mu=1}^{k-1} \hat{w}_\mu^G \frac{1}{2} \mathrm{d}_t \|I_\tau^{GL} e^1_{\tau,h}(t^G_{n,\mu})\|^2.\end{aligned}$$

Using the exactness of the $(k-1)$-point Gauss quadrature formula on $\mathbb{P}_{2k-3}(I_n; \mathbb{R})$ again, we obtain

$$\begin{aligned}T_1 &= \int_{I_n} \left(\frac{1}{2} \mathrm{d}_t \|A_h^{1/2} I_\tau^{GL} e^0_{\tau,h}(t)\|^2 + \frac{1}{2} \mathrm{d}_t \|I_\tau^{GL} e^1_{\tau,h}(t)\|^2\right) \mathrm{d}t \\ &= \frac{1}{2} \left(\|A_h^{1/2} e^0_{\tau,h}(t_n)\|^2 - \|A_h^{1/2} e^0_{\tau,h}(t_{n-1})\|^2 + \|e^1_{\tau,h}(t_n)\|^2 - \|e^1_{\tau,h}(t_{n-1})\|^2\right).\end{aligned} \tag{20}$$

By using Lemma 7, we have that

$$
\begin{aligned}
T_2 &= \int_{I_n} \left((-P_h I_\tau^{GL} e^1_{\tau,h}, A_h I_\tau^{GL} e^0_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h}) \right) \mathrm{d}t \\
&= \int_{I_n} \left((-\Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h}, \Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}), (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h}) \right) \mathrm{d}t \\
&= \int_{I_n} \left(-\Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h}, \Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h} \right) + \left(\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h} \right) \mathrm{d}t \\
&= 0.
\end{aligned} \tag{21}
$$

Inserting the equations (20) and (21) into (19) along with the identity $\|A_h^{1/2} v_h\| = \|\Delta v_h\|$ for $v_h \in V_h$ finally yields the assertion.

Lemma 17 *(Boundedness). Let* $V_{\tau,h} = (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})$. *Then the estimate*

$$|B_n^{GL}(\Theta, V_{\tau,h})| \leq C \tau_n^{1/2} (\tau_n^{k+1} + h^4) \{ \tau_n \|E_{\tau,h}\|^2 + \tau_n^2 Q_n^G(\|\partial_t E_{\tau,h}\|^2) \}^{1/2}$$

is valid for all $n = 1, \dots, N$.

Proof. The proof of this lemma differs from that of [16, Lemma 5.11] only in the definition of the elliptic projection operator and the applied projection error estimates.

Lemma 18 *(Estimates on right-hand side term). Let* $V_{\tau,h} = (\Pi_n^{k-2} A_h I_\tau^{GL} e^0_{\tau,h}, \Pi_n^{k-2} I_\tau^{GL} e^1_{\tau,h})$. *Then, the estimation*

$$
\begin{aligned}
&Q_n^{GL}(((0, f - I_\tau^H f), V_{\tau,h})) \\
&\quad \leq C \tau_n^{1/2} \tau_n^{k+1} \{ \tau_n \|E_{\tau,h}(t_{n-1})\|^2 + \tau_n^2 Q_n^G(\|\partial_t E_{\tau,h}\|^2) \}^{1/2}
\end{aligned}
$$

is valid for $n = 1, \dots, N$.

Proof. The proof differs from proof [6, Lemma 5.11] only in the definition of the operator A_h.

Lemma 19 *(Estimates on* $E_{\tau,h}$*). Let* $U_{0,h} = (R_h, u_0, R_h, u_1)$. *Then*

$$\|e^0_{\tau,h}(t_n)\|_1^2 + \|e^1_{\tau,h}\|^2 \leq C \left(\tau^{k+1} + h^4 \right)^2$$

holds true for all $n = 1, \dots, N$. *Additionally,*

$$
\begin{aligned}
\|\Delta e^0_{\tau,h}(t)\| &\leq C(\tau^{k+1} + h^3) \\
\|e^0_{\tau,h}(t)\| + \|e^1_{\tau,h}(t)\| &\leq C(\tau^{k+1} + h^4)
\end{aligned}
$$

is satisfied for all $t \in \overline{I}$.

Proof. The proof follows the lines of [6, Lemma 5.12].

Theorem 20 *(Error estimate for* $U_{\tau,h}$*). Let* $U = (u, \partial_t u)$ *be the solution of the problem* (1) *and* $U_{\tau,h}$ *be the discrete solution of problem* (11) *along with the initial condition* $U_{0,h} = (R_h u_0, R_h u_1)$. *Then the following estimates for the error* $E(t) = (e^0(t), e^1(t)) = U(t) - U_{\tau,h}(t)$ *apply for all* $t \in \overline{I}$*:*

$$
\begin{aligned}
\|e^0(t)\| + \|e^1(t)\| &\leq C(\tau^{k+1} + h^4), \\
\|\Delta e^0(t)\| &\leq C(\tau^{k+1} + h^3).
\end{aligned}
$$

Additionally, the estimates

$$
\begin{aligned}
\|e^0\|_{L^2(I;H)} + \|e^1\|_{L^2(I;H)} &\leq C(\tau^{k+1} + h^4), \\
\|\Delta e^0\|_{L^2(I;H)} &\leq C(\tau^{k+1} + h^3)
\end{aligned} \tag{22}
$$

are satisfied.

Proof. The proof of this theorem differs from the proof of [6, Theorem 5.13] only in the definition of the operators.

6. Numerical experiments

The aim of the numerical experiments included in this section is:

(i) to compute the numerical convergence orders for the time discretizations discussed in the previous sections;
(ii) to compare the solutions obtained when using the cGP-C^1(3)-, the cGP(2)- and the Crank–Nicolson (cGP(1)-) method for time discretization.

The numerical experiments in the first part of this section are used to confirm the error estimates proven in Section 5 and to show the faster convergence of the cGP-C^1(3)-method compared with the other algorithms investigated. In the second part of this section, we perform a comparative study in order to demonstrate the superiority of the cGP-C^1(3)-method over the considered only continuous in time approximation schemes.

We assume that the spatial domain $\Omega \subset \mathbb{R}^2$ is either the unit square $(0,1)^2$ or the square $(-1,1)^2$ which during the discretization process has been partitioned as bisections of N^2 squares with mesh size $h = \sqrt{2}/N$ or $h = 2\sqrt{2}/N$, respectively. Furthermore, we use the Bogner–Fox–Schmit element throughout for spatial discretization.

All the numerical tests included in this section have been conducted in NGSolve, see https://ngsolve.org.

6.1. Numerical convergence study

We will utilize the first example to provide numerical evidence for the error estimates proven in Section 5, on the one hand, and to present the better convergence properties of the cGP-C^1(3)-method, on the other. For this purpose, we compare the experimental order of convergence of different time discretization schemes.

Thereby, we expect to obtain an order of convergence of 2 for the Crank-Nicolson (cGP(1)-) method, an order of 3 for the cGP(2)-method, and an order of 4 for the cGP-C^1(3)-method. In addition, as in [16] for the wave equation, we expect to observe superconvergence for the cGP(2)-method in the discrete time points. While we predict a convergence order of 4 only in the discrete points for the cGP(2)-method, we anticipate a convergence order of 4 in all time points for the cGP-C^1(3)-method.

In our first example, we solve system (1) for the spatial domain $\Omega = (0,1)^2$ and the temporal domain $I = (0,1)$. For a right-hand side

$$
\begin{aligned}
f(\boldsymbol{x}, t) = &-4\pi^2 \sin(2\pi t) \cdot \sin^2(\pi x_1) \cdot \sin^2(\pi x_2) \\
&+ \pi^4 \sin(2\pi t) \cdot \left[16 \sin^2(\pi x_1) - 8 \right] \cdot \sin^2(\pi x_2) \\
&+ 2\pi^4 \sin(2\pi t) \left[2 - 4 \sin^2(\pi x_1) \right] \cdot \left[2 - 4 \sin^2(\pi x_2) \right] \\
&+ \pi^4 \sin(2\pi t) \cdot \sin^2(\pi x_1) \cdot \left[16 \sin^2(\pi x_2) - 8 \right]
\end{aligned}
$$

and initial values

$$
\begin{aligned}
u_0(\boldsymbol{x}) &= 0, \\
u_1(\boldsymbol{x}) &= 2\pi \cdot \sin^2(\pi x_1) \cdot \sin^2(\pi x_2),
\end{aligned}
$$

the exact solution u is given by

$$u(\boldsymbol{x}, t) = \sin(2\pi t) \cdot \sin^2(\pi x_1) \cdot \sin^2(\pi x_2). \tag{23}$$

For a comparison of the solutions of different time discretization methods, we consider the norms

$$\|u - u_{\tau,h}\|_{L^\infty(L^2)} = \max_{t \in [0,T]} \left(\int_\Omega |u(\boldsymbol{x}, t) - u_{\tau,h}(\boldsymbol{x}, t)|^2 \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}}$$

Table 1
Numerical errors and convergence orders for the Crank–Nicolson (cGP(1)), cGP(2)- and cGP-C^1(3)-method for the function (23).

τ	h	$\|u-u_{\tau,h}\|_{L^\infty_\tau(L^2)}$	Order	$\|u-u_{\tau,h}\|_{L^\infty(L^2)}$	Order	$\|u-u_{\tau,h}\|_{L^2(L^2)}$	Order
cGP(1)							
$\tau_0/2^0$	$h_0/2^0$	3.296e-03	–	1.794e-02	–	9.081e-03	–
$\tau_0/2^1$	$h_0/2^1$	7.835e-04	2.07	4.794e-03	1.90	2.281e-03	1.99
$\tau_0/2^2$	$h_0/2^2$	1.877e-04	2.06	1.233e-03	1.96	5.746e-04	1.99
$\tau_0/2^3$	$h_0/2^3$	4.746e-05	1.98	3.080e-04	2.00	1.435e-04	2.00
$\tau_0/2^4$	$h_0/2^4$	1.194e-05	1.99	7.701e-05	2.00	3.589e-05	2.00
cGP(2)							
$\tau_0/2^0$	$h_0/2^0$	1.058e-03	–	1.059e-03	–	8.369e-04	–
$\tau_0/2^1$	$h_0/2^1$	7.258e-05	3.87	1.063e-04	3.32	6.731e-05	3.64
$\tau_0/2^2$	$h_0/2^2$	4.553e-06	3.99	1.247e-05	3.09	6.635e-06	3.34
$\tau_0/2^3$	$h_0/2^3$	2.867e-07	3.99	1.510e-06	3.05	7.625e-07	3.12
$\tau_0/2^4$	$h_0/2^4$	1.798e-08	4.00	1.857e-07	3.02	9.310e-08	3.03
cGP-C^1(3)							
$\tau_0/2^0$	$h_0/2^0$	1.165e-03	–	1.231e-03	–	8.533e-04	–
$\tau_0/2^1$	$h_0/2^1$	8.141e-05	3.84	8.151e-05	3.92	5.673e-05	3.91
$\tau_0/2^2$	$h_0/2^2$	5.314e-06	3.94	5.314e-06	3.94	3.363e-06	4.08
$\tau_0/2^3$	$h_0/2^3$	2.998e-07	4.15	2.999e-07	4.15	2.005e-07	4.07
$\tau_0/2^4$	$h_0/2^4$	1.823e-08	4.04	1.824e-08	4.04	1.222e-08	4.04

and

$$\|u-u_{\tau,h}\|_{L^2(L^2)} = \left(\int_I \int_\Omega |u(\boldsymbol{x},t)-u_{\tau,h}(\boldsymbol{x},t)|^2 \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \right)^{\frac{1}{2}}.$$

In order to approximate the $\|\cdot\|_{L^\infty(L^2)}$ norm, we first evaluate the maximum only in the discrete temporal points in which we have computed the discrete solution $u_{\tau,h}$ and denote this value by $\|\cdot\|_{L^\infty_\tau(L^2)}$. Secondly, we determine the maximum by additionally evaluating the discrete solution in the time points

$$I_\tau = \{t_{n,j} : t_{n,j} = t_{n-1} + j \cdot \frac{1}{100} \cdot \tau_n,\ j = 1,\dots,99,\ n = 1,\dots,N\},$$

which we denote by $\|\cdot\|_{L^\infty(L^2)}$.

We compute both norms on a sequence of spatial and temporal meshes in order to determine the numerical convergence orders. We start with $\tau_0 = 0.1$, $h_0 = \frac{\sqrt{2}}{5}$ and halve both after each pass. With $e_{\tau,h}$ we denote the error with time step τ and mesh size h, then the following formula

$$EOC = \log_2\left(\frac{e_{\tau,h}}{e_{\frac{\tau}{2},\frac{h}{2}}}\right)$$

is applied to compute the experimental order of convergence (EOC).

The discretization errors and the corresponding convergence orders for the Crank–Nicolson, cGP(2)- and cGP-C^1(3)-method in case of function (23) are presented in Table 1.

As seen from this table, the discretization errors for the cGP(2)- and the cGP-C^1(3)-method are smaller than the corresponding values for the Crank–Nicolson method. Moreover, the cGP(2)-method with global continuous and piecewise quadratic functions gives higher convergence orders than the Crank–Nicolson method with piecewise linear functions. The cGP-C^1(3)-method has the highest convergence orders in both norms of all the methods studied.

The numerical convergence orders in the $\|\cdot\|_{L^2(L^2)}$ norm tend to 2 for the Crank–Nicolson method and for the cGP(2)-method to 3. In the $\|\cdot\|_{L^\infty_\tau(L^2)}$ norm it can be seen even that the convergence order is 4 for the cGP(2)-method whereas it is 2 for the Crank–Nicolson algorithm. This observed superconvergence in the $\|\cdot\|_{L^\infty_\tau(L^2)}$ norm is due to the evaluation in the Gauss-Lobatto points. Comparing the evaluation of the maximum only in the time points where we have computed the discrete solution as in $\|\cdot\|_{L^\infty_\tau(L^2)}$ with the evaluation on a finer temporal mesh as performed with $\|\cdot\|_{L^\infty(L^2)}$, verifies this superconvergence effect. For a more detailed study of the superconvergence in the case of the wave equation, we refer to, e.g., [16]. All convergence orders for the cGP-C^1(3)-method tend to 4. This confirms the convergence order proven in Theorem 20.

Consequently, we obtain better convergence results in all discrete time points for the cGP-C^1(3)-method, although we have the same numerical costs in terms of degrees of freedom as for the cGP(2)-method.

6.2. Vibration in heterogeneous media

In this section we apply the cGP-C^1(3)-method to a more sophisticated problem with arising complex wave phenomena. Additionally, we stress the superiority of this method by analyzing and comparing the number of non-zero entries in the system matrix which is involved in the linear system of equations in every time step. This is of importance with respect to the application of iterative solvers to the algebraic systems.

As a second example in this work, we consider the problem

$$\partial_{tt}u(\boldsymbol{x},t) + \Delta(c(\boldsymbol{x})\Delta u(\boldsymbol{x},t)) = f(\boldsymbol{x},t) \quad \text{in } \Omega\times(0,T], \tag{24a}$$
$$u(\boldsymbol{x},0) = u_0(\boldsymbol{x}) \quad \text{in } \Omega, \tag{24b}$$
$$\partial_t u(\boldsymbol{x},0) = u_1(\boldsymbol{x}) \quad \text{in } \Omega, \tag{24c}$$
$$u(\boldsymbol{x},t) = 0 \quad \text{on } \partial\Omega\times(0,T], \tag{24d}$$
$$\partial_{\boldsymbol{n}}u(\boldsymbol{x},t) = 0 \quad \text{on } \partial\Omega\times(0,T]. \tag{24e}$$

The coefficient $c > 0$ encodes the stiffness of the involved materials. We use the setting

$$\Omega = (-1,1)^2, \qquad T = \frac{3}{100}, \qquad c(\boldsymbol{x}) = \begin{cases} 1, & \text{if } x_2 < 0.2, \\ 9, & \text{if } x_2 \geq 0.2, \end{cases} \qquad f = 0,$$

with the initial values

$$u_0 = \exp(-|20\cdot\boldsymbol{x}|^2)\cdot(1-x_1^2)^2\cdot(1-x_2^2)^2, \qquad u_1 = 0. \tag{25}$$

The initial value is a regularized Dirac impulse and stimulates the system's dynamics. The solution of (24) with the initial values (25) is illustrated in Fig. 1 for the cGP-C^1(3)-method for the time step size 10^{-3}. We observe the development of complex wave phenomena with sharp fronts. We define the control region $\Omega_c = (0.75-l_c, 0.75+l_c)\times(-l_c, l_c)$ with $l_c = 1/32$ that simulates a senor and evaluate the quantity

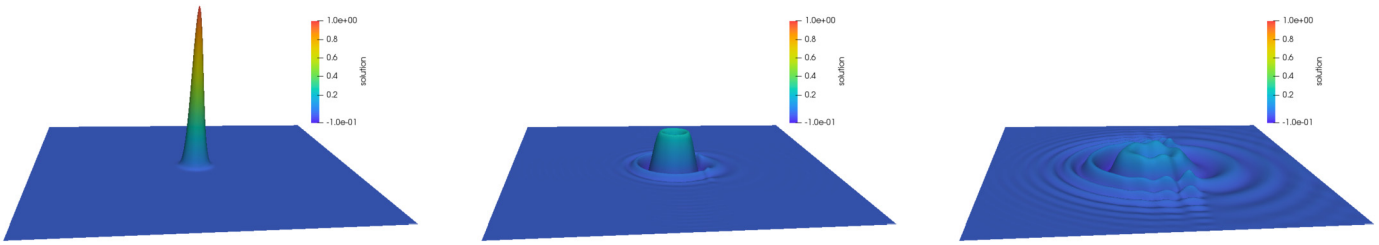


Fig. 1. Numerical solution of Problem (24) with initial values (25) for the cGP-C^1(3)-method with $\tau = 10^{-3}$.

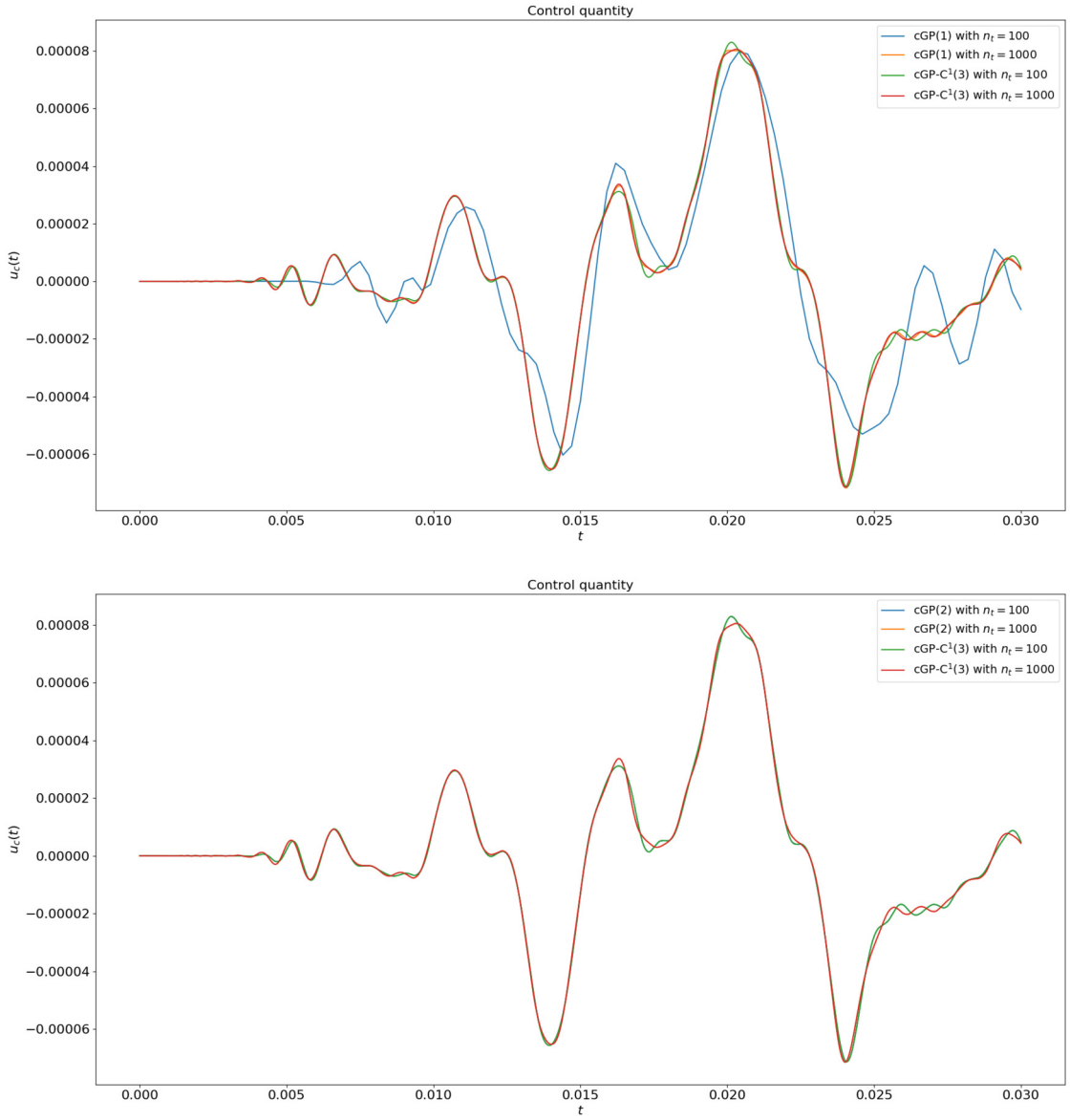


Fig. 2. Evaluation of the control quantity (26) with initial values (25) for the Crank–Nicolson (cGP(1)-), cGP(2)- and cGP-C^1(3)-method with different time step sizes.

$$u_c(t) = \int_{\Omega_c} u_{\tau,h}(\boldsymbol{x}, t)\, \mathrm{d}\boldsymbol{x}\,. \tag{26}$$

Our first aim is to examine and compare the signal arrival at the sensor position for the three time discretization methods and different time step sizes. All calculations were performed on a fixed 64×64 spatial mesh. A direct solver for the linear system of equations was used. Time step sizes were chosen as $\tau = 1/n_t$, where n_t denotes the number of prescribed time subintervals. The corresponding evaluations of the control quantity are presented in Fig. 2. As can be seen, the cGP(2)- and cGP-C^1(3)-method lead to similar graphs, whereas the cGP(1)-method requires a smaller time step size for a comparable accuracy. As demonstrated in the first numerical example, the cGP-C^1(3)-method shows convergence order of order four in all time points, whereas the cGP(2)-method yields fourth-order superconvergence in the discrete time step

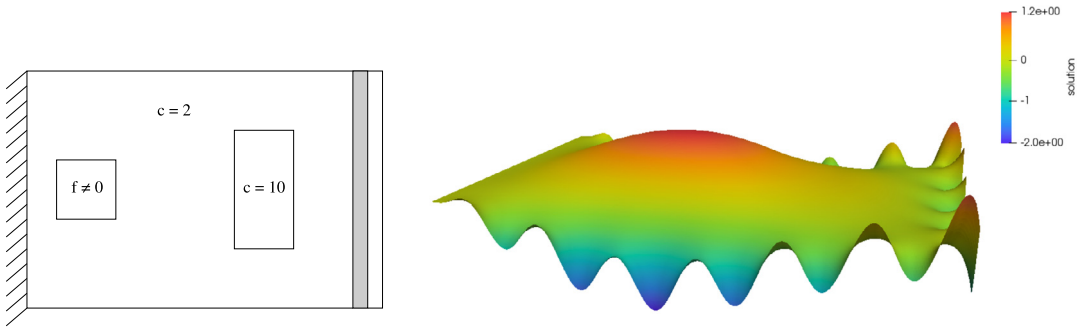


Fig. 3. Test setting (left) and numerical solution of Problem (27) for the cGP-C^1(3)-method with $\tau = 1/2000$ (right).

Table 2
Number of non-zero entries in the respective system matrix and number of degrees of freedom for the Crank–Nicolson (cGP(1)-), cGP(2)- and cGP-C^1(3)-method using different spatial refinements.

$\lvert\mathcal{T}_h\rvert$	cGP(1)		cGP(2)		cGP-C^1(3)	
	dof	nze	dof	nze	dof	nze
256	2312	1.4e+05	4624	5.0e+05	4624	3.6e+05
1024	8712	5.7e+05	17424	2.0e+06	17424	1.4e+06
4096	33800	2.3e+06	67600	7.9e+06	67600	5.7e+06
16384	133128	9.1e+06	266256	3.2e+07	266256	2.3e+07
65536	528392	3.7e+07	1056784	1.3e+08	1056784	9.3e+07
262144	2105352	1.5e+08	4210704	5.2e+08	4210704	3.8e+08

points only. Therefore, a similar accuracy of the either schemes can realistically be expected.

To evaluate these schemes further, we address the density of the arising system matrices which represents a further criteria for their computational efficiency. For this, we compare the number of non-zero entries in the system matrices resulting from each of the time discretization schemes. The number of non-zero entries (nze) as well as the number of degrees of freedom (dof) for the Crank–Nicolson, cGP(2)- and cGP-C^1(3)-method are summarized in Table 2. The Crank–Nicolson method has fewer degrees of freedom and a smaller number of non-zero entries, but provides less accuracy, as already shown before. Although the system matrices of the cGP(2)- and cGP-C^1(3)-method have the same size, the cGP-C^1(3)-method leads to less non-zero entries in the system matrix. This reduced number of non-zero entries is especially advantageous for the application of iterative solvers. In [5], iterative solvers are applied for solving the linear systems of the cGP-C^1(3) approximation of the wave equation. In this case, a strong superiority of the cGP-C^1(3) approach over the cGP(2) one is observed with respect to accuracy and runtime of the simulations.

6.3. Wave propagation in structural health monitoring

Our final numerical experiment considers an application of practical interest. We mimic structural health monitoring of mechanical engineering. The test setting is illustrated in Fig. 3. For $\Omega = (0,12)\times(0,8)$, and $T = 200$ we consider the system

$$\partial_{tt}u(\boldsymbol{x},t) + \Delta(c(\boldsymbol{x})\Delta u(\boldsymbol{x},t)) = f(\boldsymbol{x},t) \quad \text{in } \Omega\times(0,T], \tag{27a}$$
$$u(\boldsymbol{x},0) = 0 \quad \text{in } \Omega, \tag{27b}$$
$$\partial_t u(\boldsymbol{x},0) = 0 \quad \text{in } \Omega \tag{27c}$$
$$u(\boldsymbol{x},t) = 0 \quad \text{on } \Gamma_D\times(0,T], \tag{27d}$$
$$\Delta u(\boldsymbol{x},t) = 0 \quad \text{on } \Gamma_N\times(0,T], \tag{27e}$$
$$\partial_{\boldsymbol{n}}\Delta u(\boldsymbol{x},t) = 0 \quad \text{on } \Gamma_N\times(0,T]. \tag{27f}$$

Here we impose a homogeneous Dirichlet boundary condition on $\Gamma_D = \{0\}\times(0,8)$ and use natural (do-nothing) boundary conditions for the biharmonic operator on $\Gamma_N = \partial\Omega\setminus\Gamma_D$. We put

$$c(\boldsymbol{x}) = \begin{cases} 10, & \text{if } \boldsymbol{x}\in[7,9]\times[2,6], \\ 2, & \text{otherwise.} \end{cases}$$

The heterogeneity in the coefficient c can be regarded as a material defect. Wave propagation is stimulated by the right-hand side function

$$f(\boldsymbol{x},t) = \begin{cases} f_1(t)\cdot f_2(\boldsymbol{x}), & \text{if } \boldsymbol{x}\in[1,3]\times[3,5], \\ 0, & \text{otherwise,} \end{cases}$$

where

$$f_1(t) = \frac{1}{4}\cdot\left(1-\cos\left(\frac{2\pi}{50}\cdot t\right)\right)\cdot\cos\left(\frac{2\pi}{10}\cdot t\right),$$
$$f_2(\boldsymbol{x}) = (x_1-1)\cdot(x_1-3)\cdot(x_2-3)\cdot(x_2-5).$$

The temporal function f_1 is a so-called burst signal that is typically used in structural health monitoring. Our goal is to evaluate the control quantity

$$u_c(t) = \int_{\Omega_c} u_{\tau,h}(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} \tag{28}$$

on the control region $\Omega_c = (11,11.5)\times(0,8)$. This is done for different time discretization methods. The computed numerical solution of Problem (27) by using the cGP-C^1(3)-method with time step size $\tau = 1/2000$ is shown in the right plot of Fig. 3.

We measure the control quantity for the cGP(1)-, cGP(2)- and cGP-C^1(3)-method with different time step sizes. Again, the time step size be given by $\tau = 1/n_t$, where n_t denotes the number of time intervals. The problem was solved respectively on a fixed mesh of 48×32 cells. The results are shown in Fig. 4. The cGP(2)- and cGP-C^1(3)-methods produce comparable graphs again. The cGP(1)-method requires a much smaller time step size. The loss of accuracy of the cGP(1)-method compared to the higher-order ones leads to strong dispersion effects that increase in time. Thus, the cGP-C^1(3)-method provides a solution of higher accuracy and with higher regularity, even for larger time step sizes. To quantify differences between the cGP(2)- and cGP-C^1(3)-method might require much finer spatial meshes if wave profiles of the complexity used here have to be resolved. This does not become feasible without parallel computations and is beyond the scope of this work.

Acknowledgement

The second and third authors acknowledge the support of this work by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the project “Physics-oriented solvers for multicompartmental poromechanics” under grant number 456235063.

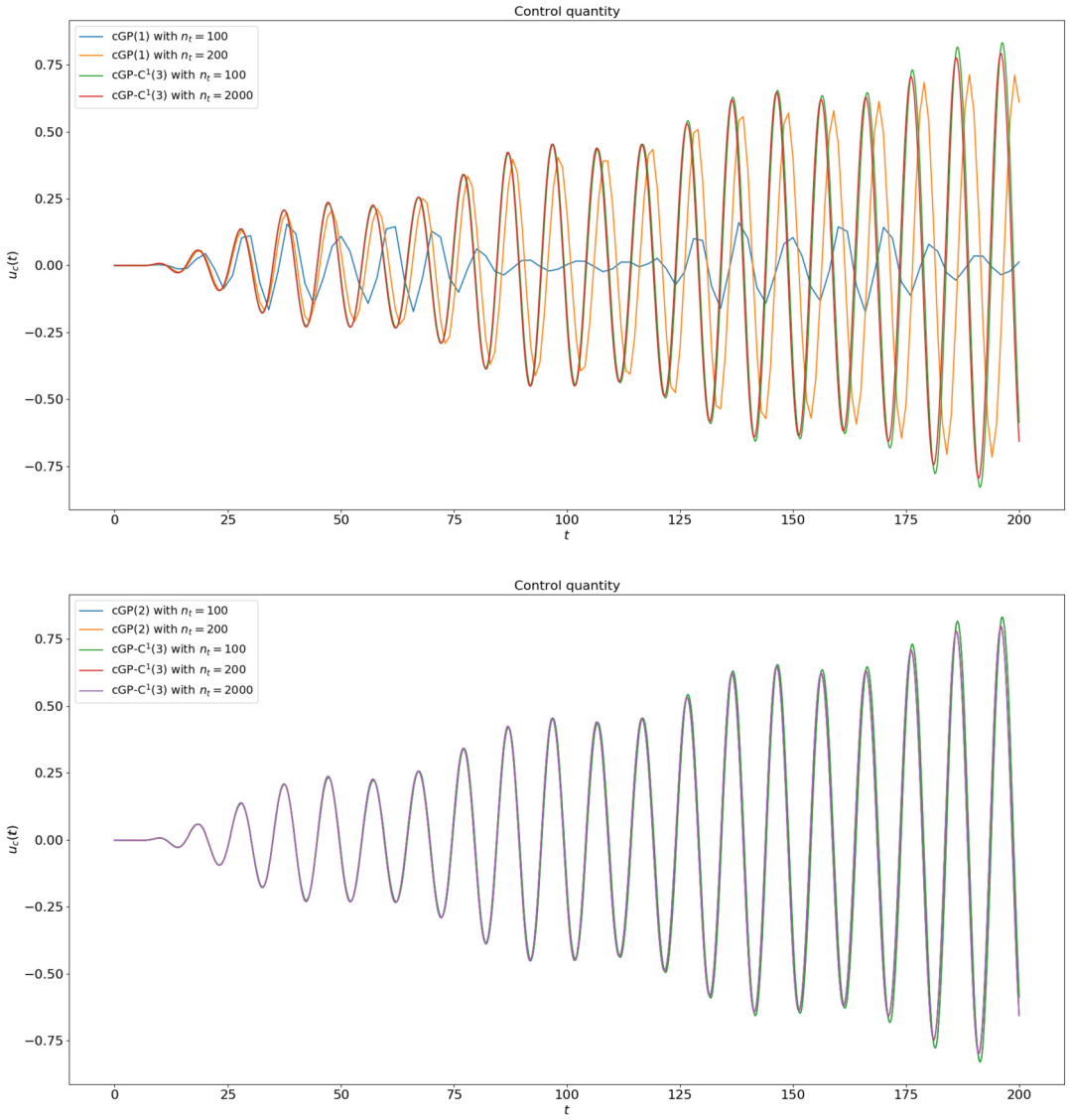


Fig. 4. Evaluation of the control quantity (28) for the Crank–Nicolson (cGP(1)-), cGP(2)- and cGP-C^1(3)-method with different time step sizes.

References

[1] R.A. Adams, J.J.F. Fournier, Sobolev Spaces, 2 edition, Elsevier, Amsterdam, 2003.

[2] G. Akrivis, C. Makridakis, R.H. Nochetto, Optimal order a posteriori error estimates for a class of Runge–Kutta and Galerkin methods, Numer. Math. 114 (2009) 133–160.

[3] G. Akrivis, C. Makridakis, R.H. Nochetto, Galerkin and Runge–Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence, Numer. Math. 118 (2011) 429–456.

[4] M. Anselmann, M. Bause, Higher order Galerkin-collocation time discretization with Nitsche's method for the Navier–Stokes equations, Math. Comput. Simul. 189 (2021) 141–162.

[5] M. Anselmann, M. Bause, Numerical study of Galerkin-collocation approximation in time for the wave equation, in: W. Dörfler, et al. (Eds.), Mathematics of Wave Phenomena. Trends in Mathematics, Birkhäuser, Cham, 2020, pp. 15–36.

[6] M. Anselmann, M. Bause, S. Becher, G. Matthies, Galerkin-collocation approximation in time for the wave equation and its post-processing, ESAIM: Math. Model. Numer. Anal. 54 (6) (2020) 2099–2123.

[7] P.F. Antonietti, L. Beirão da Veiga, S. Scacchi, M. Verani, A C^1 virtual element method for the Cahn-Hilliard equation with polygonal meshes, SIAM J. Numer. Anal. 54 (1) (2016) 34–56.

[8] P.F. Antonietti, B.A. De Dios, I. Mazzieri, A. Quarteroni, Stability analysis of discontinuous Galerkin approximations to the elastodynamics problem, J. Sci. Comput. 68 (1) (2016) 143–170.

[9] P.F. Antonietti, G. Manzini, M. Verani, The fully nonconforming virtual element method for biharmonic problems, Math. Models Methods Appl. Sci. 28 (2) (2018) 387–407.

[10] P.F. Antonietti, G. Manzini, M. Verani, The conforming virtual element method for polyharmonic problems, Comput. Math. Appl. 79 (7) (2020) 2021–2034.

[11] J.H. Argyris, I. Fried, D.W. Scharpf, The tuba family of plate elements for the matrix displacement method, J. R. Aeronaut. Soc. 72 (1969) 701–709.

[12] L. Bales, I. Lasiecka, Continuous finite elements in space and time for the nonhomogeneous wave equation, Comput. Math. Appl. 27 (1994) 91–102.

[13] W. Bangerth, M. Geiger, R. Rannacher, Adaptive Galerkin finite element methods for the wave equation, Comput. Methods Appl. Math. 10 (2010) 3–48.

[14] M. Bause, M. Anselmann, Comparative study of continuously differentiable Galerkin time discretizations for the wave equation, PAMM 19 (2019).

[15] M. Bause, M.P. Bruchhäuser, U. Köcher, Flexible goal-oriented adaptivity for higher-order space-time discretizations of transport problems with coupled flow, Comput. Math. Appl. 91 (2021) 17–35.

[16] M. Bause, U. Köcher, F.A. Radu, F. Schieweck, Post-processed Galerkin approximation of improved order for wave equations, Math. Comput. 89 (322) (2020) 595–627.

[17] M. Bause, M. Lymbery, K. Osthues, C^1-conforming variational discretization of the biharmonic wave equation, arXiv:2107.03906, 2021.

[18] M. Bause, F.A. Radu, U. Köcher, Error analysis for discretizations of parabolic problems using continuous finite elements in time and mixed finite elements in space, Numer. Math. 137 (4) (2017) 773–818.

[19] S. Becher, G. Matthies, Variational time discretizations of higher order and higher regularity, arXiv:2003.04056, 2020.

[20] S. Becher, G. Matthies, D. Wenzel, Variational methods for stable time discretization of first-order differential equations, in: K. Georgiev, M. Todorov, I. Georgiev (Eds.), Advanced Computing in Industrial Mathematics. BGSIAM, Springer, Cham, 2018, pp. 63–75.

[21] F.K. Bogner, R.L. Fox, L.A. Schmit, The generation of interelement compatible stiffness and mass matrices by the use of interpolation formulae, in: Proc. Conf. Matrix Methods in Struct. Mech., AirForce Inst. of Tech., Wright Patterson AF Base, Ohio, 1965, pp. 397–444.
[22] F. Bonaldi, D.A. Di Pietro, G. Geymonat, F. Krasucki, A hybrid high-order method for Kirchhoff-Love plate bending problems, ESAIM: Math. Model. Numer. Anal. 52 (2018) 393–421.
[23] S.C. Brenner, L.-Y. Sung, C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains, J. Sci. Comput. 22/23 (2005) 83–118.
[24] F. Brezzi, L.D. Marini, Virtual element methods for plate bending problems, Comput. Methods Appl. Mech. Eng. 253 (2013) 455–462.
[25] E. Burman, P. Hansbo, M.G. Larson, Cut Bogner-Fox-Schmit elements for plates, Adv. Model. Simul. Eng. Sci. 7 (27) (2020).
[26] C. Carstensen, N. Nataraj, Lower-order equivalent nonstandard finite element methods for biharmonic plates, arXiv:2102.08125, 2021, pp. 1–37.
[27] Long Chen, Xuehai Huang, Nonconforming virtual element method for $2m$th order partial differential equations in $\mathbb{R}^n$, Math. Comput. 89 (324) (2020) 1711–1744.
[28] C. Chinosi, L.D. Marini, Virtual element method for fourth order problems: L^2-estimates, Comput. Math. Appl. 72 (2016) 1959–1967.
[29] R.W. Clough, J.L. Tocher, Finite element stiffness matrices for analysis of plate bending, in: Matrix Methods in Structural Mechanics (AFFDL-TR-66–80), 1966, pp. 515–545.
[30] Z. Dong, A. Ern, Hybrid high-order and weak Galerkin methods for the biharmonic problem, arXiv:2103.16404, 2021, pp. 1–28.
[31] A. Ern, F. Schieweck, Discontinuous Galerkin method in time combined with a stabilized finite element method in space for linear first-order PDEs, Math. Comput. 85 (301) (2016) 2099–2129.
[32] I. Faragó, Convergence and stability constant of the theta-method, in: Conference Applications of Mathematics, 2013.
[33] D.A. French, T.E. Peterson, A continuous space-time finite element method for the wave equation, Math. Comput. 65 (1996) 491–506.
[34] E.H. Georgoulis, P. Houston, Discontinuous Galerkin methods for the biharmonic problem, IMA J. Numer. Anal. 29 (2009) 573–594.
[35] M.J. Grote, A. Schneebeli, D. Schötzau, Discontinuous Galerkin finite element method for the wave equation, SIAM J. Numer. Anal. 44 (6) (2006) 2408–2431.
[36] T.J.R. Hughes, The Finite Element Method, Dover Publications, 2000.
[37] C. Johnson, Discontinuous Galerkin finite element methods for second order hyperbolic problems, Comput. Methods Appl. Mech. Eng. 107 (1993) 117–129.
[38] H. Joulak, B. Beckermann, On Gautschi's conjecture for generalized Gauss–Radau and Gauss–Lobatto formulae, J. Comput. Appl. Math. 233 (2009) 768–774.
[39] O. Karakashian, C. Makridakis, Convergence of a continuous Galerkin method with mesh modification for nonlinear wave equations, Math. Comput. 74 (249) (2005) 85–102.
[40] U. Köcher, M. Bause, Variational space-time methods for the wave equation, J. Sci. Comput. 61 (2014) 424–453.
[41] U. Köcher, M.P. Bruchhäuser, M. Bause, Efficient and scalable data structures and algorithms for goal-oriented adaptivity of space–time FEM codes, SoftwareX 10 (2019) 100239.
[42] J.-L. Lions, Optimal Control of Systems Governed by Partial Differential Equations, Die Grundlehren der mathematischen Wissenschaften, Band 170, Springer-Verlag, New York-Berlin, 1971, translated from the French by S.K. Mitter.
[43] J.-L. Lions, E. Magenes, Non-homogeneous Boundary Value Problems and Applications. Vol. II, Die Grundlehren der mathematischen Wissenschaften, Band 182, Springer-Verlag, New York-Heidelberg, 1972, translated from the French by P. Kenneth.
[44] L. Morley, The triangular equilibrium element in the solution of plate bending problems, Aeronaut. Q. 19 (1968) 149–169.
[45] I. Mozolevski, E. Süli, A priori error analysis for the hp-version of the discontinuous Galerkin finite element method for the biharmonic equation, Comput. Methods Appl. Math. 3 (2003) 596–607.
[46] L. Mu, J. Wang, X. Ye, Weak Galerkin finite element methods for the biharmonic equation on polytopal meshes, Numer. Methods Partial Differ. Equ. 30 (2014) 1003–1029.
[47] E. Süli, I. Mozolevski, hp-version interior penalty DGFEMs for the biharmonic equation, Comput. Methods Appl. Mech. Eng. 196 (2007) 1851–1863.
[48] M. Wang, J. Xu, The Morley element for fourth order elliptic equations in any dimensions, Numer. Math. 103 (2006) 155–169.
[49] W.L. Wood, A unified set of single step algorithms. Part II: theory, Int. J. Numer. Methods Eng. 20 (1984) 2303–2309.
[50] W.L. Wood, Practical Time-Stepping Schemes, Clarendon Press, 1990.
[51] X. Ye, S. Zhang, Z. Zhang, A new P1 weak Galerkin method for the biharmonic equation, J. Comput. Appl. Math. 364 (2020) 112337.
[52] R. Zhang, Q. Zhai, A weak Galerkin finite element scheme for the biharmonic equations by using polynomials of reduced order, J. Sci. Comput. 64 (2015) 559–585.
[53] J. Zhao, S. Chen, B. Zhang, The nonconforming virtual element method for plate bending problems, Math. Models Methods Appl. Sci. 26 (2016) 1671–1687.
[54] J. Zhao, B. Zhang, S. Chen, S. Mao, The Morley-type virtual element for plate bending problems, J. Sci. Comput. 76 (2018) 610–629.
[55] O.C. Zienkiewicz, R.L. Taylor, The Finite Element Method. Vol. 2: Solid Mechaincs, 5 edition, Butterworth–Heinemann, 2000.

AUXILIARY SPACE MULTIGRID METHOD BASED ON ADDITIVE SCHUR COMPLEMENT APPROXIMATION

NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2015; **22**:965–986
Published online 14 October 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/nla.1959

Auxiliary space multigrid method based on additive Schur complement approximation

J. Kraus[1], M. Lymbery[2,*,†] and S. Margenov[2]

[1]*Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenberger Str. 69, A-4040 Linz, Austria*
[2]*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 25A, 1113 Sofia, Bulgaria*

SUMMARY

In this paper, the idea of auxiliary space multigrid methods is introduced. The construction is based on a two-level block factorization of local (finite element stiffness) matrices associated with a partitioning of the domain into overlapping or non-overlapping subdomains. The two-level method utilizes a coarse-grid operator obtained from additive Schur complement approximation. Its analysis is carried out in the framework of auxiliary space preconditioning and condition number estimates for both the two-level preconditioner and the additive Schur complement approximation are derived. The two-level method is recursively extended to define the auxiliary space multigrid algorithm. In particular, so-called Krylov cycles are considered. The theoretical results are supported by a representative collection of numerical tests that further demonstrate the efficiency of the new algorithm for multiscale problems. Copyright © 2014 John Wiley & Sons, Ltd.

Received 31 October 2013; Revised 19 August 2014; Accepted 22 August 2014

KEY WORDS: auxiliary space multigrid; algebraic multilevel iteration; additive Schur complement approximation

1. INTRODUCTION

Partial differential equations (PDEs) play a key role in the modeling of various processes that occur in fields as diverse as physics, chemistry, biology, economics, engineering, and life sciences.

The numerical solution of PDE based on discretization techniques such as finite difference, finite volume, and finite element methods typically reduces a continuous problem to a discrete problem that finally is represented in the form of one or more systems of linear algebraic equations.

In many applications, the arising linear systems are sparse and very large. Hence, it is important to construct efficient iterative solution methods that converge uniformly with respect to problem size and parameters. The most successful approaches for achieving this goal are domain decomposition (DD) (see, e.g., [1, 2]) and multigrid (MG)/multilevel methods (see, e.g., [3–5]).

As has been shown in [2, 6], two-level DD methods are robust as long as the variations of the coefficients of the scalar elliptic equation are bounded inside coarse-grid cells. Recently, this robustness has been achieved also for problems with general coefficient variations using coarse spaces based on local generalized eigenvalue problems [7, 8]. The latter approach has been generalized for the mixed form and the stream function formulations of Stokes's and Brinkman's equations [9]. Other related techniques for constructing suitable coarse spaces for PDE modeling heterogeneous media have been considered in [10, 11].

*Correspondence to: IICT, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 25A, 1113 Sofia, Bulgaria.
†E-mail: mariq@parallel.bas.bg

Regarding computational complexity, MG methods have asserted to be most efficient because they have been demonstrated to be optimal with respect to the problem size; see [3, 5] and the references therein. However, their design needs careful adaptation for problems with certain 'bad' parameters in the PDE model. From this perspective, it is desirable to enhance their robustness in the sense of covering wider problem classes [12].

The algebraic multilevel iteration (AMLI) framework provides useful tools to achieve this goal, for example, more general polynomial acceleration techniques or Krylov cycles resulting in nonlinear so-called variable-step preconditioners [13–16].

In the present paper, a non-variational auxiliary space MG (ASMG) algorithm for general symmetric positive definite (SPD) problems is introduced. The method is based on exact two-by-two block factorization of local (stiffness) matrices that correspond to a sequence of coverings of the domain by overlapping or non-overlapping subdomains. The coarse-grid matrix is defined via additive Schur complement approximation (ASCA) [17–19]. Its sparsity can be controlled by the size and overlap of the subdomains. The coarse-grid correction step, as used in classical MG methods, however, is replaced by a correction that involves the application of an auxiliary space preconditioner. For that reason, the method studied in this paper is referred to as the ASMG method. The idea of integrating DD techniques into MG algorithms was performed as early as in [20].

The remainder of the paper is organized as follows. In Section 2, a fictitious space preconditioner based on ASCA is constructed and analyzed and further complemented by a smoothing process defining an auxiliary space two-level preconditioner and a related stationary two-grid method. In Section 3, a condition number estimate of the auxiliary space preconditioner is proven followed by a theorem characterizing the ASCA. The recursive extension of the auxiliary space two-grid method is defined and described algorithmically in Section 4. As known from the AMLI theory [13, 14, 21, 22], the convergence of the multilevel algorithm depends on uniform two-level estimates. In the present context, the decisive quantity, analogous to the Cauchy–Bunyakowski–Schwarz constant in the hierarchical basis methods, is given by the energy norm of a certain projection operator. Its efficient computation by a multilevel algorithm is addressed in Section 5. Finally, several numerical tests are presented, addressing both the performance of the ASMG method on a collection of challenging high-frequency high-contrast problems and the computation of the spectral bounds of interest.

2. AUXILIARY SPACE TWO-GRID METHOD

2.1. Fictitious space preconditioner

Let $\{\Omega_{G_i} : i = 1, 2, \ldots, n_{\mathcal{G}}\}$ be a covering of the domain Ω by non-overlapping or overlapping subdomains Ω_{G_i}, that is,

$$\overline{\Omega} = \bigcup_{i=1}^{n_{\mathcal{G}}} \overline{\Omega}_{G_i}, \tag{1}$$

where $\mathcal{G} = \{G_i : i = 1, 2, \ldots, n_{\mathcal{G}}\}$ denotes a set of macrostructures that correspond to the adjacency graphs associated with the subdomains Ω_{G_i}. The construction is such that the global stiffness matrix A can be assembled from small-sized (local) symmetric positive semi-definite (SPSD) (stiffness) matrices A_{G_i} corresponding to the subdomains Ω_{G_i} [18]. Then A can be written in the form

$$A = \sum_{i=1}^{n_{\mathcal{G}}} R_{G_i}^T A_{G_i} R_{G_i}, \tag{2}$$

where the operator R_{G_i} restricts a global vector $\mathbf{v} \in V = \mathbb{R}^N$ to the local space $V_{G_i} = \mathbb{R}^{n_{G_i}}$ related to the subdomain Ω_{G_i}.

Consider a partitioning of the set $\mathcal{D}$ of degrees of freedom (DOF) into two subsets

$$\mathcal{D} = \mathcal{D}_{\mathrm{f}} \oplus \mathcal{D}_{\mathrm{c}}, \tag{3}$$

DOI: 10.1002/nla

where $\mathcal{D}_\mathrm{f}$ consists of fine DOF and $\mathcal{D}_\mathrm{c}$ is the set of coarse DOF. Let the cardinalities of these sets be denoted by $N_1 := |\mathcal{D}_\mathrm{f}|$ and $N_2 := |\mathcal{D}_\mathrm{c}|$.

Further, let $n_{G_i:1}$ and $n_{G_i:2}$ be the number of fine and coarse DOF, respectively, that are associated with the subdomain Ω_{G_i}. The dimension of the local space $\dim(V_{G_i}) = n_{G_i}$ then can be presented as the sum

$$n_{G_i} = n_{G_i:1} + n_{G_i:2}. \tag{4}$$

Next, the auxiliary (fictitious) space $\tilde{V} = \mathbb{R}^{\tilde{N}}$ of dimension $\tilde{N} = \left(\sum_{i=1}^{n_\mathcal{G}} n_{G_i:1}\right) + N_2$ is introduced, and a surjective mapping $\Pi_{\tilde{D}} : \tilde{V} \to V$ is defined via the relations

$$R_1^T = \begin{bmatrix} R_{G_1:1} \\ R_{G_2:1} \\ \vdots \\ R_{G_{n_\mathcal{G}}:1} \end{bmatrix}, \qquad R = \begin{bmatrix} R_1 & 0 \\ 0 & I_2 \end{bmatrix}, \qquad \Pi_{\tilde{D}} = \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}, \tag{5}$$

where R_1^T is of size $(\sum_{i=1}^{n_\mathcal{G}} n_{G_i:1}) \times N_1$, the identity matrix I_2 is of size $N_2 \times N_2$, and R and $\Pi_{\tilde{D}}$ are of dimension $N \times \tilde{N}$. Here, $\tilde{D}$ is a block-diagonal matrix of size $\tilde{N} \times \tilde{N}$ to be specified later.

Given the introduced splitting of the DOF into fine and coarse the matrices $A_{G_i}, i = 1, \ldots, n_\mathcal{G}$ and A can be written in a two-by-two block form

$$A_{G_i} = \begin{bmatrix} A_{G_i:11} & A_{G_i:12} \\ A_{G_i:21} & A_{G_i:22} \end{bmatrix} \; i = 1, \ldots, n_\mathcal{G}, \qquad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{6}$$

Let the $\tilde{N} \times \tilde{N}$ DD auxiliary matrix $\tilde{A}$ be defined by

$$\tilde{A} = \begin{bmatrix} A_{G_1:11} & & & & A_{G_1:12}R_{G_1:2} \\ & A_{G_2:11} & & & A_{G_2:12}R_{G_2:2} \\ & & \ddots & & \vdots \\ & & & A_{G_{n_\mathcal{G}}:11} & A_{G_{n_\mathcal{G}}:12}R_{G_{n_\mathcal{G}}:2} \\ R_{G_1:2}^T A_{G_1:21} & R_{G_2:2}^T A_{G_2:21} & \ldots & R_{G_{n_\mathcal{G}}:2}^T A_{G_{n_\mathcal{G}}:21} & \sum_{i=1}^{n_\mathcal{G}} R_{G_i:2}^T A_{G_i:22} R_{G_i:2} \end{bmatrix}, \tag{7}$$

where the matrices A_{G_i} are assumed to be SPSD with $A_{G_i:11}$ SPD. Then the matrix $\tilde{A}$ is SPSD, and the Schur complement

$$Q := S_{\tilde{A}} = \tilde{A}_{22} - \tilde{A}_{21}\tilde{A}_{11}^{-1}\tilde{A}_{12} \tag{8}$$

exists and is SPSD. If in addition $\tilde{A}$ is SPD, which is the case for example if A and $\tilde{A}$ are irreducible, it introduces an energy inner product on the auxiliary space $\tilde{V}$. Moreover, from (5) and (7), it follows that the relation

$$A = R\tilde{A}R^T \tag{9}$$

holds. Note that whereas each fine DOF from $\mathcal{D}_\mathrm{f}$ adds one to the dimension of A, it increases the dimension of $\tilde{A}$ by the number of subdomains to which it belongs. At the same time, the blocks that correspond to the coarse DOF are identical for both the original matrix A and the auxiliary matrix $\tilde{A}$, that is,

$$A_{22} = \tilde{A}_{22} = \sum_{i=1}^{n_\mathcal{G}} R_{G_i:2}^T A_{G_i:22} R_{G_i:2}. \tag{10}$$

The matrix $\tilde{A}_{11}$ is of block diagonal form with blocks of size $n_{G_i:1} \times n_{G_i:1}$ for $i = 1, 2, \ldots, n_\mathcal{G}$ and thus allows for a cheap computation of the interpolation matrix

$$P = \begin{bmatrix} -\tilde{A}_{11}^{-1}\tilde{A}_{12} \\ I_2 \end{bmatrix},$$

DOI: 10.1002/nla

which for any given coarse vector $\mathbf{v}_2$ provides the minimum energy extension $\mathbf{v} = P\mathbf{v}_2$ in the sense that $\mathbf{v} = \operatorname{argmin}_{\mathbf{w}:\,\mathbf{w}_2=\mathbf{v}_2} \|\mathbf{w}\|_{\tilde{A}} = \operatorname{argmin}_{\mathbf{w}:\,\mathbf{w}_2=\mathbf{v}_2} \left(\mathbf{w}^T \tilde{A}\mathbf{w}\right)^{1/2}$. Hence, the exact Schur complement of $\tilde{A}$ defines the Galerkin coarse-grid matrix of the variational two-grid method corresponding to the energy-minimizing interpolation P on the auxiliary space $\tilde{V}$, that is,

$$Q = A_c = P^T \tilde{A} P = S_{\tilde{A}} = \tilde{A}_{22} - \tilde{A}_{21}\tilde{A}_{11}^{-1}\tilde{A}_{12}. \tag{11}$$

It is important to note that A_c can be determined without computing the (global) triple matrix product. Instead, the coarse-grid matrix can be assembled from its subdomain contributions, the corresponding local Schur complements, which can be computed in parallel for all subdomains, that is,

$$A_c = \sum_{i=1}^{n_{\mathcal{G}}} R_{G_i:2}^T \left(A_{G_i:22} - A_{G_i:21}A_{G_i:11}^{-1}A_{G_i:12}\right) R_{G_i:2}.$$

The number of non-zero entries in A_c can be controlled by limiting the size n_{G_i} of the subdomains Ω_{G_i}, which guarantees the sparsity of the coarse-grid matrix.

Remark 1
The spectral equivalence of the Schur complements S_A and $S_{\tilde{A}}$ of A and $\tilde{A}$ has been established in [17, 23] for SPD stiffness matrices A arising from conforming finite element discretizations of second-order scalar elliptic PDEs with piecewise constant coefficients under the assumption that the jumps are aligned with the coarse mesh. There, the construction of $\tilde{A}$ has been for a non-overlapping DD, and the set of coarse DOF has been associated with a coarse mesh resulting from standard (full) coarsening. In a more recent paper [18], it has been proven that with a proper overlap of the subdomains, the spectral equivalence of S_A and $S_{\tilde{A}}$ holds uniformly and is independent of jumps of a piecewise constant diffusion coefficient even if these appear across arbitrary element interfaces on the finest grid.

In the following, let C denote the fictitious (auxiliary) space preconditioner defined via the relation

$$C^{-1} := \Pi_{\tilde{D}} \tilde{A}^{-1} \Pi_{\tilde{D}}^T. \tag{12}$$

The idea of fictitious space preconditioning goes back to Sergei Nepomnyaschikh [24]. An important tool for deriving condition number estimates in this context is the so-called fictitious space lemma [24], which is as follows.

Lemma 1
Let V be a Hilbert space equipped with inner product $\langle\cdot,\cdot\rangle$ and $A : V \mapsto V$ an SPD (w.r.t. $\langle\cdot,\cdot\rangle$) linear operator. Let $\tilde{V}$ be a second Hilbert space (auxiliary space) equipped with inner product $\langle\cdot,\cdot\rangle_\sim$ and $\tilde{A} : \tilde{V} \mapsto \tilde{V}$ a second SPD linear operator. Further, let $\Pi : \tilde{V} \mapsto V$ be a surjective mapping satisfying the following conditions:

(a) For all $\mathbf{v} \in V$, there exists $\tilde{\mathbf{v}} \in \tilde{V}$ such that $\Pi\tilde{\mathbf{v}} = \mathbf{v}$ and $\tilde{c}\left\langle \tilde{A}\tilde{\mathbf{v}}, \tilde{\mathbf{v}}\right\rangle_\sim \leqslant \langle A\mathbf{v}, \mathbf{v}\rangle$.
(b) $\langle A\Pi\tilde{\mathbf{u}}, \Pi\tilde{\mathbf{u}}\rangle \leqslant c\left\langle \tilde{A}\tilde{\mathbf{u}}, \tilde{\mathbf{u}}\right\rangle_\sim$ for all $\tilde{\mathbf{u}} \in \tilde{V}$.

Introduce the adjoint operator $\Pi^\star : V \mapsto \tilde{V}$ by

$$\langle \Pi\tilde{\mathbf{u}}, \mathbf{v}\rangle = \left\langle \tilde{\mathbf{u}}, \Pi^\star\mathbf{v}\right\rangle_\sim \quad \text{for all } \tilde{\mathbf{u}} \in \tilde{V}, \mathbf{v} \in V.$$

Then

$$\tilde{c}\left\langle A^{-1}\mathbf{u}, \mathbf{u}\right\rangle \leqslant \left\langle \Pi\tilde{A}^{-1}\Pi^\star\mathbf{u}, \mathbf{u}\right\rangle \leqslant c\left\langle A^{-1}\mathbf{u}, \mathbf{u}\right\rangle \quad \text{for all } \mathbf{u} \in V. \tag{13}$$

DOI: 10.1002/nla

From now on, if not stated otherwise, we denote by

$$\langle \mathbf{u}, \mathbf{v} \rangle := \sum_{i=1}^{N} u_i v_i \text{ for all } \mathbf{u}, \mathbf{v} \in V \quad \text{and} \quad \langle \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_{\sim} := \sum_{i=1}^{\tilde{N}} \tilde{u}_i \tilde{v}_i \text{ for all } \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \tilde{V},$$

the Euclidean inner products on $V = \mathbb{R}^N$ and $\tilde{V} = \mathbb{R}^{\tilde{N}}$. In this case, $\Pi^\star = \Pi^T$ and the following corollary can be proven.

Corollary 1
Let $\Pi = \Pi_{\tilde{D}}$ be defined according to (5), where $\tilde{D} \in \mathbb{R}^{\tilde{N} \times \tilde{N}}$ is an SPD matrix. Then the fictitious space preconditioner defined in (12) with auxiliary matrix $\tilde{A}$ as given in (7) satisfies

$$\left\langle A^{-1}\mathbf{u}, \mathbf{u} \right\rangle \leqslant \left\langle \Pi \tilde{A}^{-1} \Pi^T \mathbf{u}, \mathbf{u} \right\rangle \leqslant \|\pi_{\tilde{D}}\|^2_{\tilde{A}} \left\langle A^{-1}\mathbf{u}, \mathbf{u} \right\rangle \quad \text{for all } \mathbf{u} \in V, \tag{14}$$

where $\pi_{\tilde{D}} := R^T \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}$.

Proof
The estimate (14) follows from Lemma 1 because in the present context conditions (a) and (b) hold with constants $\tilde{c} = 1$ and $c = \|\pi_{\tilde{D}}\|^2_{\tilde{A}}$:

(a) For all $\mathbf{v} \in V = \mathbb{R}^N$, define $\tilde{\mathbf{v}} := R^T \mathbf{v}$. Then

$$\Pi \tilde{\mathbf{v}} = \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}\tilde{\mathbf{v}} = \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}R^T\mathbf{v} = \mathbf{v}.$$

Hence,

$$\langle A\mathbf{v}, \mathbf{v} \rangle = \left\langle R\tilde{A}R^T\mathbf{v}, \mathbf{v} \right\rangle = \left\langle R\tilde{A}\tilde{\mathbf{v}}, \mathbf{v} \right\rangle = \left\langle \tilde{A}\tilde{\mathbf{v}}, R^T\mathbf{v} \right\rangle_{\sim} = \left\langle \tilde{A}\tilde{\mathbf{v}}, \tilde{\mathbf{v}} \right\rangle_{\sim},$$

and thus condition (a) holds with $\tilde{c} = 1$.

(b) Further, as

$$\begin{aligned} \left\langle A\Pi_{\tilde{D}}\tilde{\mathbf{u}}, \Pi_{\tilde{D}}\tilde{\mathbf{u}} \right\rangle &= \left\langle R\tilde{A}R^T \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}\tilde{\mathbf{u}}, \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}\tilde{\mathbf{u}} \right\rangle \\ &= \left\langle \tilde{A}R^T \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}\tilde{\mathbf{u}}, R^T \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}\tilde{\mathbf{u}} \right\rangle_{\sim} \\ &=: \left\langle \tilde{A}\pi_{\tilde{D}}\tilde{\mathbf{u}}, \pi_{\tilde{D}}\tilde{\mathbf{u}} \right\rangle_{\sim} \end{aligned}$$

we see that the inequality (b) is sharp for

$$c := \sup_{\tilde{\mathbf{u}} \neq \mathbf{0}} \frac{\left\langle \tilde{A}\pi_{\tilde{D}}\tilde{\mathbf{u}}, \pi_{\tilde{D}}\tilde{\mathbf{u}} \right\rangle_{\sim}}{\left\langle \tilde{A}\tilde{\mathbf{u}}, \tilde{\mathbf{u}} \right\rangle_{\sim}} = \|\pi_{\tilde{D}}\|^2_{\tilde{A}}, \tag{15}$$

which completes the proof.

□

Remark 2
The operator $\pi_{\tilde{D}} = R^T \left(R\tilde{D}R^T\right)^{-1} R\tilde{D}$ is a projection, that is,

$$\pi^2_{\tilde{D}} = \pi_{\tilde{D}}.$$

Moreover, as $\pi_{\tilde{D}}$ is non-trivial, nor does it equal identity, owing to Kato's lemma [25],

$$\|\pi_{\tilde{D}}\|_{\tilde{A}} = \|I - \pi_{\tilde{D}}\|_{\tilde{A}}, \tag{16}$$

where $\|\cdot\|_{\tilde{A}}$ is the $\tilde{A}$ inner product norm, that is,

$$\|\tilde{\mathbf{v}}\|_{\tilde{A}} := \sqrt{\left\langle \tilde{A}\tilde{\mathbf{v}}, \tilde{\mathbf{v}} \right\rangle_{\sim}}, \quad \forall \tilde{\mathbf{v}} \in \tilde{V}.$$

DOI: 10.1002/nla

For a proof of (16), see [5] where it has also been shown that $\|\pi_{\tilde{D}}\|_{\tilde{A}}$ is related to the cosine γ of the angle between the two spaces Range $\left(\pi_{\tilde{D}}\right)$ and Range $\left(I-\pi_{\tilde{D}}\right)$ in the $\tilde{A}$ inner product, that is,

$$\|\pi_{\tilde{D}}\|_{\tilde{A}}=\frac{1}{\sqrt{1-\gamma^2}}.$$

Hence, the relative condition number $\kappa\left(C^{-1}A\right)$ of the fictitious space preconditioner defined via $C^{-1}:=\Pi_{\tilde{D}}\tilde{A}^{-1}\Pi_{\tilde{D}}^T$ can be estimated by

$$\kappa\left(C^{-1}A\right)\leqslant c=\|\pi_{\tilde{D}}\|_{\tilde{A}}^2=\frac{1}{1-\gamma^2}.$$

The idea of fictitious space preconditioning has been further developed in the setting of auxiliary space preconditioning by incorporating an additional smoother, hence relaxing the constraints on the choice of the auxiliary space $\tilde{V}$. For details, see [26].

2.2. *Two-grid method*

The proposed auxiliary space two-grid method determines a stationary iterative procedure

$$\mathbf{x}_{k+1}=\mathbf{x}_k+B^{-1}\mathbf{r}_k, \tag{17}$$

where the k-th iterate and the k-th residual have been denoted by $\mathbf{x}_k$ and $\mathbf{r}_k$, respectively. Assume that M is an A-convergent smoother, that is,

$$\|I-M^{-1}A\|_A<1.$$

Then the symmetrized smoother $\overline{M}=M\left(M+M^T-A\right)^{-1}M^T$ is also A convergent, that is,

$$\|I-\overline{M}^{-1}A\|_A=\|\left(I-M^{-T}A\right)\left(I-M^{-1}A\right)\|_A<1,$$

and the two-grid preconditioner utilizing the fictitious space preconditioner (12) is defined by

$$B^{-1}:=\overline{M}^{-1}+\left(I-M^{-T}A\right)C^{-1}\left(I-AM^{-1}\right). \tag{18}$$

As $I-B^{-1}A=\left(I-M^{-T}A\right)\left(I-C^{-1}A\right)\left(I-M^{-1}A\right)$, the two-grid method is convergent, that is,

$$\|I-B^{-1}A\|_A<1, \tag{19}$$

if the auxiliary space correction is non-expansive in A norm, that is,

$$\|I-C^{-1}A\|_A\leqslant 1. \tag{20}$$

From Corollary 1, we have

$$\frac{1}{c}\mathbf{v}^T\mathbf{v}\leqslant\frac{1}{c}\mathbf{v}^T\left(\Pi_{\tilde{D}}\tilde{A}^{-1}\Pi_{\tilde{D}}^T\right)A\mathbf{v}\leqslant\mathbf{v}^T\mathbf{v}\quad\forall\mathbf{v}\in V,$$

and thus (20) and finally (19) are satisfied, for example, if the matrix C in (18) is defined by

$$C^{-1}=\tau^{-1}\Pi_{\tilde{D}}\tilde{A}^{-1}\Pi_{\tilde{D}}^T, \tag{21}$$

where τ is a scaling parameter satisfying

$$\tau\geqslant c:=\|\pi_{\tilde{D}}\|_{\tilde{A}}^2. \tag{22}$$

Another way of defining B^{-1} is via the product matrix

$$\hat{B}=\begin{bmatrix}M & 0\\ \Pi_{\tilde{D}}^TA & I\end{bmatrix}\begin{bmatrix}\left(M+M^T-A\right)^{-1} & 0\\ 0 & \tau\tilde{A}\end{bmatrix}\begin{bmatrix}M^T & A\Pi_{\tilde{D}}\\ 0 & I\end{bmatrix}.$$

DOI: 10.1002/nla

Then

$$\hat{B}^{-1} = \begin{bmatrix} M^{-T} & -M^{-T}A\Pi_{\tilde{D}} \\ 0 & I \end{bmatrix} \begin{bmatrix} M + M^T - A & 0 \\ 0 & \tau^{-1}\tilde{A}^{-1} \end{bmatrix} \begin{bmatrix} M^{-1} & 0 \\ -\Pi_{\tilde{D}}^T A M^{-1} & I \end{bmatrix},$$

and

$$B^{-1} = \begin{bmatrix} I & \Pi_{\tilde{D}} \end{bmatrix} \hat{B}^{-1} \begin{bmatrix} I \\ \Pi_{\tilde{D}}^T \end{bmatrix}.$$

Note that the preconditioner (18) can also be written in the form

$$B^{-1} = \overline{M}^{-1} + \tau^{-1}\Pi\tilde{A}^{-1}\Pi^T, \tag{23}$$

where

$$\Pi = \left(I - M^{-T}A\right)\Pi_{\tilde{D}} = \left(I - M^{-T}A\right)\left(R\tilde{D}R^T\right)^{-1} R\tilde{D}. \tag{24}$$

Comparing classical two-grid methods with the proposed auxiliary space two-grid method the main difference is that in the latter, the coarse-grid correction step is replaced by a subspace correction with iteration matrix $I - C^{-1}A$, where C is the fictitious space preconditioner defined in (21).

Remark 3
From the XZ identity [27], we have the following relation

$$\begin{aligned} \mathbf{v}^T B\mathbf{v} &= \min_{\mathbf{v}=\mathbf{w}+\Pi_{\tilde{D}}\tilde{\mathbf{w}}} \left[\tau\tilde{\mathbf{w}}^T\tilde{A}\tilde{\mathbf{w}} + \left(M^T\mathbf{w} + A\Pi_{\tilde{D}}\tilde{\mathbf{w}}\right)^T \left(M + M^T - A\right)^{-1} \left(M^T\mathbf{w} + A\Pi_{\tilde{D}}\tilde{\mathbf{w}}\right)\right] \\ &= \min_{\mathbf{v}=\mathbf{w}+\Pi_{\tilde{D}}\tilde{\mathbf{w}}} \left[\tau\|\tilde{\mathbf{w}}\|_{\tilde{A}}^2 + \|M^T\mathbf{w} + A\Pi_{\tilde{D}}\tilde{\mathbf{w}}\|_{(M+M^T-A)^{-1}}^2\right]. \end{aligned} \tag{25}$$

3. CONDITION NUMBER ESTIMATES

A condition number estimate of the two-grid preconditioner B defined by (23) and (24) can be based on the following assumptions. For the smoother, assume that

$$\underline{c}\,\langle\mathbf{v},\mathbf{v}\rangle \leqslant \rho_A\left\langle\overline{M}^{-1}\mathbf{v},\mathbf{v}\right\rangle \leqslant \bar{c}\,\langle\mathbf{v},\mathbf{v}\rangle \tag{26}$$

and

$$\|M^{-T}A\mathbf{v}\|^2 \leqslant \frac{\eta}{\rho_A}\|\mathbf{v}\|_A^2, \tag{27}$$

where $\rho_A = \lambda_{\max}(A)$ denotes the spectral radius of A and η is a non-negative constant. Further, let the operator Π defined in (24) satisfy

$$\|\Pi\tilde{\mathbf{v}}\|_A^2 \leqslant c_\Pi\|\tilde{\mathbf{v}}\|_{\tilde{A}}^2 \quad \forall\tilde{\mathbf{v}} \in \tilde{V}, \tag{28}$$

which, owing to $\|\Pi^\star\Pi\| = \|\Pi\Pi^\star\|$, is equivalent to

$$\|\Pi^\star\mathbf{v}\|_{\tilde{A}}^2 \leqslant c_\Pi\|\mathbf{v}\|_A^2 \quad \forall\mathbf{v} \in V, \tag{29}$$

where

$$\Pi^\star = \tilde{A}^{-1}\Pi^T A \tag{30}$$

denotes the adjoint operator, that is,

$$\langle\Pi\tilde{\mathbf{u}},\mathbf{v}\rangle_A = \left\langle\tilde{\mathbf{u}},\Pi^\star\mathbf{v}\right\rangle_{\tilde{A}} \quad \forall\tilde{\mathbf{u}} \in \tilde{V}, \mathbf{v} \in V. \tag{31}$$

Then the following theorem holds (cf. [26]).

DOI: 10.1002/nla

Theorem 1
Under assumptions (26)–(28), the two-grid preconditioner B defined in (23) and (24) satisfies

$$\lambda_{\max}\left(B^{-1}A\right) \leqslant \bar{c} + c_{\Pi}/\tau \tag{32}$$

and

$$\lambda_{\min}\left(B^{-1}A\right) \geqslant \frac{1}{\tau + \eta/\underline{c}}, \tag{33}$$

that is, $\kappa\left(B^{-1}A\right) \leqslant (\bar{c} + c_{\Pi}/\tau)\,(\tau + \eta/\underline{c})$.

Proof
Using (23), (26), (29), and (30), it follows that

$$\begin{aligned}\left\langle B^{-1}A\mathbf{v}, \mathbf{v}\right\rangle_A &= \left\langle \overline{M}^{-1}A\mathbf{v}, \mathbf{v}\right\rangle_A + \tau^{-1}\left\langle \Pi\tilde{A}^{-1}\Pi^T A\mathbf{v}, \mathbf{v}\right\rangle_A \\ &= \left\langle \overline{M}^{-1}A\mathbf{v}, \mathbf{v}\right\rangle_A + \tau^{-1}\left\langle \Pi^{\star}\mathbf{v}, \Pi^{\star}\mathbf{v}\right\rangle_{\tilde{A}} \\ &\leqslant \frac{\bar{c}}{\rho_A}\left\langle A\mathbf{v}, A\mathbf{v}\right\rangle + \frac{c_{\Pi}}{\tau}\|\mathbf{v}\|_A^2 \\ &\leqslant (\bar{c} + c_{\Pi}/\tau)\,\|\mathbf{v}\|_A^2,\end{aligned} \tag{34}$$

which proves (32).

On the other hand, from (26), we have

$$\left\langle \overline{M}\mathbf{v}, \mathbf{v}\right\rangle \leqslant \frac{\rho_A}{\underline{c}}\left\langle \mathbf{v}, \mathbf{v}\right\rangle, \quad \forall \mathbf{v},$$

which is equivalent to

$$\left\langle M^T\mathbf{v}, \left(M + M^T - A\right)^{-1}M^T\mathbf{v}\right\rangle \leqslant \frac{\rho_A}{\underline{c}}\left\langle \mathbf{v}, \mathbf{v}\right\rangle, \quad \forall \mathbf{v},$$

and, by setting $M^T\mathbf{v} = \mathbf{w}$, yields

$$\left\langle \mathbf{w}, \left(M + M^T - A\right)^{-1}\mathbf{w}\right\rangle \leqslant \frac{\rho_A}{\underline{c}}\left\langle M^{-T}\mathbf{w}, M^{-T}\mathbf{w}\right\rangle, \quad \forall \mathbf{w}. \tag{35}$$

Inserting $\mathbf{w} = A\mathbf{v}$ in (35) and afterwards using (27), one obtains

$$\left\langle A\mathbf{v}, \left(M + M^T - A\right)^{-1}A\mathbf{v}\right\rangle \leqslant \frac{\eta}{\underline{c}}\left\langle A\mathbf{v}, \mathbf{v}\right\rangle, \quad \forall \mathbf{v}. \tag{36}$$

Now, choose $\widetilde{\mathbf{w}} = R^T\mathbf{v}$, and let $\mathbf{v} = \mathbf{w} + \Pi_{\tilde{D}}\widetilde{\mathbf{w}} = \mathbf{w} + \Pi_{\tilde{D}}R^T\mathbf{v} = \mathbf{0} + \mathbf{v}$ be a particular decomposition of $\mathbf{v}$, that is, the one obtained for $\mathbf{w} = \mathbf{0}$. Then $\left\langle \tilde{A}\widetilde{\mathbf{w}}, \widetilde{\mathbf{w}}\right\rangle = \left\langle A\mathbf{v}, \mathbf{v}\right\rangle$, and thus from (25) and (36), it follows that

$$\begin{aligned}\left\langle B\mathbf{v}, \mathbf{v}\right\rangle &\leqslant \tau\left\langle A\mathbf{v}, \mathbf{v}\right\rangle + \|M^T\mathbf{w} + A\Pi_{\tilde{D}}R^T\mathbf{v}\|^2_{\left(M+M^T-A\right)^{-1}} \\ &= \tau\left\langle A\mathbf{v}, \mathbf{v}\right\rangle + \|A\mathbf{v}\|^2_{\left(M+M^T-A\right)^{-1}} \\ &\leqslant \left(\tau + \frac{\eta}{\underline{c}}\right)\left\langle A\mathbf{v}, \mathbf{v}\right\rangle, \quad \forall \mathbf{v},\end{aligned}$$

which proves (33). □

Remark 4
Note that when no smoothing is applied ($B = C$), the condition number estimate provided in Theorem 1 reduces to $\kappa\left(B^{-1}A\right) \leqslant c_{\Pi} = c = \|\pi_{\tilde{D}}\|^2_{\tilde{A}}$.

Now, the following theorem can be proved.

DOI: 10.1002/nla

Theorem 2
Let $\tilde{D}$ be a two-by-two block-diagonal SPD matrix, that is,

$$\tilde{D} := \begin{bmatrix} \tilde{D}_{11} & 0 \\ 0 & \tilde{D}_{22} \end{bmatrix},$$

such that for the fictitious space preconditioner (12), there holds

$$\left\langle \Pi_{\tilde{D}} \tilde{A}^{-1} \Pi_{\tilde{D}}^T \mathbf{u}, \mathbf{u} \right\rangle \leqslant c \left\langle A^{-1}\mathbf{u}, \mathbf{u} \right\rangle$$

for all $\mathbf{u} \in V$ with $c = \|\pi_{\tilde{D}}\|^2_{\tilde{A}}$.

Then the ASCA Q, as defined in (8), satisfies the relations

$$\frac{1}{c} S \leqslant Q \leqslant S, \tag{37}$$

where S is the exact Schur complement of A. Moreover, the lower bound in (37) is sharp for

$$\tilde{D} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ 0 & I \end{bmatrix}.$$

Proof
As

$$\Pi_{\tilde{D}}^T = \tilde{D} R^T \left(R \tilde{D} R^T \right)^{-1} = \begin{bmatrix} \tilde{D}_{11} R_1^T \left(R_1 \tilde{D}_{11} R_1^T \right)^{-1} & 0 \\ 0 & I_2 \end{bmatrix}, \tag{38}$$

with $\mathbf{u} = \begin{pmatrix} \mathbf{0} \\ \mathbf{w} \end{pmatrix}$, it follows that

$$\begin{pmatrix} \mathbf{0} \\ \mathbf{w} \end{pmatrix}^T \Pi_{\tilde{D}} \tilde{A}^{-1} \Pi_{\tilde{D}}^T \begin{pmatrix} \mathbf{0} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{0}} \\ \mathbf{w} \end{pmatrix}^T \tilde{A}^{-1} \begin{pmatrix} \tilde{\mathbf{0}} \\ \mathbf{w} \end{pmatrix} = \left\langle Q^{-1}\mathbf{w}, \mathbf{w} \right\rangle.$$

Moreover, $\mathbf{u}^T A^{-1} \mathbf{u} = \left\langle S^{-1}\mathbf{w}, \mathbf{w} \right\rangle$, and thus, Corollary 1 implies the estimate (37).

In the remainder of the proof, let

$$\tilde{D} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ 0 & I \end{bmatrix}.$$

In view of (38) and the relations $A_{11} = R_1 \tilde{A}_{11} R_1^T$, $A_{12} = R_1 \tilde{A}_{12}$, and $A_{21} = \tilde{A}_{21} R_1^T$, we have

$$\begin{aligned} \Pi_{\tilde{D}} \tilde{A}^{-1} \Pi_{\tilde{D}}^T &= \begin{bmatrix} \left(R_1 \tilde{A}_{11} R_1^T\right)^{-1} R_1 \tilde{A}_{11} & -\left(R_1 \tilde{A}_{11} R_1^T\right)^{-1} R_1 \tilde{A}_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{A}_{11}^{-1} & 0 \\ 0 & Q^{-1} \end{bmatrix} \\ &\times \begin{bmatrix} \tilde{A}_{11} R_1^T \left(R_1 \tilde{A}_{11} R_1^T\right)^{-1} & 0 \\ -\tilde{A}_{21} R_1^T \left(R_1 \tilde{A}_{11} R_1^T\right)^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1} A_{12} Q^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{12} Q^{-1} \\ -Q^{-1} A_{21} A_{11}^{-1} & Q^{-1} \end{bmatrix}, \end{aligned}$$

and hence

$$\mathbf{v}^T \Pi_{\tilde{D}} \tilde{A}^{-1} \Pi_{\tilde{D}}^T \mathbf{v} = \mathbf{v}^T \begin{bmatrix} I & -A_{11}^{-1} A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & Q^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21} A_{11}^{-1} & I \end{bmatrix} \mathbf{v} \tag{39}$$

and

$$\mathbf{v}^T A^{-1} \mathbf{v} = \mathbf{v}^T \begin{bmatrix} I & -A_{11}^{-1} A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21} A_{11}^{-1} & I \end{bmatrix} \mathbf{v}. \tag{40}$$

DOI: 10.1002/nla

Now, let $c = \|\pi_{\tilde{D}}\|^2_{\tilde{A}} > 1$. Then from Corollary 1, it follows that

$$\mathbf{v}^T \left(cA^{-1} - \Pi_{\tilde{D}} \tilde{A}^{-1} \Pi_{\tilde{D}}^T \right) \mathbf{v} \geqslant 0 \quad \forall \mathbf{v} \in V, \tag{41}$$

and there exists $\bar{\mathbf{v}} \in V$, $\bar{\mathbf{v}} \neq \mathbf{0}$, such that (see (15))

$$\bar{\mathbf{v}}^T \left(cA^{-1} - \Pi_{\tilde{D}} \tilde{A}^{-1} \Pi_{\tilde{D}}^T \right) \bar{\mathbf{v}} = 0.$$

Next, by using (39) and (40) in (41), it can be seen that

$$\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}^T \begin{bmatrix} (c-1)A_{11}^{-1} & 0 \\ 0 & cS^{-1} - Q^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \geqslant 0 \quad \forall \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \in V, \tag{42}$$

and further, there exists

$$\bar{\mathbf{w}} = \begin{bmatrix} \bar{\mathbf{w}}_1 \\ \bar{\mathbf{w}}_2 \end{bmatrix} \neq \mathbf{0}$$

for which (42) holds with equality. Moreover, as A_{11} is SPD and $\left(cS^{-1} - Q^{-1}\right)$ is SPSD, as (37) shows, it follows that $\bar{\mathbf{w}}_1 = \mathbf{0}$ and

$$\bar{\mathbf{w}}_2^T \left(cS^{-1} - Q^{-1} \right) \bar{\mathbf{w}}_2 = 0$$

for a certain vector $\bar{\mathbf{w}}_2 = \bar{\mathbf{v}}_2 - A_{21} A_{11}^{-1} \bar{\mathbf{v}}_1 \neq \mathbf{0}$. This, however, finally results in

$$\lambda_{\max} \left(Q^{-1} S \right) = c.$$

□

4. AUXILIARY SPACE MULTIGRID METHOD

Consider the sequence of auxiliary space stiffness matrices $\tilde{A}^k$, $k = 0, 1, \ldots, \ell - 1$. In an exact factorization form, they are as follows:

$$\left(\tilde{A}^{(k)} \right)^{-1} = \left(\tilde{L}^{(k)} \right)^T \tilde{D}^{(k)} \tilde{L}^{(k)}, \tag{43}$$

where

$$\tilde{L}^{(k)} = \begin{bmatrix} I & \\ -\tilde{A}_{21}^{(k)} \left(\tilde{A}_{11}^{(k)} \right)^{-1} & I \end{bmatrix}, \quad \tilde{D}^{(k)} = \begin{bmatrix} \left(\tilde{A}_{11}^{(k)} \right)^{-1} & \\ & Q^{(k)^{-1}} \end{bmatrix}, \tag{44}$$

and the index k refers to a particular level of mesh refinement. The matrix $Q^{(k)}$ is associated with the stiffness matrix on the next coarser level, that is,

$$A^{(k+1)} := Q^{(k)}. \tag{45}$$

The AMLI-cycle ASMG preconditioner approximating (43) is defined recursively as follows:

$$B^{(k)^{-1}} := \overline{M}^{(k)^{-1}} + \left(I - M^{(k)^{-T}} A^{(k)} \right) \Pi^{(k)} \left(\tilde{L}^{(k)} \right)^T \overline{D}^{(k)} \tilde{L}^{(k)} \Pi^{(k)^T} \left(I - A^{(k)} M^{(k)^{-1}} \right),$$

where

$$\overline{D}^{(k)} := \begin{bmatrix} \left(\tilde{A}_{11}^{(k)} \right)^{-1} & \\ & B_v^{(k+1)} \end{bmatrix} \tag{46}$$

and

$$B_v^{(\ell)} := A^{(\ell)^{-1}}. \tag{47}$$

DOI: 10.1002/nla

In the linear AMLI cycle, $B_\nu^{(k+1)}$ is a polynomial approximation of the inverse of the coarse-level matrix $A^{(k+1)} = Q^{(k)}$, that is,

$$\begin{aligned} B_\nu^{(k+1)} &:= \left(I - p^{(k)}\left(B^{(k+1)^{-1}} A^{(k+1)}\right)\right) A^{(k+1)^{-1}} \\ &=: q^{(k)}\left(B^{(k+1)^{-1}} A^{(k+1)}\right)\right) B^{(k+1)^{-1}}, \end{aligned}$$

where $p^{(k)}(t)$ is a scaled and shifted Chebyshev polynomial of degree ν_k and

$$p^{(k)}(0) = 1, \qquad q^{(k)}(t) := \frac{1 - p^{(k)}(t)}{t} \approx \frac{1}{t};$$

see [16].

In case of the nonlinear AMLI cycle, the action of $B_\nu^{(k+1)} = B_\nu^{(k+1)}[\cdot]$ on a vector defines a nonlinear mapping, which is realized by ν iterations of a Krylov subspace (here, a generalized conjugate gradient (GCG)) method, thereby utilizing the preconditioner $B^{(k+1)}$ from the coarse level. The resulting AMLI-cycle MG method is therefore sometimes referred to as a K-cycle MG (cf. [22]). The convergence analysis of the multiplicative nonlinear AMLI has first been presented in [21]. A description in the MG framework and comparative analysis can be found in [5, 22, 28]. The numerical results presented in Section 6 have been obtained on the basis of implementing the following algorithms (cf. [28]).

Given a nonlinear preconditioner $\tilde{B}^{(k)}[\cdot]$ the action of $B_{\mathrm{GCG},\nu}^{(k)}[\cdot]$, the preconditioned GCG preconditioner at level k, on a vector $\mathbf{d} \in V^{(k)}$ is defined via the following algorithm; see for example, [15].

Algorithm 1
Generalized conjugate gradient preconditioner: Definition of $B_{\mathrm{GCG},\nu}^{(k)}[\mathbf{d}]$

Step 1: $\mathbf{u}_{(0)} = \mathbf{0}, \quad \mathbf{r}_{(0)} = \mathbf{d}, \quad \mathbf{p}_{(0)} = \tilde{B}^{(k)}[\mathbf{r}_{(0)}]$

$\alpha_0 = \dfrac{\langle \mathbf{r}_{(0)}, \mathbf{p}_{(0)} \rangle}{\langle \mathbf{p}_{(0)}, A^{(k)}\mathbf{p}_{(0)} \rangle}, \quad \mathbf{u}_{(1)} = \alpha_0 \mathbf{p}_{(0)}, \quad \mathbf{r}_{(1)} = \mathbf{r}_{(0)} - \alpha_0 A^{(k)}\mathbf{p}_{(0)}$

Step 2: For $i = 1, 2, \ldots, \nu - 1$

$\beta_{ij} = \dfrac{\langle \tilde{B}^{(k)}[\mathbf{r}_{(i)}], A^{(k)}\mathbf{p}_{(j)} \rangle}{\langle \mathbf{p}_{(j)}, A^{(k)}\mathbf{p}_{(j)} \rangle}$

$\mathbf{p}_{(i)} = \tilde{B}^{(k)}[\mathbf{r}_{(i)}] - \sum_{j=0}^{i-1} \beta_{ij}\mathbf{p}_{(j)}$

$\alpha_i = \dfrac{\langle \mathbf{r}_{(i)}, \mathbf{p}_{(i)} \rangle}{\langle \mathbf{p}_{(i)}, A^{(k)}\mathbf{p}_{(i)} \rangle}$

$\mathbf{u}_{(i+1)} = \mathbf{u}_{(i)} + \alpha_i \mathbf{p}_{(i)}$

$\mathbf{r}_{(i+1)} = \mathbf{r}_{(i)} - \alpha_i A^{(k)}\mathbf{p}_{(i)}$

Step 3: $B_{\mathrm{GCG},\nu}^{(k)}[\mathbf{d}] := \mathbf{u}_{(\nu)}$

Finally, let

$$B_\nu^{(\ell)}[\cdot] = \left(A^{(\ell)}\right)^{-1}$$

and define the action $B_\nu^{(k)}[\mathbf{d}]$ of the nonlinear AMLI-cycle ASMG preconditioner $B_\nu^{(k)}[\cdot] : V^{(k)} \to V^{(k)}$ at level $k < \ell$ on a vector $\mathbf{d} \in V^{(k)}$ via the following algorithm.

DOI: 10.1002/nla

Algorithm 2
Nonlinear AMLI-cycle ASMG preconditioner: Definition of $B_\nu^{(k)}[\mathbf{d}]$

Pre-smoothing:	$\mathbf{u} = {M^{(k)}}^{-1}\mathbf{d}$
Auxiliary space correction:	$\begin{cases} \begin{pmatrix} \tilde{\mathbf{q}}_1 \\ \tilde{\mathbf{q}}_2 \end{pmatrix} := \tilde{\mathbf{q}} = \Pi_{\tilde{D}^{(k)}}^T \left(\mathbf{d} - A^{(k)}\mathbf{u}\right) \\ \tilde{\mathbf{p}}_1 = \left(\tilde{A}_{11}^{(k)}\right)^{-1} \tilde{\mathbf{q}}_1 \\ \tilde{\mathbf{p}}_2 = B_{\mathrm{GCG},\nu}^{(k+1)} \left[\left(\tilde{\mathbf{q}}_2 - \tilde{A}_{21}^{(k)}\tilde{\mathbf{p}}_1\right)\right] \\ \tilde{\mathbf{q}}_1 = \tilde{\mathbf{p}}_1 - \left(\tilde{A}_{11}^{(k)}\right)^{-1} \tilde{A}_{12}^{(k)}\tilde{\mathbf{p}}_2 \\ \tilde{\mathbf{q}}_2 = \tilde{\mathbf{p}}_2 \\ \mathbf{v} = \mathbf{u} + \Pi_{\tilde{D}^{(k)}}\tilde{\mathbf{q}} \end{cases}$
Post-smoothing:	$B_\nu^{(k)}[\mathbf{d}] := \mathbf{v} + {M^{(k)}}^{-T}\left(\mathbf{d} - A^{(k)}\mathbf{v}\right)$

At a given level k, the nonlinear AMLI-cycle ASMG method employs the GCG method with the particular preconditioner $\tilde{B}^{(k+1)}[\cdot] := B_\nu^{(k+1)}[\cdot]$ at the coarse level $k+1$.

Remark 5
For the exact two-level method, the auxiliary space correction step (at level 0) updates the approximation $\mathbf{u}$ according to

$$\mathbf{u} \leftarrow \mathbf{u} + \Pi_{\tilde{D}^{(0)}} \left(\tilde{A}^{(0)}\right)^{-1} \Pi_{\tilde{D}^{(0)}}^T \left(\mathbf{d} - A^{(0)}\mathbf{u}\right).$$

5. ESTIMATION OF $\|\pi_{\tilde{D}^{(k_0)}}\|^2_{\tilde{A}^{(k_0)}}$

In order to estimate $\|\pi_{\tilde{D}}\|^2_{\tilde{A}}$, it suffices to find an upper bound Λ for the maximum eigenvalue $\lambda_{\max}$ of

$$\pi_{\tilde{D}}^T \tilde{A} \pi_{\tilde{D}} \tilde{\mathbf{v}} = \lambda \tilde{A} \tilde{\mathbf{v}}.$$

Then $\Lambda \geqslant \lambda_{\max}$ implies $\|\pi_{\tilde{D}}\|^2_{\tilde{A}} \leqslant \Lambda$.

As

$$\tilde{A} = \sum_{G \in \mathcal{G}} \tilde{R}_G^T A_G \tilde{R}_G$$

for a certain set of restriction matrices $\{\tilde{R}_G\}$ *local estimates* can be derived by computing the maximum eigenvalues $\lambda_{G,\max}$ of the low-rank generalized eigenvalue problems

$$\pi_{\tilde{D}}^T \tilde{R}_G^T A_G \tilde{R}_G \pi_{\tilde{D}} \tilde{\mathbf{v}} = \lambda_G \tilde{A} \tilde{\mathbf{v}}, \qquad \forall G \in \mathcal{G}, \tag{48}$$

which results in

$$\lambda_{\max} \leqslant \max_{G \in \mathcal{G}} \lambda_{G,\max}\, n_{\mathrm{color}} =: \Lambda, \tag{49}$$

where n_{color} is the coloring integer constant for the adjacency graph of subdomains; two subdomains are adjacent if and only if they share at least one DOF.

As the auxiliary matrix $\tilde{A}$ is symmetric and positive definite, the generalized eigenvalue problems (48) can be equivalently written as

$$\tilde{A}^{-\frac{1}{2}} \pi_{\tilde{D}}^T \tilde{R}_G^T A_G^{\frac{1}{2}} A_G^{\frac{1}{2}} \tilde{R}_G \pi_{\tilde{D}} \tilde{A}^{-\frac{1}{2}} \tilde{\mathbf{v}} = \lambda_G \tilde{\mathbf{v}}, \qquad \forall G \in \mathcal{G}. \tag{50}$$

Finding the non-zero eigenvalues of (50), however, is equivalent to finding the eigenvalues of the small-sized eigenvalue problems

$$A_G^{\frac{1}{2}} \tilde{R}_G \pi_{\tilde{D}} \tilde{A}^{-1} \pi_{\tilde{D}}^T \tilde{R}_G^T A_G^{\frac{1}{2}} \mathbf{v}_G = \lambda_G \mathbf{v}_G, \qquad \forall G \in \mathcal{G}. \tag{51}$$

DOI: 10.1002/nla

The major remaining difficulty is the efficient inversion of the auxiliary matrix $\tilde{A}$. A cost-efficient upper bound can be computed based on the following multilevel procedure.

Consider Equation (51) for a fixed level $k_0 \in \{0, \ldots, \ell - 1\}$, that is,

$$\begin{aligned} &\text{for all } G^{(k_0)} \in \mathcal{G}^{(k_0)}, \\ &A^{\frac{1}{2}}_{G^{(k_0)}} \tilde{R}_{G^{(k_0)}} \pi_{\tilde{D}^{(k_0)}} \left(\tilde{A}^{(k_0)}\right)^{-1} \pi^T_{\tilde{D}^{(k_0)}} \tilde{R}^T_{G^{(k_0)}} A^{\frac{1}{2}}_{G^{(k_0)}} \mathbf{v}_{G^{(k_0)}} = \lambda_{G^{(k_0)}} \mathbf{v}_{G^{(k_0)}}, \end{aligned} \tag{52}$$

where $\pi_{\tilde{D}^{(k_0)}}$ is the projection operator for level k_0 and $G^{(k_0)}$ are the related subdomains.

In order to estimate the largest eigenvalue of (52), the auxiliary matrix $\left(\tilde{A}^{(k_0)}\right)^{-1}$ can be replaced by a 'larger' matrix $\left(\tilde{B}^{(k_0)}\right)^{-1}$, that is,

$$\tilde{\mathbf{v}}^T \tilde{A}^{(k_0)} \tilde{\mathbf{v}} \geqslant \tilde{\mathbf{v}}^T \tilde{B}^{(k_0)} \tilde{\mathbf{v}}, \qquad \forall \tilde{\mathbf{v}} \in \tilde{V},$$

thus considering the eigenvalue problems

$$\begin{aligned} &\text{for all } G^{(k_0)} \in \mathcal{G}^{(k_0)}. \\ &A^{\frac{1}{2}}_{G^{(k_0)}} \tilde{R}_{G^{(k_0)}} \pi_{\tilde{D}^{(k_0)}} \left(\tilde{B}^{(k_0)}\right)^{-1} \pi^T_{\tilde{D}^{(k_0)}} \tilde{R}^T_{G^{(k_0)}} A^{\frac{1}{2}}_{G^{(k_0)}} \mathbf{v}_{G^{(k_0)}} = \xi_{G^{(k_0)}} \mathbf{v}_{G^{(k_0)}}, \end{aligned} \tag{53}$$

Then, given the exact factorization (43)–(45) of the auxiliary matrix $\tilde{A}^{(k_0)}$, that is,

$$\left(\tilde{A}^{(k_0)}\right)^{-1} = \left(\tilde{L}^{(k_0)}\right)^T \begin{bmatrix} \left(\tilde{A}^{(k_0)}_{11}\right)^{-1} & \\ & A^{(k_0+1)^{-1}} \end{bmatrix} \left(\tilde{L}^{(k_0)}\right), \tag{54}$$

the left-hand side inequality in (14) implies that the following estimate holds on all levels:

$$\mathbf{v}^T \left(A^{(k)}\right)^{-1} \mathbf{v} \leqslant \mathbf{v}^T \Pi_{\tilde{D}^{(k)}} \left(\tilde{A}^{(k)}\right)^{-1} \Pi^T_{\tilde{D}^{(k)}} \mathbf{v}, \qquad \forall \mathbf{v} \in V, \quad k = 0, \ldots, \ell - 1. \tag{55}$$

Therefore, the matrix

$$\begin{aligned} \left(\tilde{B}^{(k_0)}\right)^{-1} &:= \left(\tilde{L}^{(k_0)}\right)^T \begin{bmatrix} \left(\tilde{A}^{(k_0)}_{11}\right)^{-1} & \\ & \Pi_{\tilde{D}^{(k_0+1)}} \left(\tilde{A}^{(k_0+1)}\right)^{-1} \Pi^T_{\tilde{D}^{(k_0+1)}} \end{bmatrix} \left(\tilde{L}^{(k_0)}\right) \\ &= \left(\tilde{L}^{(k_0)}\right)^T \begin{bmatrix} I & \\ & \Pi_{\tilde{D}^{(k_0+1)}} \end{bmatrix} \begin{bmatrix} \left(\tilde{A}^{(k_0)}_{11}\right)^{-1} & \\ & \left(\tilde{A}^{(k_0+1)}\right)^{-1} \end{bmatrix} \begin{bmatrix} I & \\ & \Pi^T_{\tilde{D}^{(k_0+1)}} \end{bmatrix} \left(\tilde{L}^{(k_0)}\right) \end{aligned} \tag{56}$$

can be used in (53). Note that the matrix in the middle of the right-hand side of (56) is of greater dimension than the auxiliary matrix at level k_0.

Moreover, if (55) is further applied recursively in (56) for $k = k_0 + 1, \ldots, \ell - 1$, the following multilevel estimate is obtained.

Let

$$Y^{(k)} = \begin{bmatrix} I & \\ & \Pi_{\tilde{D}^{(k)}} \end{bmatrix} \quad Z^{(k)} = \begin{bmatrix} I & \\ & \tilde{L}^{(k)} \end{bmatrix} \quad k = k_0{+}1, k_0{+}2, \ldots, \ell{-}1, \quad Y^{(\ell)} = I, \quad Z^{(k_0)} = \tilde{L}^{(k_0)}$$

and

$$X^{(k_0)} = \begin{bmatrix} \left(\tilde{A}^{(k_0)}_{11}\right)^{-1} & & & \\ & \left(\tilde{A}^{(k_0+1)}_{11}\right)^{-1} & & \\ & & \ddots & \\ & & & \left(\tilde{A}^{\ell}\right)^{-1} \end{bmatrix}.$$

DOI: 10.1002/nla

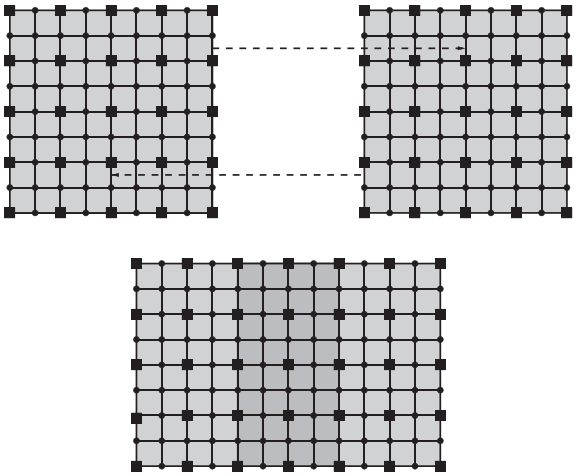


Figure 1. Two subdomains composed of 8×8 elements each overlapping with half of their width.

Then the matrix to be used in (53) can be also defined as

$$\left(\tilde{B}^{(k_0)}\right)^{-1} := \prod_{k=k_0}^{\ell-1} Z^{(k)^T} Y^{(k+1)^T} X^{(k_0)} Y^{(k+1)} Z^{(k)}. \tag{57}$$

Remark 6
Note that the computation of $\left(\tilde{B}^{(k_0)}\right)^{-1}$ requires the inversion of block diagonal matrices with a small, uniformly bounded, semi-bandwidth and a small-sized coarse-grid matrix only. Hence, solving the eigenvalue problems (53) is computationally much cheaper than solving the problems (52).

A numerical example comparing the estimates (49) with

$$\xi_{\max} \leqslant \max_{G^{(0)} \in \mathcal{G}^{(0)}} \xi_{G^{(0)},\max}\, n_{\text{color}} =: \Xi, \tag{58}$$

and $\|\pi_{\tilde{D}^{(0)}}\|^2_{\tilde{A}^{(0)}}$ is presented in the following section.

6. NUMERICAL TESTS

The presented numerical tests refer to the second-order elliptic boundary-value problem

$$-\nabla \cdot (\boldsymbol{k}(\boldsymbol{x})\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}) \text{ in } \Omega, \tag{59a}$$

$$u = 0 \quad \text{on } \Gamma, \tag{59b}$$

where the polygonal domain Ω is defined in $\mathbb{R}^2$, f is a function in $L_2(\Omega)$ and $\boldsymbol{k}(\boldsymbol{x}) = \alpha(x)I$.

Note that the imposed Dirichlet boundary conditions upon the entire boundary are not a restriction as for other boundary conditions the numerical results are quite similar.

Problem (59) is discretized using piecewise bilinear functions resulting in the linear system of algebraic equations

$$A\mathbf{u} = \mathbf{f}. \tag{60}$$

The considered mesh is uniform and consists of $n \times n$ elements (squares), where $n = 2^{\ell+2}$, that is, $n = 8, \ldots, 512$. The covering (1) of the domain is by subdomains composed of 8×8 elements (Examples 1–2) (Figure 1) or 4×4 elements (Example 3) that overlap with half of their width (and height).

The right-hand side vector of (60) has been chosen to be the vector of all zeros, and the outer nonlinear preconditioned conjugate gradient iteration has been initialized with a random vector.

Subject to numerical testing are four representative cases of problems characterized by a highly varying diffusion coefficient α, namely:

DOI: 10.1002/nla

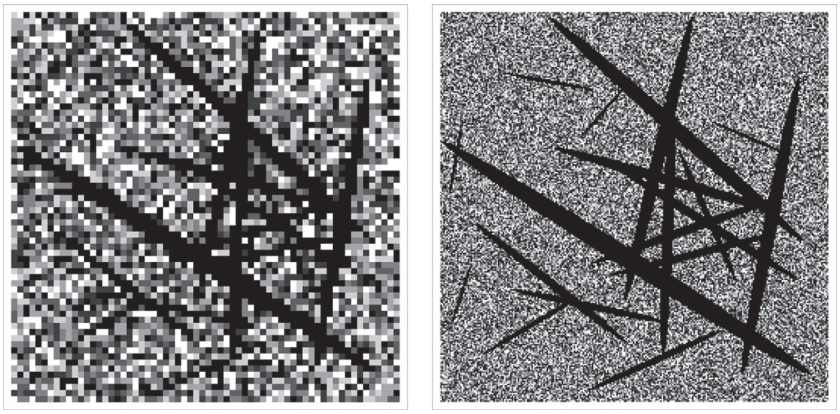
[a] a random diffusion coefficient $\alpha_e = 10^{p_{rand}}$, $p_{rand} \in \{0, 1, 2, \ldots, q\}$, that is, $\alpha_{\max}/\alpha_{\min} = 10^q$ where α_e is constant on the given element e;
[b] alternating layers of high ($\alpha_{\max}$) and low (1) permeability;
[c] islands of high permeability $\alpha_{\max} = 10^q$ against a background as in [a]; see Figure 2;
[d] islands of high permeability $\alpha_{\max} = 10^q$ against a background as in [b]; see Figure 3.

Note that test cases [b] and [d] in the present setting of full coarsening result in highly anisotropic coarse-grid problems and thus add an additional difficulty for robust preconditioning to the one introduced by the high-frequency high-contrast coefficient.

Two variants of the surjective mapping $\Pi_{\tilde{D}}$ as defined in (5) are tested numerically:

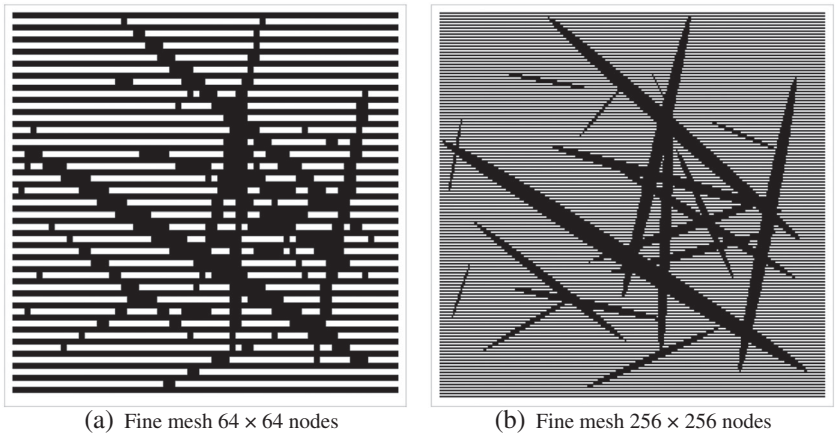
[I] $\tilde{D} = \operatorname{diag}(\tilde{A})$. Note that this choice of $\tilde{D}$ leads to a cheap computation of $\Pi_{\tilde{D}}$ as the matrix $R\tilde{D}R^T$ to be inverted becomes diagonal;
[II] $\tilde{D} = \operatorname{blockdiag}(\tilde{A})$, where the blocks are chosen in accordance to the groups of fine DOF associated with different macrostructures; in rows corresponding to coarse DOF $\tilde{D} = \operatorname{diag}(\tilde{A})$. The evaluation of $\left(R\tilde{D}R^T\right)^{-1}$ then requires an efficient preconditioner.

Example 1 (Auxiliary space two-grid method)
The first set of numerical tests (Tables I and II) shows the performance of the auxiliary space two-grid method as described and analyzed in Sections 2 and 3 for test cases [c] and [d] and $\Pi_{\tilde{D}}$ as in [I]. The size of the coarse grid and the fine grid, respectively, has been denoted by h and H, where



(a) Fine mesh 64 × 64 nodes (b) Fine mesh 256 × 256 nodes

Figure 2. Islands of high permeability $\alpha_{\max} = 10^q$ against a background as in [a].



(a) Fine mesh 64 × 64 nodes (b) Fine mesh 256 × 256 nodes

Figure 3. Islands of high permeability $\alpha_{\max} = 10^q$ against background as in [b].

DOI: 10.1002/nla

Table I. Number of iterations for residual reduction by 10^6.

2-Level method: $H = 2h$ (case [c] [I])					
q	h				
	1/16	1/32	1/64	1/128	1/256
0	9	9	9	9	9
1	10	10	10	10	10
2	10	10	10	10	10
3	10	11	11	11	11
4	10	11	11	11	11
5	10	11	11	11	11
6	10	11	11	11	11

Table II. Number of iterations for residual reduction by 10^6.

2-Level method: $H = 2h$ (case [d] [I])					
q	h				
	1/16	1/32	1/64	1/128	1/256
0	9	9	9	9	9
1	9	10	9	9	9
2	9	10	10	9	9
3	10	10	10	9	9
4	10	10	10	9	9
5	9	10	10	9	9
6	10	10	10	9	9

$H = 2h$ and h take values from the set $\{1/16, 1/32, 1/64, 1/128, 1/256\}$. In order to fully confirm the robustness of the auxiliary space preconditioner, no additional smoothing has been performed.

Example 2 (Nonlinear AMLI-cycle ASMG method)
The second set of numerical tests illustrates the performance of the nonlinear AMLI-cycle ASMG method based on the recursive application of an auxiliary space preconditioner and a point Gauss–Seidel smoother for different test cases and mapping operators. The coarsest level is $\ell = 1$, which corresponds to a uniform mesh with $2^{1+2} \times 2^{1+2} = 64$ elements and 81 coarse-grid nodes.

The finest mesh is obtained by performing $\ell - 1 = 1, \ldots, 6$ steps of uniform mesh refinement. For $\ell = 7$, the finest mesh is composed of 512×512 bilinear elements with $(512 + 1)\times(512 + 1)$ nodes. The ℓ-level V-cycle, W-cycle, and threefold V-cycle methods are tested with different choices of the parameter m indicating the number of pre-point and post-point Gauss–Seidel smoothing steps per one GCG iteration on each grid (except on the coarsest one where an exact solve is performed). That is, $m = 0$ corresponds to the case in which no smoothing is applied.

Tables III–VIII demonstrate the performance of the algorithm, variant [I], for test cases [a] and [c]. As can be seen for a moderate contrast ($q \leqslant 3$), no additional smoothing is required in order to achieve a uniform convergence. Further, the application of a point Gauss–Seidel smoother significantly improves the performance. This finally leads to an optimal order solution process for the nonlinear threefold AMLI V-cycle for any given magnitude q of the maximum contrast.

Table IX presents a comparison between variants [I] and [II] of the ℓ-level W-cycle with two pre-smoothing and post-smoothing steps for test case [b].

The obtained numerical results clearly demonstrate how crucial the choice of $\tilde{D}$ in (5) and consequently of the surjective mapping $\Pi_{\tilde{D}}$ is. As can be observed, for variant [I], the high contrast deteriorates the performance of the method. In some cases, the multilevel algorithm does not reach the prescribed accuracy within 250 iterations (denoted by * in Table IX). At the same time, the proposed ASMG algorithm with variant [II] shows robustness with respect to the contrast.

DOI: 10.1002/nla

Table III. Number of iterations for residual reduction by 10^6.

Nonlinear AMLI V-cycle (case [a] [I])																		
q	$m=1$						$m=2$						$m=3$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	4	5	6	6	7	8	4	4	5	5	6	7	3	4	4	5	6	6
1	5	5	6	6	7	8	4	4	5	5	6	7	4	4	4	5	6	6
2	5	6	6	7	7	8	4	5	5	5	6	7	4	4	5	5	6	6
3	5	6	7	8	8	8	4	5	5	6	6	6	4	4	5	5	6	6
4	5	7	8	8	9	9	4	5	6	6	7	7	4	5	5	6	6	6
5	5	7	9	10	10	13	4	5	7	8	7	10	4	5	6	7	6	8
6	6	7	9	14	14	18	5	6	7	10	10	13	4	5	7	8	8	10

Table IV. Number of iterations for residual reduction by 10^6.

Nonlinear AMLI W-cycle (case [a] [I])																		
q	$m=0$						$m=1$						$m=2$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	9	10	10	10	10	10	4	5	5	5	5	5	4	4	4	4	4	4
1	10	10	10	11	11	11	5	5	5	5	5	5	4	4	4	4	4	4
2	10	11	11	11	11	11	5	5	6	6	6	6	4	4	4	5	5	5
3	10	11	11	12	12	12	5	6	6	6	6	6	4	5	5	5	5	5
4	10	11	12	13	15	17	5	6	6	6	7	7	4	5	5	5	5	5
5	10	11	13	19	21	22	5	6	7	8	7	8	4	5	5	6	5	6
6	10	11	14	21	32	40	6	6	7	10	12	11	5	5	6	7	7	8

Table V. Number of iterations for residual reduction by 10^6.

Nonlinear AMLI 3-fold V-cycle (case [a] [I])																		
q	$m=0$						$m=1$						$m=2$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	9	10	10	10	10	10	4	5	5	5	5	5	4	4	4	4	4	4
1	10	10	10	10	10	10	5	5	5	5	5	5	4	4	4	4	4	4
2	10	10	11	11	11	11	5	5	6	6	6	6	4	4	5	5	5	5
3	10	11	11	11	11	11	5	6	6	6	6	6	4	5	5	5	5	5
4	10	11	11	12	12	13	5	6	6	6	7	7	4	5	5	5	5	5
5	10	11	12	15	17	16	5	6	7	7	7	7	4	5	5	5	5	5
6	10	11	12	15	27	28	6	6	7	7	8	8	5	5	5	6	6	6

In Tables X and XI, the ℓ-level V-cycle and W-cycle methods are tested for case [d] with a mapping operator variant [II] with different choices of the parameter m . The numerical results in Table XI show uniform convergence and robustness.

The computational expense of the ASMG iteration crucially depends on the cost of solving systems with $R\tilde{D}R^T$ that are required in each application of $\Pi_{\tilde{D}}$ and its transpose. By testing a single iteration of variant [I], we have found that the cost of one W-cycle without additional smoothing ($m = 0$) is about seven to eight, and the cost of one V-cycle about four times the cost of a single matrix–vector multiply on the finest mesh. For variant [II], when using a (scaled) one-level additive Schwarz preconditioner to solve iteratively the systems with $R\tilde{D}R^T$ on all levels (with the same tolerance as used in the outer iteration), the cost of a single ASMG iteration increases by a factor five to ten as compared to variant [I].

DOI: 10.1002/nla

Table VI. Number of iterations for residual reduction by 10^6

Nonlinear AMLI V-cycle (case [c] [I])																		
q	$m=1$						$m=2$						$m=3$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	4	5	6	6	7	8	4	4	5	5	6	7	3	4	4	5	6	6
1	5	5	6	6	7	8	4	4	5	5	6	7	4	4	4	5	6	7
2	5	5	6	7	7	8	4	5	5	5	6	7	4	4	4	5	6	6
3	5	6	6	7	7	8	4	5	6	6	6	7	4	4	5	5	6	6
4	5	6	7	8	8	9	4	5	5	6	7	7	4	4	5	5	6	6
5	5	6	7	9	10	12	4	5	7	7	7	10	4	4	6	6	6	8
6	6	7	8	12	12	17	5	5	7	8	9	12	4	4	6	7	7	10

Table VII. Number of iterations for residual reduction by 10^6.

Nonlinear AMLI W-cycle (case [c] [I])																		
q	$m=0$						$m=1$						$m=2$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	9	10	10	10	10	10	4	5	5	5	5	5	4	4	4	4	4	4
1	10	10	10	10	11	11	5	5	5	5	5	5	4	4	4	4	4	4
2	10	11	11	11	11	11	5	5	5	6	6	6	4	4	4	4	5	5
3	10	11	11	11	12	12	5	6	6	6	6	6	4	5	5	5	5	5
4	10	11	11	12	15	15	5	6	6	6	6	6	4	5	5	5	5	5
5	10	11	13	19	21	22	5	6	6	6	7	8	4	5	5	5	5	6
6	10	12	14	24	33	46	6	6	6	8	10	10	5	5	5	5	6	8

Table VIII. Number of iterations for residual reduction by 10^6.

Nonlinear AMLI 3-fold V-cycle (case [c] [I])																		
q	$m=0$						$m=1$						$m=2$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	9	10	10	10	10	10	4	5	5	5	5	5	4	4	4	4	4	4
1	10	10	10	10	10	10	5	5	5	5	5	5	4	4	4	4	4	4
2	10	10	10	11	11	11	5	5	5	6	6	6	4	4	4	4	5	5
3	10	11	11	11	11	11	5	6	6	6	6	6	4	5	5	5	5	5
4	10	11	11	11	11	12	5	6	6	6	6	6	4	5	5	5	5	5
5	10	11	12	14	17	17	5	6	6	6	7	7	4	5	5	5	5	6
6	10	11	12	17	24	36	6	6	6	7	8	8	5	5	5	5	5	6

Example 3 (Recursive estimate of $\|\pi_{\tilde{D}^{(k_0)}}\|^2_{\tilde{A}^{(k_0)}}$)
Finally, an example demonstrating the accuracy of the proposed multilevel technique for estimating $\|\pi_{\tilde{D}^{(k_0)}}\|^2_{\tilde{A}^{(k_0)}}$ is provided for test case [a] with $q = 4$, fixed level $k_0 = 0$ and mapping operator $\Pi_{\tilde{D}}$ as defined according to [I] and [II].[‡] The fine mesh in this example consists of 16×16 elements and 49 overlapping subdomains. One recursive step in (57) has been performed.

On Figures 4 and 5, the sparsity patterns of the original and auxiliary matrices at different levels are shown.

[‡]Mathematica © has been used in the presented computations.

DOI: 10.1002/nla

Table IX. Number of iterations for residual reduction by 10^6.

Nonlinear AMLI V-cycle, $m = 2$ (case [b])												
q	[I]						[II]					
	ℓ											
	2	3	4	5	6	7	2	3	4	5	6	7
0	4	4	4	4	4	4	4	4	4	4	4	4
1	4	4	4	4	4	4	4	4	4	4	4	4
2	4	4	4	4	4	4	4	4	4	4	4	4
3	4	5	6	7	7	7	4	4	4	5	5	5
4	4	9	14	19	21	20	4	4	4	4	4	5
5	4	20	54	109	*	*	4	4	4	4	4	4
6	4	25	51	114	*	*	4	4	4	4	4	4

Table X. Number of iterations for residual reduction by 10^6.

Nonlinear AMLI V-cycle (case [d] [II])																		
q	$m = 1$						$m = 2$						$m = 3$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	4	5	5	6	8	8	4	4	5	5	5	7	3	4	4	5	6	6
1	5	5	6	7	8	8	4	4	5	5	7	7	3	4	4	5	6	7
2	5	5	7	8	10	10	4	4	6	7	8	10	3	4	5	6	7	9
3	5	6	8	10	11	12	4	5	7	8	10	11	3	4	6	8	9	10
4	5	6	8	10	12	14	4	5	7	9	11	11	3	4	7	9	10	11
5	5	6	8	10	12	15	4	5	7	9	11	13	3	4	7	9	11	13
6	5	6	8	10	12	15	4	5	7	9	11	13	3	4	7	9	11	13

Table XI. Number of iterations for residual reduction by 10^6.

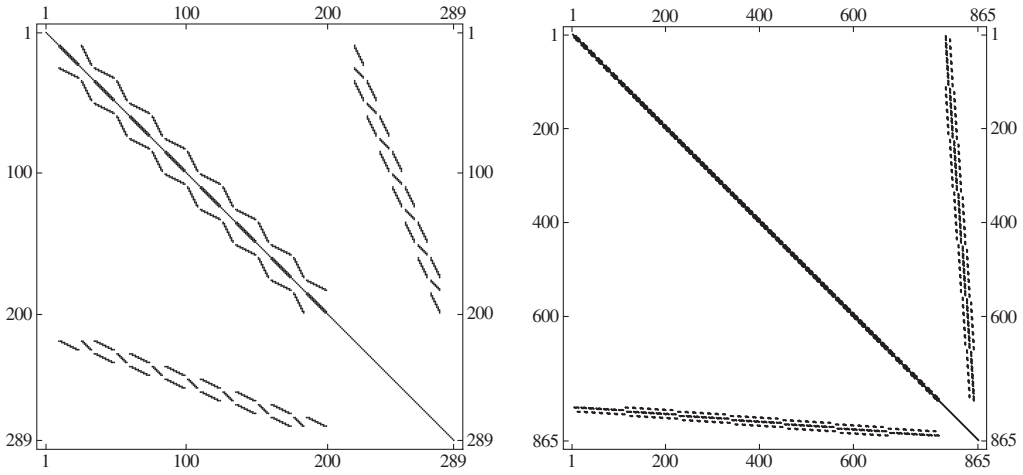
Nonlinear AMLI W-cycle (case [d] [II])																		
q	$m = 0$						$m = 1$						$m = 2$					
	ℓ																	
	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
0	9	9	9	9	9	9	5	5	5	5	5	5	4	4	4	4	4	4
1	9	10	10	10	10	10	5	5	5	5	5	5	4	4	4	4	4	4
2	9	10	10	10	10	10	5	5	5	5	5	5	4	4	4	4	4	4
3	9	10	10	10	10	10	5	5	5	5	6	6	4	4	4	4	5	5
4	9	10	10	10	11	11	5	5	5	5	6	6	4	4	4	5	5	5
5	9	10	10	10	11	11	5	5	5	6	6	6	4	4	4	5	5	5
6	9	10	10	10	11	11	5	5	5	6	6	6	4	4	4	5	5	5

The coloring integer constant in this example is $n_{\text{color}} = 9$. In order to obtain a tight upper bound for the maximum eigenvalue in (52) and (53), one can assume that the subdomains touching the boundary further overlap with degenerated subdomains of smaller size. Note that in this case n_{color} does not change. Further, it is sufficient to solve (52) and (53) only for the non-degenerated subdomains, that is, the number of local eigenvalue problems does not increase.

On Figure 6, the maximum eigenvalues of (52) and (53) are depicted for the two variants [I] and [II] of projection operators for which it is found that

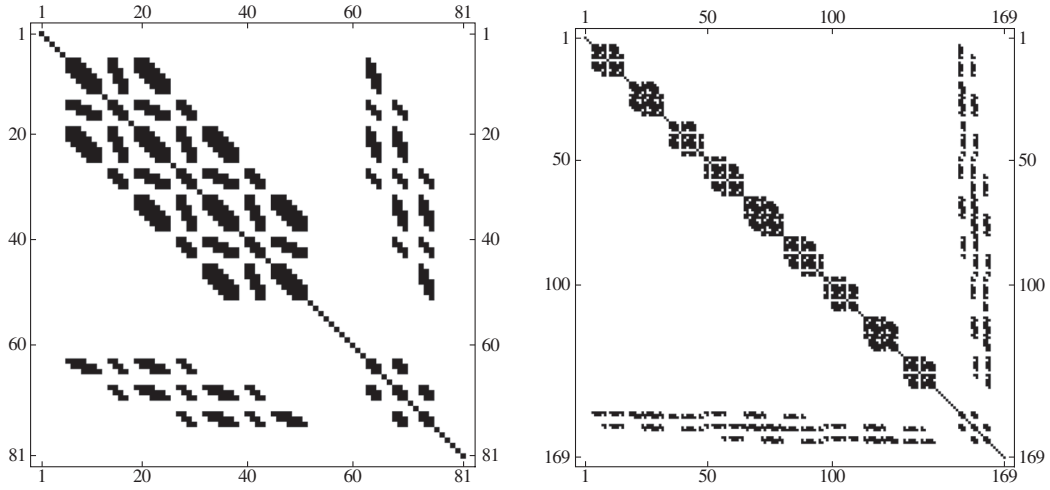
$$\max_{G\in\mathcal{G}} \lambda^{[I]}_{G,max} = 0.515764, \qquad \max_{G\in\mathcal{G}} \xi^{[I]}_{G,max} = 0.590758,$$

DOI: 10.1002/nla



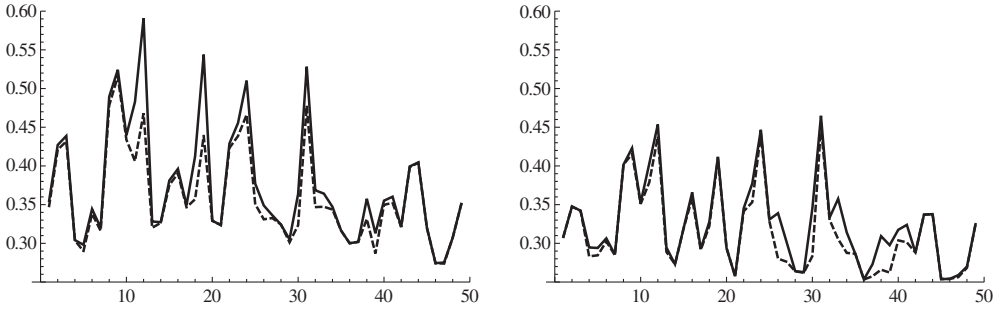
(a) Original matrix (b) Auxiliary matrix

Figure 4. Sparsity pattern of the fine grid matrices.



(a) Original matrix (b) Auxiliary matrix

Figure 5. Sparsity pattern of the coarser grid matrices.



(a) Variant [I]. (b) Variant [II].

Figure 6. Distribution of the maximum eigenvalues of (53) (solid lines) and of (52) (dashed lines).

DOI: 10.1002/nla

$$\max_{G\in\mathcal{G}} \lambda_{G,max}^{[II]} = 0.450956, \qquad \max_{G\in\mathcal{G}} \xi_{G,max}^{[II]} = 0.464827.$$

The computed norms of the projections are

$$\|\pi_{\tilde{D}^{(k_0)}}^{[I]}\|_{\tilde{A}^{(k_0)}}^2 = 2.1893390511486, \qquad \|\pi_{\tilde{D}^{(k_0)}}^{[II]}\|_{\tilde{A}^{(k_0)}}^2 = 1.9827749765716.$$

Evaluating the respective estimates gives

$$\Lambda^{[I]} = 4.64184, \qquad \Lambda^{[II]} = 4.058604,$$

$$\Xi^{[I]} = 5.316822, \qquad \Xi^{[II]} = 4.183443,$$

where $\Xi^{[I]}$ and $\Xi^{[II]}$ correspond to (53), whereas $\Lambda^{[I]}$ and $\Lambda^{[II]}$ are for (52); see also (49).

7. CONCLUSIONS

A new MG method employing an auxiliary space and an ASCA has been constructed and analyzed. The presented condition number estimate for the two-grid preconditioner implies robust convergence of the related two-grid method. Also established has been the spectral equivalence between the ASCA and the exact Schur complement. The upper bound in this relation is always sharp. The lower bound is given in terms of the energy norm of a projection operator that involves an SPD block-diagonal matrix $\tilde{D}$. Further, for a particular choice of $\tilde{D}$ also, the lower bound has been shown to be sharp. Its efficient computation has been addressed, and a particular multilevel algorithm has been proposed for this purpose.

A main contribution of this work is the definition and formulation of an AMLI-cycle ASMG method, which differs from classical MG methods in replacing coarse-grid correction by auxiliary space correction. A representative collection of numerical test has been presented. The obtained numerical results not only demonstrate the efficiency of the proposed algorithm but also reveal possibilities for further development, for example, incorporating different smoothers and transfer mappings or shifting the focus to different problem classes.

Although not in the scope of this study, it should be mentioned that the proposed ASMG method is suitable for implementation on parallel computer architectures.

ACKNOWLEDGEMENTS

This work has been supported by the Austrian Science Fund, grant P22989; the Bulgarian Science Fund, grant DCVP 02/01; and the Project AComIn, grant 316087, funded by the FP7 Capacity Programme.

REFERENCES

1. Mathew TPA. Lecture Notes in Computational Science and Engineering. *Domain decomposition methods for the numerical solution of partial differential equations* 2008; **61**.
2. Toselli A, Widlund O. *Domain Decomposition methods—algorithms and theory*, Springer Series in Computational Mathematics. Springer, 2001.
3. Hackbusch W. *Multi-grid methods and applications*. Springer: Berlin-Heidelberg-New York-Tokyo, 1985.
4. Trottenberg U, Oosterlee C W, Schüller A. *Multigrid*. Academic Press Inc.: San Diego, CA, 2005.
5. Vassilevski P. *Multilevel Block Factorization Preconditioners*. Springer: New York, 2008.
6. Graham IG, Lecher PO, Scheichl R. Domain decomposition for multiscale PDEs. *Numerische Mathematik* 2007; **106**(4):489–626.
7. Galvis J, Efendiev Y. Domain decomposition preconditioners for multiscale flows in high-contrast media. *Multiscale Modelling and Simulation* 2010; **8**(4):1461–1483.
8. Galvis J, Efendiev Y. Domain decomposition preconditioners for multiscale flows in high-contrast media: reduced dimension coarse spaces. *Multiscale Modelling and Simulation* 2010; **8**(5):1621–1644.
9. Efendiev Y, Galvis J, Lazarov R, Willems J. Robust domain decomposition preconditioners for abstract symmetric positive definite bilinear forms. *Mathematical Modelling and Numerical Analysis* 2012; **46**:1175–1199.
10. Spillane N, Dolean V, Hauret P, Nataf F, Pechstein C, Scheichl R. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numerische Mathematik* 2014; **126**:741–770.

DOI: 10.1002/nla

11. Scheichl R, Vassilevski P, Zikatanov L. Weak approximation properties of elliptic projections with functional constraints. *Multiscale Modelling and Simulation* 2011; **9**(4):1677–1699.
12. Kraus J, Margenov S. *Robust Algebraic Multilevel Methods and Algorithms*: De Gruyter, Berlin, Germany, 2009.
13. Axelsson O, Vassilevski P. Algebraic multilevel preconditioning methods I. *Numerische Mathematik* 1989; **56**: 1569–1590.
14. Axelsson O, Vassilevski P. Algebraic multilevel preconditioning methods II. *SIAM Journal on Numerical Analysis* 1990; **27**(6):1569–1590.
15. Axelsson O, Vassilevski P. Variable-step multilevel preconditioning methods, I: self-adjoint and positive definite elliptic problems. *Numerical Linear Algebra with Applications* 1994; **1**:75–101.
16. Kraus J, Vassilevski P, Zikatanov L. Polynomial of best uniform approximation to 1/x and smoothing for two-level methods. *Computational Methods in Applied Mathematics* 2012; **12**:448–468.
17. Kraus J. Algebraic multilevel preconditioning of finite element matrices using local Schur complements. *Numerical Linear Algebra with Applications* 2006; **13**:49–70.
18. Kraus J. Additive Schur complement approximation and application to multilevel preconditioning. *SIAM Journal on Scientific Computing* 2012; **34**(6):A2872–A2895.
19. Kraus J, Lymbery M, Margenov S. Robust multilevel methods for quadratic finite element anisotropic elliptic problems. *Numerical Linear Algebra with Applications* 2014; **21**:375–398.
20. Kuznetsov Y. Algebraic multigrid domain decomposition methods. *Soviet Journal on Numerical Analysis and Mathematical Modelling* 1989; **4**(5):351–379.
21. Kraus J. An algebraic preconditioning method for M-matrices: linear versus non-linear multilevel iteration. *Numerical Linear Algebra with Applications* 2002; **9**:599–618.
22. Notay Y, Vassilevski P. Recursive Krylov-based multigrid cycles. *Numerical Linear Algebra with Applications* 2008; **15**:473–487.
23. Axelsson O, Blaheta R, Neytcheva M. Preconditioning of boundary value problems using elementwise Schur complements. *SIAM Journal on Matrix Analysis and Applications* 2009; **31**:767–789.
24. Nepomnyaschikh S. Mesh theorems on traces, normalizations of function traces and their inversion. *Russian Journal of Numerical Analysis and Mathematical Modelling* 1991; **6**:223–242.
25. Kato T. Estimation of iterated matrices, with application to the von Neumann condition. *Numerische Mathematik* 1960; **2**(1):22–29.
26. Xu J. The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids. *Computing* 1996; **56**:215–235.
27. Xu J, Zikatanov L. The method of alternating projections and the method of subspace corrections in Hilbert space. *Journal of the American Mathematical Society* 2002; **15**:573–597.
28. Hu X, Vassilevski P, Xu J. Comparative convergence analysis of nonlinear AMLI-cycle multigrid. *SIAM Journal on Numerical Analysis* 2013; **51**(2):1349–1369.

DOI: 10.1002/nla

AUXILIARY SPACE MULTIGRID METHOD FOR ELLIPTIC PROBLEMS WITH HIGHLY VARYING COEFFICIENTS

Auxiliary space multigrid method for elliptic problems with highly varying coefficients

Johannes Kraus[1] and Maria Lymbery[2]

1 Introduction

The robust preconditioning of linear systems of algebraic equations arising from discretizations of partial differential equations (PDE) is a fastly developing area of scientific research. In many applications these systems are very large, sparse and therefore it is vital to construct (quasi-)optimal iterative methods that converge independently of problem parameters.

The most established techniques to accomplish this objective are domain decomposition (DD), see, e.g., Toselli and Widlund [2005], Mathew [2008], and multigrid (MG)/algebraic multilevel iteration (AMLI) methods, see, e.g., Hackbusch [2003], Trottenberg et al. [2001], Vassilevski [2008].

As demonstrated by Klawonn et al. [2002], Toselli and Widlund [2005], Graham et al. [2007], two-level DD methods can be proven to be robust for scalar elliptic PDE with varying coefficient if the variations of the coefficient inside the coarse grid cells are assumed to be bounded. A key tool in the classical analysis of overlapping DD methods is the Poincaré inequality or its weighted analog as for problems with highly varying coefficients. It is well-known that the weighted Poincaré inequality holds only under certain conditions, e.g., in case of quasi-monotonic coefficients, see Sarkis [1994]. The concept of quasi-monotonic coefficients has been further developed in Pechstein and Scheichl [2008] for the convergence analysis of finite element tearing and interconnecting (FETI) methods.

Recently the robustness of DD methods has also been achieved for problems with general coefficient variations using coarse spaces that are constructed by solving local generalized eigenvalue problems, see, e.g., Efendiev et al. [2012], Galvis and Efendiev [2010], Spillane et al. [2014].

RICAM, Altenberger Str. 69, 4040 Linz, Austria johannes.kraus@oeaw.ac.at · IICT, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 25A, 1113 Sofia, Bulgaria mariq@parallel.bas.bg

In view of computational complexity, MG methods have been known to be most efficient as they have demonstrated optimality with respect to the problem size, see Hackbusch [2003], Vassilevski [2008]. Their design, however, needs careful adaptation for problems with large variations in the physical problem parameters. The AMLI framework contributes in achieving this goal, e.g. by providing more general polynomial acceleration techniques or Krylov cycles, see Axelsson and Vassilevski [1989, 1990, 1994], Kraus et al. [2012].

The idea of integrating domain decomposition techniques into multigrid methods can be found as early as in Kuznetsov [1989]. The method that is presented in the following combines DD and MG techniques with those from auxiliary space preconditioning, see Xu [1996]. It is related to substructuring methods like FETI, see Farhat and Roux [1991], and balancing domain decomposition (BDD) methods, see Mandel [1993].

The most advanced of these methods, BDDC (BDD based on constraints), see Dohrmann [2003], and FETI-DP (FETI dual-primal), see Farhat et al. [2001], can be formulated and analyzed in a common algebraic framework, see Mandel and Dohrmann [2003], Mandel et al. [2005], Mandel and Sousedík [2007]. The BDDC method enforces continuity across substructure interfaces by a certain averaging operator. The additional constraints can be interpreted as subspace corrections where coarse basis functions are subject to energy minimization. From this point of view the BDDC method has a high degree of similarity with the present approach.

However, contrary to BDDC, the auxiliary space multigrid (ASMG) method considered here naturally allows overlapping of subdomains and coarse degrees of freedom (DOF) are associated in general not only with the interfaces of subdomains but also with their interior. Moreover, the aim is to define a full multilevel method by recursive application of a two-level method. In contrary to standard (variational) multigrid algorithms coarse-grid correction is replaced by an auxiliary space correction. The coarse-grid operator then appears as the exact Schur complement of the auxiliary matrix and defines an additive approximation of the Schur complement of the original system, see Kraus [2006, 2012].

The purpose of the present paper is to summarize the main steps of the construction of the ASMG method recently proposed in Kraus et al. [2014] on a less technical level (Sections 2 and 4) and further to discuss its spectral properties and robustness with respect to highly varying coefficients (Section 3). The latter issue is also illustrated by numerical tests (Section 5).

2 Auxiliary space two-grid preconditioner

Consider the linear system of algebraic equations

$$A\mathbf{u} = \mathbf{f} \tag{1}$$

obtained after a finite element (FE) discretization of a partial differential equation (PDE) defined over a domain Ω, where A denotes the global stiffness matrix and $\mathbf{f}$ is a given right-hand side vector.

Consider a covering of Ω by n (overlapping) subdomains Ω_i, i.e., $\overline{\Omega} = \bigcup_{i=1}^{n} \overline{\Omega}_i$. Assume that for each subdomain Ω_i there is a symmetric positive semi-definite (SPSD) subdomain matrix A_i and that $A = \sum_{i=1}^{n} R_i^T A_i R_i$ where R_i restricts a global vector $\mathbf{v} \in V = \mathbb{R}^N$ to the local space $V_i = \mathbb{R}^{n_i}$ related to Ω_i. In practice the matrices A_i are assembled from scaled element matrices where the scaling factors account for the overlap of the subdomains. The DOF are split into two groups, coarse and fine, and the matrices A and A_i are partitioned accordingly into two-by-two blocks, where the lower right blocks (with index 22) are associated with coarse DOF, i.e.,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad A_i = \begin{bmatrix} A_{i:11} & A_{i:12} \\ A_{i:21} & A_{i:22} \end{bmatrix}, \quad i = 1, \ldots, n.$$

Introduce the following auxiliary domain decomposition matrix

$$\widetilde{A} = \begin{bmatrix} A_{1:11} & & & & A_{1:12}R_{1:2} \\ & A_{2:11} & & & A_{2:12}R_{2:2} \\ & & \ddots & & \vdots \\ & & & A_{n:11} & A_{n:12}R_{n:2} \\ R_{1:2}^T A_{1:21} & R_{2:2}^T A_{2:21} & \ldots & R_{n:2}^T A_{n:21} & \sum_{i=1}^{n} R_{i:2}^T A_{i:22} R_{i:2} \end{bmatrix}. \tag{2}$$

Denote $\widetilde{A}_{11} = \operatorname{diag}\{A_{1:11}, \ldots, A_{n:11}\}$, $\widetilde{A}_{22} = A_{22} = \sum_{i=1}^{n} R_{i:2}^T A_{i:22} R_{i:2}$, i.e., $\widetilde{A} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix}$. The matrices $A \in \mathbb{R}^{N \times N}$ and $\widetilde{A} \in \mathbb{R}^{\widetilde{N} \times \widetilde{N}}$ are related via

$$A = R\widetilde{A}R^T \text{ where } R = \begin{bmatrix} R_1 & 0 \\ 0 & I_2 \end{bmatrix}, \; R_1 = \begin{bmatrix} R_{1:1}^T & \ldots & R_{n:1}^T \end{bmatrix}, \; A_{11} = R_1 \widetilde{A}_{11} R_1^T.$$

Definition 1. (Kraus [2012]) The additive Schur complement approximation (ASCA) of $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is defined as the Schur complement Q of $\widetilde{A}$:

$$Q := \widetilde{A}_{22} - \widetilde{A}_{21}\widetilde{A}_{11}^{-1}\widetilde{A}_{12} = \sum_{i=1}^{n} R_{i:2}^T (A_{i:22} - A_{i:21}A_{i:11}^{-1}A_{i:12}) R_{i:2} \tag{3}$$

Next define a surjective mapping $\mathit{\Pi}_{\widetilde{D}} : \widetilde{V} \to V$ by

$$\mathit{\Pi}_{\widetilde{D}} = (R\widetilde{D}R^T)^{-1} R\widetilde{D}, \tag{4}$$

where $\widetilde{V} = \mathbb{R}^{\widetilde{N}}$ and $\widetilde{D}$ is a two-by-two block-diagonal SPD matrix.

The proposed auxiliary space two-grid preconditioner is defined by

$$B^{-1} := \overline{M}^{-1} + (I - M^{-T}A)C^{-1}(I - AM^{-1}) \tag{5}$$

where the operator M in (5) denotes an A-norm convergent smoother, i.e. $\|I - M^{-1}A\|_A \leq 1$, and $\overline{M} = M(M + M^T - A)^{-1}M^T$ is the corresponding symmetrized smoother. The matrix C defines a fictitious (auxiliary) space preconditioner approximating A and is given by

$$C^{-1} = \Pi_{\widetilde{D}} \widetilde{A}^{-1} \Pi_{\widetilde{D}}^T. \tag{6}$$

Denote $\Pi = (I - M^{-T}A)\Pi_{\widetilde{D}} = (I - M^{-T}A)(R\widetilde{D}R^T)^{-1}R\widetilde{D}$, then the preconditioner (5) can also be presented as

$$B^{-1} = \overline{M}^{-1} + \Pi \widetilde{A}^{-1} \Pi^T. \tag{7}$$

The proposed auxiliary space two-grid method differs from the classical two-grid methods in the replacement of the coarse grid correction step by a subspace correction with iteration matrix $I - C^{-1}A$.

3 Spectral properties and robustness

As it has been shown in Kraus et al. [2014] the condition number of the two-grid preconditioner defined in (7) satisfies the estimate

$$\kappa(B^{-1}A) \leq (\bar{c} + c_\Pi)(\underline{c} + \eta)/\underline{c},$$

where $\rho_A = \lambda_{\max}(A)$, c_Π is the constant in the estimate $\|\Pi\tilde{\mathbf{v}}\|_A^2 \leq c_\Pi \|\tilde{\mathbf{v}}\|_{\widetilde{A}}^2$ for all $\tilde{\mathbf{v}} \in \widetilde{V}$, and the constants $\bar{c}$, $\underline{c}$ and η characterize the smoother, i.e.,

$$\underline{c}\langle \mathbf{v}, \mathbf{v}\rangle \leq \rho_A \langle \overline{M}^{-1}\mathbf{v}, \mathbf{v}\rangle \leq \bar{c}\langle \mathbf{v}, \mathbf{v}\rangle \quad \text{and} \quad \|M^{-T}A\mathbf{v}\|^2 \leq \frac{\eta}{\rho_A}\|\mathbf{v}\|_A^2.$$

Moreover, the ASCA defined in (3) is spectrally equivalent to S, i.e. $Q \simeq S$:

Theorem 1. *(Kraus et al. [2014]) Denote $\pi_{\widetilde{D}} = R^T\Pi_{\widetilde{D}}$ where $\Pi_{\widetilde{D}}$ is defined as in* (4) *and $\widetilde{D}$ is an arbitrary two-by-two block-diagonal SPD matrix for the same fine-coarse partitioning of DOF as used in the construction of $\widetilde{A}$. Then $\langle A^{-1}\mathbf{u}, \mathbf{u}\rangle \leq \langle \Pi_{\widetilde{D}}\widetilde{A}^{-1}\Pi_{\widetilde{D}}^T\mathbf{u}, \mathbf{u}\rangle \leq c\,\langle A^{-1}\mathbf{u}, \mathbf{u}\rangle \ \ \forall \mathbf{u} \in V$ where $c := \|\pi_{\widetilde{D}}\|_{\widetilde{A}}^2$. Hence,*

$$\frac{1}{c}\langle S\mathbf{v}_2, \mathbf{v}_2\rangle \leq \langle Q\mathbf{v}_2, \mathbf{v}_2\rangle \leq \langle S\mathbf{v}_2, \mathbf{v}_2\rangle \quad \forall \mathbf{v}_2. \tag{8}$$

The upper bound in (8) *is sharp, the lower bound is sharp for* $\widetilde{D} = \begin{bmatrix} \widetilde{A}_{11} & 0 \\ 0 & I \end{bmatrix}$.

To verify that $\langle S\mathbf{v}_2, \mathbf{v}_2\rangle \le c\langle Q\mathbf{v}_2, \mathbf{v}_2\rangle$ is robust with respect to an arbitrary variation of an elementwise constant coefficient $\alpha(\boldsymbol{x}) = \alpha_e$ for all $\boldsymbol{x} \in e$ and all elements e, see (15), one has to consider all possible distributions of $\{\alpha_e\}$ on the finest mesh. However, in the following we will show that the worst condition number (largest values of c) is obtained for a certain binary distribution of $\{\alpha_e\}$ so it suffices to study distributions of this type.

Let n_e denote the number of elements e and consider first an arbitrary distribution $\{\alpha_e\}$ of a piecewise constant coefficient where $\alpha_e \in (0, 1]$ for all e. Further, let A denote the global stiffness matrix corresponding to this distribution. Then there exists a set of binary distributions $\{\mathcal{C}_i : i = 1, 2, \ldots, n_e\}$ with $\mathcal{C}_i = \{\alpha_{e_j} : j = 1, 2, \ldots, n_e,\ \alpha_{e_j} = \beta_{e_i}$ if $j = i$ and $\alpha_{e_j} = \delta$ else$\}$ for some constants $0 < \delta \le \beta_{e_i} \le 1$ such that $A = \sum_{i=1}^{n_e} A_i$ where A_i is the global stiffness matrix corresponding to the distribution $\mathcal{C}_i$. It is easy to see that if A is SPD then A_i is SPD for all i. Now, let S_i denote the exact Schur complement of A_i and S be the Schur complement of A. Moreover, let Q_i denote the ASCA corresponding to A_i, i.e., $Q_i \simeq S_i$ where Q_i is the exact Schur complement of $\widetilde{A}_i$, cf. (2).

Lemma 1. *Using the above notation, assume that*

$$\frac{1}{c_j}\langle S_j\mathbf{v}_2, \mathbf{v}_2\rangle \le \langle Q_j\mathbf{v}_2, \mathbf{v}_2\rangle \le \langle S_j\mathbf{v}_2, \mathbf{v}_2\rangle \quad \forall \mathbf{v}_2 \text{ and } j = 1, \ldots, n_e. \tag{9}$$

Further, denote $c_{\max} = \max_{i\in\{1,\ldots,n_e\}}\{c_i\}$. *Then the following relations hold:*

$$\frac{1}{c_{\max}}\langle S\mathbf{v}_2, \mathbf{v}_2\rangle \le \langle Q\mathbf{v}_2, \mathbf{v}_2\rangle \le \langle S\mathbf{v}_2, \mathbf{v}_2\rangle \quad \forall \mathbf{v}_2. \tag{10}$$

Proof. The right inequality in (10) follows directly from the energy minimization property of Schur complements. In order to prove the left inequality we assume that (10) is wrong. Then there exists a vector $\mathbf{v}_2 \neq \mathbf{0}$ such that $\mathbf{v}_2^T S\mathbf{v}_2 \ge \bar{c}\,\mathbf{v}_2^T Q\mathbf{v}_2 > c_{\max}\mathbf{v}_2^T Q\mathbf{v}_2$, the left inequality of which can also be written in the form $\min_{\mathbf{v}_1} \begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix}^T A \begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix} \ge \bar{c}\,\mathbf{v}_2^T Q\mathbf{v}_2$, or, equivalently as $\min_{\mathbf{v}_1} \begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix}^T \left(\sum_{j=1}^{n_e} A_j\right) \begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix} \ge \bar{c}\, \min_{\tilde{\mathbf{v}}_1} \begin{pmatrix}\tilde{\mathbf{v}}_1\\ \mathbf{v}_2\end{pmatrix}^T \left(\sum_{j=1}^{n_e} \tilde{A}_j\right) \begin{pmatrix}\tilde{\mathbf{v}}_1\\ \mathbf{v}_2\end{pmatrix}$. From the latter inequality it follows that

$$\begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix}^T \left(\sum_{j=1}^{n_e} A_j\right) \begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix} \ge \bar{c} \sum_{j=1}^{n_e} \min_{\tilde{\mathbf{v}}_1} \begin{pmatrix}\tilde{\mathbf{v}}_1\\ \mathbf{v}_2\end{pmatrix}^T \tilde{A}_j \begin{pmatrix}\tilde{\mathbf{v}}_1\\ \mathbf{v}_2\end{pmatrix} \quad \forall \mathbf{v}_1,$$

which is equivalent to

$$\sum_{j=1}^{n_e} \begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix}^T A_j \begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix} \ge \bar{c} \sum_{j=1}^{n_e} \mathbf{v}_2^T Q_j\mathbf{v}_2 \quad \forall \mathbf{v}_1. \tag{11}$$

Then, since all matrices A_j and Q_j are SPSD, it follows from (11) that there exists at least one index $j_0 \in \{1, 2, \ldots, n_e\}$ such that

$$\begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix}^T A_{j_0} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix} \geq \bar{c}\, \mathbf{v}_2^T Q_{j_0} \mathbf{v}_2 \quad \forall \mathbf{v}_1.$$

Hence $\mathbf{v}_2^T S_{j_0} \mathbf{v}_2 = \min_{\mathbf{v}_1} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix}^T A_{j_0} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix} \geq \bar{c}\, \mathbf{v}_2^T Q_{j_0} \mathbf{v}_2$ which is in contradiction to (9) since $\bar{c} > c_{\max}$.

A crucial step in the application of the two-level preconditioner is the realization of the operator $\mathit{\Pi}_{\widetilde{D}}$. We propose two different variants that correspond to the following choices of $\widetilde{D}$:

[I] $\widetilde{D} = \mathrm{diag}(\widetilde{\mathrm{A}})$;

[II] $\widetilde{D} = \mathrm{blockdiag}(\widetilde{\mathrm{A}})$. The diagonal blocks are determined by the groups of fine DOF related to different macro structures whereas $\widetilde{D} = \mathrm{diag}(\widetilde{\mathrm{A}})$ in rows corresponding to coarse DOF.

In variant [I] the matrix $R\widetilde{D}R^T$ is diagonal, which makes the application of $\mathit{\Pi}_{\widetilde{D}}$ notably simple and cost-efficient. In case of variant [II] the action of $(R\widetilde{D}R^T)^{-1}$ can be implemented via an inner iterative method such as a preconditioned conjugate gradient (PCG) method, which then for reasons of efficiency requires a uniform preconditioner. A possible candidate is the one-level additive Schwarz (AS) preconditioner which however has to be adapted in order to be robust with respect to coefficient jumps. For this reason we study the scaled one-level AS preconditioner B_{AS} defined via

$$B_{\mathrm{AS}}^{-1} = SR\widetilde{S}^{-1}(\widetilde{S}\widetilde{D}\widetilde{S})^{-1}\widetilde{S}^{-1}R^T S \tag{12}$$

which can be applied to the scaled system with the matrix

$$D_s = SDS = SR\widetilde{D}R^T S,$$

where $S = [\mathrm{diag}(\mathrm{A})]^{-1/2}$, if the result is then rescaled. Let us further denote

$$\widetilde{D}_s = \widetilde{S}\widetilde{D}\widetilde{S} \;\text{ and }\; R_s = SR\widetilde{S}^{-1} \text{ where } \widetilde{S} = [\mathrm{diag}(\widetilde{\mathrm{A}})]^{-1/2}.$$

Then the following lemma holds:

Lemma 2. *The condition number of the preconditioned system using the scaled one-level AS preconditioner satisfies the estimate*

$$\kappa(B_{\mathrm{AS}}^{-1}D_s) \leq \kappa(\widetilde{D}_s). \tag{13}$$

Proof. First we show that $\lambda_{\min}(B_{\mathrm{AS}}^{-1}D_s) \geq 1$. Note that $D_s = R_s\widetilde{D}_s R_s^T$ and

$$R_s R_s^T = SR\widetilde{S}^{-1}\widetilde{S}^{-1}R^T S = [\mathrm{diag}(A)]^{-1/2} R\,[\mathrm{diag}(\widetilde{A})] R^T [\mathrm{diag}(A)]^{-1/2} = I.$$

Consider next the matrix

$$\begin{bmatrix} R_s\widetilde{D}_sR_s^T & I \\ I & R_s\widetilde{D}_s^{-1}R_s^T \end{bmatrix} = \begin{bmatrix} R_s & 0 \\ 0 & R_s \end{bmatrix} \begin{bmatrix} \widetilde{D}_s & I \\ I & \widetilde{D}_s^{-1} \end{bmatrix} \begin{bmatrix} R_s^T & 0 \\ 0 & R_s^T \end{bmatrix}$$

which is SPSD with an SPD pivot block $D_s = R_s\widetilde{D}_sR_s^T$. Consequently, its Schur complement is an SPSD matrix, i.e.

$$R_s\widetilde{D}_s^{-1}R_s^T - (R_s\widetilde{D}_sR_s^T)^{-1} \geq 0$$

which proves that $\lambda_{\min}(B_{\mathrm{AS}}^{-1}D_s) \geq 1$.

On the other hand we have

$$\begin{aligned} \lambda_{\max}(B_{\mathrm{AS}}^{-1}D_s) &= \lambda_{\max}(R_s\widetilde{D}_s^{-1}R_sD_s) \\ &= \lambda_{\max}(D_s^{1/2}R_s\widetilde{D}_s^{-1}R_sD_s^{1/2}) \\ &= \lambda_{\max}(\widetilde{D}_s^{-1/2}R_s^TD_sR_s\widetilde{D}_s^{-1/2}) \\ &\leq \lambda_{\max}(\widetilde{D}_s^{-1})\lambda_{\max}(R_s^TR_s\widetilde{D}_sR_s^TR_s) \\ &\leq \lambda_{\max}(\widetilde{D}_s^{-1})\lambda_{\max}(\widetilde{D}_s)\lambda_{\max}(R_s^TR_s) = \kappa(\widetilde{D}_s) \end{aligned}$$

which completes the proof.

Remark 1. For conforming FEM discretization of the second order scalar elliptic PDE it is not difficult to show that $\kappa(\widetilde{D}_s)$ is uniformly bounded with respect to jumps of an elementwise constant coefficient. Furthermore, $\widetilde{D}_s$ is block-diagonal with small-sized blocks and thus $\kappa(\widetilde{D}_s)$ is easily computable.

4 Auxiliary space multigrid method

Consider the exact block factorization of the sequence of auxiliary stiffness matrices $\widetilde{A}^k$, where the superscript $k = 0, 1, \ldots, \ell-1$ indicates the coarsening level:

$$\widetilde{A}^{(k)\,-1} = \widetilde{L}^{(k)\,T}\,\widetilde{D}^{(k)}\widetilde{L}^{(k)}, \quad A^{(k+1)} := Q^{(k)},$$

$$\widetilde{L}^{(k)} = \begin{bmatrix} I & \\ -\widetilde{A}_{21}^{(k)}\ \widetilde{A}_{11}^{(k)\,-1} & I \end{bmatrix}, \qquad \widetilde{D}^{(k)} = \begin{bmatrix} \widetilde{A}_{11}^{(k)\,-1} & \\ & Q^{(k)\,-1} \end{bmatrix}.$$

Let the algebraic multilevel iteration (AMLI)-cycle auxiliary space multigrid (ASMG) preconditioner $B^{(k)}$ be defined by (see Kraus et al. [2014]):

$$\begin{aligned} B^{(k)\,-1} &:= \overline{M}^{(k)\,-1} \\ &\quad +(I - M^{(k)\,-T}A^{(k)})\Pi^{(k)}\ \widetilde{L}^{(k)\,T}\,\overline{D}^{(k)}\widetilde{L}^{(k)}\Pi^{(k)\,T}(I - A^{(k)}M^{(k)\,-1}), \\ \overline{D}^{(k)} &:= \begin{bmatrix} \widetilde{A}_{11}^{(k)\,-1} & \\ & B_\nu^{(k+1)} \end{bmatrix}, \qquad B_\nu^{(\ell)} := A^{(\ell)\,-1}. \end{aligned}$$

In the nonlinear AMLI-cycle $B_\nu^{(k+1)} = B_\nu^{(k+1)}[\cdot]$ is a nonlinear mapping realized by ν iterations of a Krylov subspace method (e.g. the generalized conjugate gradient (GCG) method), thus employing the coarse level preconditioner $B^{(k+1)}$. In Kraus [2002] the convergence of the multiplicative nonlinear AMLI has been first analyzed, while Notay and Vassilevski [2008], Vassilevski [2008], Hu et al. [2013] have provided the multigrid framework along with a comparative analysis.

We want to stress the fact that the presented construction provides a framework for both linear and nonlinear AMLI cycle multigrid as well as classical multigrid methods.

5 Numerical Results

Subject to numerical testing is the scalar elliptic boundary-value problem

$$-\nabla \cdot (\boldsymbol{k}(\boldsymbol{x})\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}) \quad in \ \ \Omega, \tag{14a}$$

$$u = 0 \quad on \ \ \Gamma. \tag{14b}$$

Here Ω is a polygonal domain in $\mathbb{R}^2$, f is a given function in $L_2(\Omega)$ and

$$\boldsymbol{k}(\boldsymbol{x}) = \alpha(\boldsymbol{x})I = \alpha_e I. \tag{15}$$

Upon the entire boundary of the domain Dirichlet boundary conditions have been imposed as other boundary conditions would not qualitatively affect the numerical results.

Piecewise bilinear functions have been used in the process of discretization of (14) leading to the linear system of algebraic equations (1). A uniform mesh consisting of $N \times N$ elements (squares) is considered where $N = 2^{\ell+2}$, $\ell = 1, \ldots, 7$, and the covering is assumed to consist of subdomains composed of 8×8 elements that overlap with half of their width or height. The mesh hierarchy is such that the coarsest mesh corresponds to $\ell = 1$ and is composed of $2^{1+2} \times 2^{1+2} = 64$ elements whereas the finest mesh is obtained by performing $\ell - 1 = 1, \ldots, 6$ steps of uniform mesh refinement.

The vector of all zeros was chosen to be the right hand side $\mathbf{f}$ in (1) while the outer iteration was initialized with a random vector. Three representative coefficient configurations are considered (on the respective finest mesh):

[0] log-uniformly distributed coefficient $\alpha_e = 10^{p_{rand}}$ where α_e is constant on each element e and $p_{rand} \in (0, q]$;
[1] inclusions with coefficient $\alpha_\iota = 10^{p_{rand}}$ against a background as in [0] where α_ι is constant on every inclusion ι and $p_{rand} \in (0, q]$, see Fig. 2(a);
[2] stiff inclusions with coefficient $\alpha_\iota = 10^q$ against a background as in [0], see Fig. 2(b).

In Table 1 we compare the condition numbers

$$\kappa(\widetilde{D}_s) = \kappa(S\widetilde{D}S), \quad \kappa(B_{\mathrm{AS}}^{-1}D_s) = \kappa(SR\widetilde{S}^{-2}\widetilde{D}^{-1}\widetilde{S}^{-2}R^TS(SR\widetilde{D}R^TS)),$$

with that of the corresponding unscaled preconditioned system

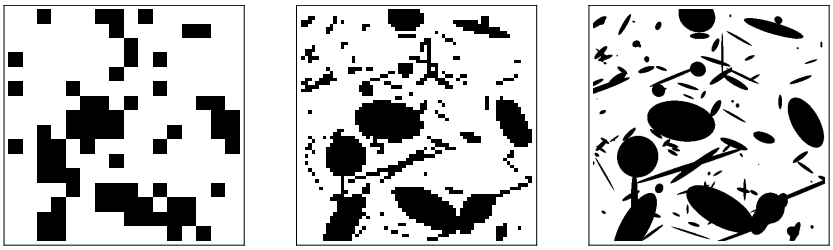
$$\kappa(R\widetilde{D}^{-1}R^T(R\widetilde{D}R^T))$$

for the coefficient distribution [0] on three different meshes with mesh size $h \in \{1/16, 1/32, 1/64\}$ and varying contrast q. The obtained numerical results are in accordance with Lemma 2; They further show that the scaled one-level additive Schwarz method yields a uniform preconditioner whereas its unscaled analog suffers from high-contrast coefficients.

Next, the numerical performance of the nonlinear (AMLI)-cycle ASMG method (V-cycle and W-cycle) utilizing the preconditioner B_{AS} is tested for:

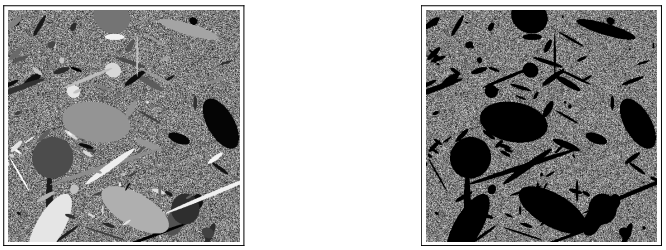
(P1) Problem (14) with coefficient distributions [1] and variants [I] and [II] of $\mathit{\Pi}_{\widetilde{D}}$. Variant [II] is realized by 10 inner PCG iterations with the scaled one-level AS preconditioner.
(P2) Same as Problem (P1) but for coefficient distribution 2.

A comparison between variant [I] and variant [II] of the ℓ-level V-cycle and W-cycle is presented in Tables 2–3. Pre- and post-smoothing is performed by



(a) 16×16 mesh (b) 64×64 mesh (c) 512×512 mesh

Fig. 1 Inclusions resolved on different fine scales (meshes)



(a) Coefficient for Problem (P1) on 512×512 mesh (b) Coefficient for Problem (P2) on 512×512 mesh

Fig. 2 Random and stiff inclusions against random background $\alpha_e = 10^{p_{rand}}$

Table 1 Condition numbers of AS-preconditioned systems versus $\kappa(\widetilde{D}_s)$

	unscaled AS method			scaled AS method			$\kappa(\widetilde{D}_s)$		
q \ h	1/16	1/32	1/64	1/16	1/32	1/64	1/16	1/32	1/64
1	9.76×10^1	9.47×10^1	9.35×10^1	1.25	1.26	1.26	4.73	4.73	4.73
2	2.25×10^2	3.69×10^2	5.89×10^2	1.28	1.27	1.29	4.73	4.73	4.73
3	6.93×10^2	2.42×10^3	3.70×10^3	1.29	1.32	1.33	4.73	4.73	4.73
4	1.93×10^4	1.97×10^4	3.77×10^4	1.33	1.33	1.33	4.73	4.73	4.73
5	1.78×10^5	1.87×10^5	2.16×10^5	1.32	1.33	1.33	4.73	4.73	4.73
6	3.07×10^5	1.34×10^6	2.15×10^6	1.33	1.33	1.33	4.73	4.73	4.73

one symmetric point Gauss-Seidel iteration on each level except the coarsest one where all linear systems are solved directly.

Table 2 Number of iterations for residual reduction by 10^6

Problem (P1)																								
	Nonlinear AMLI V-cycle												Nonlinear AMLI W-cycle											
	[I]						[II]						[I]						[II]					
q \ ℓ	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
1	4	5	6	6	7	8	5	5	6	6	7	8	4	5	5	5	5	5	5	5	5	5	5	5
2	5	5	6	6	7	8	5	5	6	6	7	8	5	5	5	5	5	5	5	5	5	5	5	5
3	5	6	6	7	7	8	5	6	6	7	7	8	5	6	6	6	6	6	5	5	5	5	5	5
4	5	6	7	8	8	9	5	6	7	8	8	8	5	6	6	6	6	6	5	6	6	6	6	6
5	5	7	7	8	9	9	5	6	7	8	8	8	5	6	6	6	7	7	5	6	6	6	6	6
6	5	7	8	9	13	15	5	7	8	8	8	9	5	6	6	7	9	10	5	6	6	6	6	6

Table 3 Number of iterations for residual reduction by 10^6

Problem (P2)																								
	Nonlinear AMLI V-cycle												Nonlinear AMLI W-cycle											
	[I]						[II]						[I]						[II]					
q \ ℓ	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7	2	3	4	5	6	7
1	5	5	6	6	7	8	5	5	6	6	7	8	5	5	5	5	5	5	5	5	5	5	5	5
2	5	5	6	6	7	8	5	5	6	6	7	8	5	5	5	5	5	5	5	5	5	5	5	5
3	5	5	6	6	7	8	5	5	6	6	7	8	5	5	5	6	5	6	5	5	5	5	5	5
4	5	6	6	7	7	8	5	5	6	7	8	8	5	5	6	6	6	6	5	6	5	5	5	6
5	5	6	7	7	9	9	5	6	7	7	8	8	5	6	6	6	6	6	5	6	6	6	6	6
6	5	6	8	8	12	13	5	6	7	8	9	9	5	6	6	6	8	9	5	6	6	6	6	6

The obtained results demonstrate that the choice of $\widetilde{D}$ and consequently of the surjective mapping $\Pi_{\widetilde{D}}$ affect the performance of the nonlinear AMLI-cycle ASMG method crucially. As for variant [I] the number of ASMG iterations required to achieve the prescribed accuracy increases with the contrast, variant [II] shows full robustness.

References

O. Axelsson and P. Vassilevski. Algebraic multilevel preconditioning methods I. *Numer. Math.*, 56(2-3):157–177, 1989.

O. Axelsson and P. Vassilevski. Algebraic multilevel preconditioning methods II. *SIAM J. Numer. Anal.*, 27(6):1569–1590, 1990.

O. Axelsson and P. Vassilevski. Variable-step multilevel preconditioning methods, I: Self-adjoint and positive definite elliptic problems. *Numer. Linear Algebra Appl.*, 1:75–101, 1994.

C.R. Dohrmann. A preconditioner for substructuring based on constrained energy minimization. *SIAM J. Sci. Comput.*, 25(1):246–258, 2003.

Y. Efendiev, J. Galvis, R. Lazarov, and J. Willems. Robust domain decomposition preconditioners for abstract symmetric positive definite bilinear forms. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(05): 1175–1199, 2012.

C. Farhat and F.X. Roux. A method of finite element tearing and interconnecting and its parallel solution algorithm. *Int. J. Numer. Methods Engrg.*, 32:1205–1227, 1991.

C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen. FETI-DP: A dual-primal unified FETI method. I. A faster alternative to the two-level FETI method. *Int. J. Numer. Methods Engrg.*, 50(7):1523–1544, 2001.

J. Galvis and Y. Efendiev. Domain decomposition preconditioners for multiscale flows in high-contrast media. *Multiscale Model. Simul.*, 8(4):1461–1483, 2010.

I.G. Graham, P.O. Lechner, and R. Scheichl. Domain decomposition for multiscale PDEs. *Numer. Math.*, 106(4):489–626, 2007.

W. Hackbusch. *Multi-Grid Methods and Applications*. Springer, Berlin Heidelberg, 2003.

X. Hu, P. Vassilevski, and J. Xu. Comparative convergence analysis of nonlinear AMLI-cycle multigrid. *SIAM J. Numer. Anal.*, 51(2):1349–1369, 2013.

A. Klawonn, O. Widlund, and M. Dryja. Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients. *SIAM J. Numer. Anal.*, 40(1):159–179, 2002.

J. Kraus. An algebraic preconditioning method for M-matrices: linear versus non-linear multilevel iteration. *Numer. Linear Algebra Appl.*, 9:599–618, 2002.

J. Kraus. Algebraic multilevel preconditioning of finite element matrices using local Schur complements. *Numer. Linear Algebra Appl.*, 13:49–70, 2006.

J. Kraus. Additive Schur complement approximation and application to multilevel preconditioning. *SIAM J. Sci. Comput.*, 34:A2872–A2895, 2012.

J. Kraus, P. Vassilevski, and L. Zikatanov. Polynomial of best uniform approximation to 1/x and smoothing for two-level methods. *Comput. Meth. Appl. Math.*, 12:448–468, 2012.

J. Kraus, M. Lymbery, and S. Margenov. Auxiliary space multigrid method based on additive Schur complement approximation. *Numer. Linear Algebra Appl.*, 2014. Online (wileyonlinelibrary.com). DOI: 10.1002/nla.1959.

Y. Kuznetsov. Algebraic multigrid domain decomposition methods. *Sov. J. Numer. Anal. Math. Modelling.*, 4(5):351–379, 1989.

J. Mandel. Balancing domain decomposition. *Comm. Numer. Methods Engrg.*, 9(3):233–241, 1993.

J. Mandel and C. R. Dohrmann. Convergence of a balancing domain decomposition by constraints and energy minimization. *Numer. Linear Algebra Appl.*, 10(7):639–659, 2003.

J. Mandel and B. Sousedík. Adaptive selection of face coarse degrees of freedom in the BDDC and FETI-DP iterative substructuring methods. *Comput. Methods Appl. Mech. Engrg.*, 196(8):1389–1399, 2007.

J. Mandel, C. R. Dohrmann, and R. Tezaur. An algebraic theory for primal and dual substructuring methods by constraints. *Appl. Numer. Math.*, 54 (2):167–193, 2005.

T.P.A. Mathew. *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations*. Springer, Berlin Heidelberg, 2008.

Y. Notay and P. Vassilevski. Recursive Krylov-based multigrid cycles. *Numer. Linear Algebra Appl.*, 15:473–487, 2008.

C. Pechstein and R. Scheichl. Analysis of FETI methods for multiscale PDEs. *Numer. Math.*, 111(2):293–333, 2008.

M. Sarkis. *Schwarz Preconditioners for Elliptic Problems with Discontinuous Coefficients Using Conforming and Non-Conforming Elements*. PhD thesis, Courant Institute, New York University, 1994.

N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numer. Math.*, 126(4):741–770, 2014.

A. Toselli and O. Widlund. *Domain Decomposition Methods–Algorithms and Theory*. Springer, Berlin Heidelberg, 2005.

U. Trottenberg, C.W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press Inc., San Diego, CA, 2001.

P. Vassilevski. *Multilevel Block Factorization Preconditioners: Matrix-based Analysis and Algorithms for Solving Finite Element Equations*. Springer, New York, 2008.

J. Xu. The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids. *Computing*, 56:215–235, 1996.

PRECONDITIONING HETEROGENEOUS $H(\mathrm{div})$ PROBLEMS BY ADDITIVE SCHUR COMPLEMENT APPROXIMATION AND APPLICATIONS

SIAM J. SCI. COMPUT.
Vol. 38, No. 2, pp. A875–A898

PRECONDITIONING HETEROGENEOUS $H(\mathrm{div})$ PROBLEMS BY ADDITIVE SCHUR COMPLEMENT APPROXIMATION AND APPLICATIONS*

JOHANNES KRAUS†, RAYTCHO LAZAROV‡, MARIA LYMBERY†, SVETOZAR MARGENOV§, AND LUDMIL ZIKATANOV¶

Abstract. In this paper we propose and analyze a preconditioner for a system arising from a mixed finite element approximation of second-order elliptic problems describing processes in highly heterogeneous media. Our approach uses the technique of multilevel methods (see, e.g., [P. Vassilevski, *Multilevel Block Factorization Preconditioners: Matrix-Based Analysis and Algorithms for Solving Finite Element Equations*, Springer, New York, 2008]) and the recently proposed preconditioner based on additive Schur complement approximation by J. Kraus [*SIAM J. Sci. Comput.*, 34 (2012), pp. A2872–A2895]. The main results are the design, study, and numerical justification of iterative algorithms for these problems that are robust with respect to the contrast of the media, defined as the ratio between the maximum and minimum values of the coefficient of the problem. Numerical tests provide experimental evidence for the high quality of the preconditioner and its desired robustness with respect to the material contrast. Such results for several representative cases are presented, one of which is related to the SPE10 (Society of Petroleum Engineers) benchmark problem.

Key words. mixed finite elements, high contrast media, robust preconditioners for weighted $\boldsymbol{H}(\mathrm{div})$-norm, discrete Poincaré inequality

AMS subject classifications. 65F10, 65N20, 65N30

DOI. 10.1137/140974092

1. Introduction.

1.1. Model problem definition. Flows in porous media appear in many industrial, scientific, engineering, and environmental applications and are a subject of significant scientific interest. Their analogous mathematical formulations are also used in the modelling of other physical processes such as heat and mass transfer, diffusion of passive chemicals, and electromagnetics. This leads to the following system of partial differential equations (PDEs) of first order for the unknown scalar function $p(x)$ and the vector function $\boldsymbol{u}(x)$:

$$\boldsymbol{u} + K(x)\nabla p = 0 \qquad \text{in } \Omega, \tag{1.1a}$$

*Submitted to the journal's Methods and Algorithms for Scientific Computing section June 23, 2014; accepted for publication (in revised form) December 18, 2015; published electronically March 17, 2016. This work was partially supported by Bulgarian NSF grant DCVP 02/1, FP7 grant AComIn, and Austrian NSF grant P22989.

http://www.siam.org/journals/sisc/38-2/97409.html

†Faculty of Mathematics, University of Duisburg-Essen, Thea-Leymann-Str. 9, 45127 Essen, Germany (johannes.kraus@uni-due.de, maria.lymbery@uni-due.de).

‡Department of Mathematics, Texas A & M University, College Station, TX 77843, and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Bl. 8, 1113 Sofia, Bulgaria (lazarov@math.tamu.edu). This author's work was supported in part by U.S. NSF grant DMS-1016525.

§Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Block 2, 1113 Sofia, Bulgaria (margenov@parallel.bas.bg).

¶Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Bl. 8, 1113 Sofia, Bulgaria (ltz@math.psu.edu). This author's work was supported in part by U.S. NSF grants DMS-1217142 and DMS-1016525.

$$\operatorname{div} \boldsymbol{u} = f \qquad \text{in } \Omega, \tag{1.1b}$$

$$p = g \qquad \text{on } \Gamma_D, \tag{1.1c}$$

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0 \qquad \text{on } \Gamma_N, \tag{1.1d}$$

where Ω is a polygonal domain in $\mathbb{R}^d$, $d = 2, 3$. In the terminology of flows in porous media the unknown scalar function $p(x)$ and the vector function $\boldsymbol{u}$ are called pressure and velocity, respectively, while $K(x) : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}$, the permeability tensor, is a symmetric and positive definite (SPD) matrix for almost all $x \in \Omega$. Equation (1.1a) is Darcy's law, while (1.1b) expresses conservation of mass.

Our study is focused on the case $K(x) = k(x)I$, where I is the identity matrix in $\mathbb{R}^d$ and $k(x)$ is a scalar function. The given forcing term f is a function in $L^2(\Omega)$. The boundary $\partial\Omega$ is split into two nonoverlapping parts Γ_D and Γ_N, and in the case of a pure Neumann problem, i.e., $\Gamma_N = \partial\Omega$, we assume that f satisfies the compatibility condition $\int_\Omega f dx = 0$. In such a case the solution is determined uniquely by taking $\int_\Omega p\, dx = 0$.

To simplify the presentation, Γ_D is assumed to be a nonempty set with strictly positive measure which is also closed with respect to $\partial\Omega$, and $g(x) \equiv 0$ on Γ_D, so the above system of equations has a unique solution $p \in H^1_D(\Omega) = \{q \in H^1(\Omega) : q = 0 \text{ on } \Gamma_D\}$, and $\boldsymbol{u}$ is defined by (1.1a).

Specifically, applications to flows in *highly heterogeneous* porous media of *high contrast* are studied. The coefficient $k(x)$ in this context represents media with multiscale features, involving many small size inclusions and/or long connected subdomains (channels), where $k(x)$ has large values (see Figure 2). A computer generated permeability coefficient $K(x)$ exhibiting such features has been used as a benchmark in petroleum engineering related simulations; cf. the SPE10 Comparative Solution Project [28] of the Society of Petroleum Engineers. Figure 3 shows the permeability field for two-dimensional (2D) slices of such media. An important characteristic is the contrast κ, defined by (2.2) as a ratio between the maximum and minimum values of $k(x)$.

In this paper we consider approximations of the problem (1.1) by the mixed finite element method (FEM) on a mesh that resolves the finest scale of the permeability. This leads to a very large indefinite symmetric system of algebraic equations. Developing, studying, and testing an optimal preconditioner with respect to the contrast κ and the mesh size h for this algebraic problem is the objective of this paper. Our considerations and numerical experiments show that the proposed preconditioner is optimal so that the number of iterations depends on neither the contrast nor the mesh size. This is the main achievement of this paper.

For the vector variable $\boldsymbol{u}$ we use the lowest-order Raviart–Thomas $\boldsymbol{H}(\mathrm{div})$-conforming finite elements (FEs). The algebraic system of linear equations for the unknown degrees of freedom associated with $\boldsymbol{u}$ and p can be written in the following block form (for more details, see subsection 4.1):

$$\begin{bmatrix} M_\alpha & -B_{\mathrm{div}}^T \\ -B_{\mathrm{div}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f} \end{bmatrix}, \tag{1.2}$$

where the matrix M_α is generated by the inner product $(\alpha\boldsymbol{u}, \boldsymbol{v})$, while B_{div} is generated by the form $(\nabla \cdot \boldsymbol{u}, q)$. It is well known (see, e.g., [1]) that the matrix mapping properties of this system are the same as those of

$$\mathcal{B}_h := \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix}, \tag{1.3}$$

where the matrix A corresponds to the weighted $\boldsymbol{H}(\mathrm{div})$-inner product $(\alpha\boldsymbol{u},\boldsymbol{v}) + (\mathrm{div}\,\boldsymbol{u},\mathrm{div}\,\boldsymbol{v})$. Therefore, for an optimal MinRes iteration, the construction of an efficient preconditioner for the bilinear form $(\alpha\boldsymbol{u},\boldsymbol{v}) + (\mathrm{div}\,\boldsymbol{u},\mathrm{div}\,\boldsymbol{v})$, which is robust with respect to both the contrast and the mesh size, is essential. In this paper we focus on the construction and study of a suitable preconditioner of A in (1.3).

1.2. Overview of existing results. The standard elliptic theory ensures the existence of a unique solution $p \in H^1_D(\Omega)$. However, since $K(x)$ is a piecewise smooth matrix and may have very large jumps, the solution p has low regularity. For example, the case $H^{1+s}(\Omega)$, where $s > 0$, could depend on the contrast κ in a subtle and unfavorable manner. This must be considered when proving the stability of discrete methods with a constant independent of κ. As a consequence, any solution or preconditioning technique, such as multigrid or domain decomposition, that is analyzed by using the solutions' regularity cannot produce theoretical results independent of the contrast.

Note that the block A corresponds to the FE approximation of the weighted $\boldsymbol{H}(\mathrm{div})$-norm generated by the weighted inner product $(K^{-1}\boldsymbol{u},\boldsymbol{v}) + (\mathrm{div}\,\boldsymbol{u},\mathrm{div}\,\boldsymbol{v})$ with $\boldsymbol{H}(\mathrm{div})$-conforming FEs. Thus, one might expect that the existing preconditioners of $\boldsymbol{H}(\mathrm{div})$-norms would be appropriate to begin with. Various scenarios for the properties of $K(x)$ are possible, such as the following.

Constant K and/or smooth variable $K(x)$. The case of $K(x)$ being an SPD matrix over Ω has been considered by Arnold, Falk, and Winther in [1, 2], and the corresponding preconditioner (based on multigrid and/or domain decomposition) was shown to be optimal with respect to the mesh size in two space dimensions. The analysis of the preconditioner relies on the approximation properties of the Raviart–Thomas projection and requires full regularity of the solution. Further, based on early work by Vassilevski and Wang [31], Börm and Hiptmair [3] and later Hiptmair and Xu [12] developed a preconditioner for the $\boldsymbol{H}(\mathrm{div})$-norm that is optimal with respect to the mesh size. This work does not consider the variable $K(x)$ and a weighted norm. Nevertheless, its analysis can be potentially extended to this case. However, the theoretical justification of this preconditioner uses in a fundamental way the approximation properties of FE projections (Raviart–Thomas in two dimensions and Nedelec in three dimensions) that require regularity of the vector field $\boldsymbol{u}$; see, e.g., [3, error bounds (2.3)–(2.5)], which may depend, in an unfavorable way, on the contrast κ. In general, such regularity is not available for problems in highly heterogeneous media with large contrast. Additionally, the main ingredient of the preconditioner in [12], stable regular decompositions, requires an extension of these results to the case of weighted norms. To the best of our knowledge, such results are still out of reach for highly heterogeneous coefficients, and the analogues of Lemmas 3.4 and 3.8 of [12], crucial for constructing preconditioners, are not yet available.

Anisotropic coefficient matrix $K(x)$. Often the coefficient matrix represents the properties of anisotropic media or heterogeneous media with highly anisotropic inclusions. Solving linear systems resulting from FE approximation of such problems is not fully mastered. Moreover, a theoretical justification of iterative solvers that are robust with respect to anisotropy is a very difficult task. Some early work in multigrid for 2D problems (see, e.g., [3, 4]) uses grids aligned with the anisotropy. In [4, conditions (1.2)–(1.4)], a certain coefficient regularity is required, while in [3] it is assumed that the grid is aligned with the anisotropy of the coefficient $K(x)$ and the line-relaxation in the strongly coupled direction is combined with semicoarsening in the other directions only.

Variable $K(x)$ with high contrast κ. The design and analysis of condition numbers for the system (1.2) preconditioned with block-diagonal preconditioners were carried out by Powell and Silvester [26] and Powell [25]. One class of preconditioners proposed in these two papers, and relevant to our constructions here, is of block-diagonal form with a weighted $\boldsymbol{H}(\mathrm{div})$ operator as one diagonal block, and a lumped (weighted) mass matrix as the other. In practice, one can use approximate solutions of the $\boldsymbol{H}(\mathrm{div})$ problem by V-cycle multigrid, e.g., as proposed by Arnold, Falk, and Winther [1]. In the case of a smooth coefficient matrix $K(x)$, such an approach produces an optimal preconditioner for the mixed system in the isotropic case, independent of the contrast κ, as seen in [26, section 2.3.1, Tables 2.3–2.7] and [25, Tables 4 and 5]. In other cases, matrix-valued anisotropic coefficients and highly heterogeneous coefficients not aligned with the grid, the resulting preconditioners are robust with respect to the coefficient variation whenever the approximation to the first diagonal block demonstrates such robustness. It is still an open theoretical question, however, whether any of the known multigrid algorithms for the weighted $\boldsymbol{H}(\mathrm{div})$ problem converge uniformly with respect to both h and the coefficient variation for all cases, and particularly in cases of matrix-valued and anisotropic coefficients with discontinuities not aligned with the coarsest mesh. This statement applies to *both* algebraic and geometric multigrid methods.

Furthermore, the framework for practical preconditioning of Powell and Silvester [25] and Powell [26] can be combined with the Schur complement preconditioning of the $\boldsymbol{H}(\mathrm{div})$ block as proposed here. While a rigorous mathematical justification of such a result is beyond our present consideration, the numerical tests presented later and the analysis in [25] and [26] show that such a combined approach has the potential to be successful and practical.

Highly heterogeneous discontinuous $K(x)$. In the existing literature there are a number of techniques for preconditioning algebraic problems with heterogeneous coefficients of high contrast or large jumps. Among the most popular are domain decomposition, e.g., [13, 8] for the standard Galerkin FEM, and multilevel methods for the hybridized mixed system, e.g., [14]. The main result in [13] concerns a domain decomposition FETI-type preconditioner which is optimal with respect to the contrast in the case when jumps of $K(x)$ are aligned with the coarse mesh (or the splitting of the domain into subdomains). Similarly, the preconditioners presented in [17], based on algebraic multilevel iteration (AMLI) methods, are theoretically proven to be robust with respect to contrast and anisotropy in the case when jumps of the coefficients $K(x)$ are aligned with the initial coarse mesh. The results shown in [17, Table 7.10, p. 163] demonstrate numerically that for highly heterogeneous media with jumps in the permeability $K(x)$ aligned with the fine mesh only, the AMLI preconditioner is not robust with respect to the contrast. In a recent work [33], J. Willems has developed a robust (with respect to the problem parameters) nonlinear multilevel method for solving general SPD systems. A crucial role in the construction of the nested spaces and the smoother is played by local generalized eigenproblems (in the manner of [10]) and four assumptions. It is not clear when these are verifiable for the case of the form Λ_α.

We share the opinion expressed in [25, section 6] that the existence of theoretically proven optimal preconditioners of the $\|\cdot\|_{\Lambda_\alpha}$-norm (defined by the weighted $\boldsymbol{H}(\mathrm{div})$-product (2.5)) in the case of a general SPD tensor $K(x)$ is an open question. Moreover, a tensor $K(x)$ with arbitrary heterogeneities and/or high anisotropy represents a genuine challenge for both theory and computational practice. Our paper is a step

in this direction for the case of a highly heterogeneous permeability tensor $K(x)$ that satisfies the condition (2.7).

1.3. Contributions of the study. The main result and novelty of this paper are the design, theoretical discussion, and experimental study of a preconditioner for a matrix corresponding to the weighted norm $\|\cdot\|_{\Lambda_\alpha}$ in the space $\boldsymbol{H}(\mathrm{div})$ (defined by (2.5)) which gives an iterative method for mixed FE systems converging independently of the contrast κ. Such a construction is based on ideas from [16].

A crucial role is played by the well-known *inf-sup* condition. In this paper we consider the case of a permeability tensor satisfying condition (2.7). The inf-sup condition for this case immediately follows from the well-studied situation where $K(x) = I$. However, in order to emphasize the dependence of the inf-sup constant on the global properties of the differential operator and the means by which it can be extended to a more general form of $K(x)$, an outline of the proof is presented. Furthermore, we present a short discussion in which $K(x) = k(x)I$ for highly heterogeneous values of $k(x)$ and $0 < k(x) < \infty$. Theorem 3.1 establishes an inf-sup condition and boundedness of the corresponding bilinear form in a discrete setting. We emphasize that under (2.7) the constant in the inf-sup condition and the boundedness of the corresponding form do not depend on the contrast of the media.

In section 4, which is central to the paper, a new preconditioning method for the FE systems is described. First, a block-diagonal preconditioner for the operator form of the mixed FEM is defined. Furthermore, subsection 4.1 presents the FE problem in matrix form. The key issue when designing a contrast-independent Krylov solver is the construction of a robust preconditioner for the weighted $\boldsymbol{H}(\mathrm{div})$-norm. An important aspect in its analysis is the norm of a suitable projection characterizing this preconditioner. Currently a general theoretical proof of the independence of this norm with respect to the contrast does not exist. However, we have presented numerical evidence (in section 5, Table 1) that this quantity is bounded independently of the contrast and the mesh size. Such a result is highly desirable and of great practical value. Moreover, all of our presented numerical experiments indirectly show such robustness. Subsection 4.2.1 introduces an auxiliary space multigrid (ASMG) method. Two variants of the algorithm, differing only in their relaxation procedure, are described in subsection 4.2.2. The work needed to compute the action of the preconditioner is proportional to the total number of nonzeros in the coarse-grid matrices (the so-called operator complexity of the preconditioner), and this is discussed in some detail in section 4.3. Finally, section 5 gives numerical results for three different examples of porous media in two dimensions in order to test the robustness of the preconditioner with respect to media contrast, and its optimality with respect to mesh size. All numerical results confirm these claims.

2. Problem formulation.

2.1. Notation and preliminaries. For functions defined on Ω we use the standard notation for Sobolev spaces; namely, $H^s(\Omega)$ for $s \geq 0$ being an integer is the space of functions having their generalized derivatives up to order s square-integrable on Ω. We denote by $(\cdot,\cdot)$ the L^2 and $[L^2]^d$ inner products. The standard norms on H^s are denoted by $\|\cdot\|_s$. For $s = 0$ we also write $\|\cdot\|$ without a subscript. When the norm is weighted with a matrix-valued function $\omega(x)$, with $\omega(x)$ being SPD for almost all $x \in \Omega$, we use the notation

$$\|\boldsymbol{v}\|_{0,\omega} := \|\omega^{1/2}\boldsymbol{v}\|, \qquad |\phi|_{1,\omega} := \|\nabla\phi\|_{0,\omega} = \|\omega^{1/2}\nabla\phi\|. \tag{2.1}$$

Occasionally, when considering only a subset of Ω, e.g., $T \subset \Omega$, this is indicated in the notation for norms and seminorms, i.e., $\|\cdot\|_{s,T}$, $\|\cdot\|_{s,\omega,T}$, $|\cdot|_{s,T}$, and $|\cdot|_{s,\omega,T}$.

To put our work into perspective, in the following, we consider $K \in \mathbb{R}^{d\times d}$ to be a symmetric matrix, the norm $\|K\|_{\ell^2}$ denoting, as usual, the spectral radius of K. We define the number

$$\kappa = \max_{x\in\Omega}(\|K(x)\|_{\ell^2}\|K^{-1}(x)\|_{\ell^2}) \quad \text{with} \quad \|K(x)\|_{\ell^2} = \sup_{\xi\in\mathbb{R}^d}(K(x)\xi\cdot\xi)/(\xi\cdot\xi) \tag{2.2}$$

to be the contrast of the media. Obviously, $\kappa = \max_{x\in\Omega} K(x)/\min_{x\in\Omega} K(x)$ for a scalar permeability coefficient $K(x)$. In many applications κ is much larger than 1, often up to 10 orders of magnitude, and the larger κ is, the more difficult it becomes to devise an efficient preconditioner.

The Hilbert space $\boldsymbol{H}(\mathrm{div})$ consists of square-integrable vector-fields on Ω with square-integrable divergence. The inner product in $\boldsymbol{H}(\mathrm{div})$ is given by

$$\Lambda(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u},\boldsymbol{v}) + (\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{v}), \quad \text{and, consequently,} \quad \|\boldsymbol{v}\|^2_{\boldsymbol{H}(\mathrm{div})} := \Lambda(\boldsymbol{v},\boldsymbol{v}). \tag{2.3}$$

Together with the Sobolev space $H^1_D(\Omega)$, we use the following notation of $\boldsymbol{H}_N(\mathrm{div})$:

$$\boldsymbol{H}_N(\mathrm{div}) := \boldsymbol{H}_N(\mathrm{div};\Omega) = \{\boldsymbol{v}\in\boldsymbol{H}(\mathrm{div};\Omega): \quad \boldsymbol{v}(x)\cdot\boldsymbol{n} = 0 \ \text{ on } \ \Gamma_N\}. \tag{2.4}$$

For $\phi \in H^1_D(\Omega)$ the seminorms $|\phi|_1 = \|\nabla\phi\|$ and $|\phi|_{1,\omega} = \|\omega^{1/2}\nabla\phi\|$ are, in fact, norms on $H^1_D(\Omega)$, and we denote these by $\|\phi\|_1$ and $\|\phi\|_{1,\omega}$.

Together with (2.3), the following weighted inner product in the space $\boldsymbol{H}(\mathrm{div})$ plays a fundamental role in our analysis:

$$\Lambda_\alpha(\boldsymbol{u},\boldsymbol{v}) = (\alpha\ \boldsymbol{u},\boldsymbol{v}) + (\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{v}), \quad \alpha(x) = K^{-1}(x). \tag{2.5}$$

It defines the norm

$$\|\boldsymbol{v}\|^2_{\Lambda_\alpha} = \Lambda_\alpha(\boldsymbol{v},\boldsymbol{v}) = \|\boldsymbol{v}\|^2_{0,\alpha} + \|\operatorname{div}\boldsymbol{v}\|^2. \tag{2.6}$$

A weighted bilinear form of the type $\Lambda_{\alpha,\beta}(\boldsymbol{u},\boldsymbol{v}) = \alpha(\boldsymbol{u},\boldsymbol{v}) + \beta(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{v})$ with constants $\alpha > 0$ and $\beta > 0$ can be preconditioned by geometric multigrid methods, uniformly with respect to the parameters α and β, as shown by Arnold, Falk, and Winther in [2]. The important difference between our bilinear form Λ_α and the form considered in [2] is that in our scheme α is a highly heterogeneous function with high contrast. This makes the proof of an inf-sup condition with the weighted $\boldsymbol{H}(\mathrm{div})$-norm more delicate, and the construction of a robust preconditioner a challenging task; see Remark 2.1.

Note that for all $\xi\in\mathbb{R}^d$, $x\in\Omega$, and $\|K\|_{\ell^2} = \sup_{\xi\in\mathbb{R}^d}(K\xi\cdot\xi)/(\xi\cdot\xi)$ we have with $K := K(x)$

$$(K\xi\cdot\xi) \geq (\xi\cdot\xi)\ \inf_{\theta\in\mathbb{R}^d}\frac{(K\theta\cdot\theta)}{(\theta\cdot\theta)} = (\xi\cdot\xi)\ \inf_{\theta\in\mathbb{R}^d}\frac{(\theta\cdot\theta)}{(K^{-1}\theta\cdot\theta)} = \frac{(\xi\cdot\xi)}{\|K^{-1}\|_{\ell^2}}.$$

Throughout the paper the following inequality is assumed:

$$1 \leq \min_{x\in\Omega}\|K(x)\|_{\ell^2}, \quad \text{which implies} \quad (\xi\cdot\xi) \leq (K(x)\xi\cdot\xi), \quad \xi\in\mathbb{R}^d. \tag{2.7}$$

As seen from the considerations above, such an assumption is fulfilled if we scale the coefficient $K(x) \leftarrow K(x)\max_{x\in\Omega}\|K^{-1}(x)\|_{\ell^2}$. This rescaling does not change the

value of κ; however, it changes the right-hand side $f(x) \leftarrow f(x) \max_{x\in\Omega} \|K^{-1}(x)\|_{\ell^2}$, and in general the stability of the solution cannot be established uniformly with respect to the contrast. Nevertheless, homogeneous equations ($f(x) \equiv 0$) represent an important class of applications. Such a scaling can also be justified when the permeability is homogeneous near the Dirichlet boundary. Another possible case is when $f(x) = 0$ in areas with very high permeability. The numerical examples of Powell and Silvester presented in [26, Tables 2.9 and 2.10] for $K(x) = k(x)I$, with $k(x)$ a scalar function and I the identity matrix in $\mathbb{R}^2$, clearly show this. Numerical experiments in section 5 consider homogeneous equations which are relevant to numerical reservoir simulations.

The case $0 < k(x) < \infty$, which could be used to model flow models in perforated domains, appears to be more complicated and less studied. For such problems more advanced techniques involving weighted L^2-norms and the weighted Poincaré inequality are needed; see Remark 2.1. Such an inequality can be established under certain restrictions on the arrangement of the jumps and the permeability distribution; see, e.g., [22, 23, 27]. Even more difficult is the case of tensor permeability $0 < \|K(x)||_{\ell^2} < \infty$, which also includes models of flows in anisotropic highly heterogeneous media. These cases represent open problems with a wide range of applications and are left for further consideration and future studies.

2.2. Weak formulations of the elliptic problem. To present the dual mixed weak form we require the following notation: $\boldsymbol{V} \equiv \boldsymbol{H}_N(\mathrm{div};\Omega)$ and $W \equiv L^2(\Omega)$.

We multiply the first equation by $\alpha(x) = K^{-1}(x)$ and a test function $\boldsymbol{v} \in \boldsymbol{V}$, integrate over Ω, and perform integration by parts to obtain

$$(\alpha(x)\boldsymbol{u}, \boldsymbol{v}) - (p, \mathrm{div}\, \boldsymbol{v}) = 0. \tag{2.8}$$

Next, multiplying the second equation by a test function $q \in W$ and integrating over Ω gives

$$(\mathrm{div}\, \boldsymbol{u}, q) = (f, q). \tag{2.9}$$

Then the weak form of problem (1.1) is as follows: find $\boldsymbol{u} \in \boldsymbol{V}$ and $p \in W$ such that

$$\mathcal{A}(\boldsymbol{u}, p; \boldsymbol{v}, q) = -(f, q) \quad \text{for all} \quad (\boldsymbol{v}, q) \in \boldsymbol{V} \times W, \tag{2.10}$$

where the bilinear form $\mathcal{A}(\boldsymbol{u}, p; \boldsymbol{v}, q) : (\boldsymbol{V}, W) \times (\boldsymbol{V}, W) \to \mathbb{R}$ is defined as

$$\mathcal{A}(\boldsymbol{u}, p; \boldsymbol{v}, q) := (\alpha\boldsymbol{u}, \boldsymbol{v}) - (p, \mathrm{div}\, \boldsymbol{v}) - (\mathrm{div}\, \boldsymbol{u}, q). \tag{2.11}$$

2.3. Stability of the weak formulations. Consider the stability of the discrete problem (2.10). From the Poincaré inequality we have that there exists a constant $C_P > 0$ such that

$$\|q\|^2 \le C_P \|\nabla q\|^2 \quad \text{for all} \quad q \in H^1_D(\Omega). \tag{2.12}$$

The constant C_P depends only on the geometry of the domain Ω and the splitting of $\partial\Omega$ into Γ_D and Γ_N. Moreover, due to (2.7), for the coefficient $K(x)$ we also have the inequality

$$\|q\|^2 \le C_P \|\nabla q\|^2 \le C_P \|\nabla q\|^2_{0,K}.$$

To show the stability of the weak formulation we need a continuity and an inf-sup condition (see, e.g., [6, 9]) for the bilinear form $\mathcal{A}(\boldsymbol{u}, p; \boldsymbol{v}, q)$ on the spaces $\boldsymbol{V}$ and $L^2(\Omega)$ equipped with the weighted norm $(\Lambda_\alpha(\boldsymbol{v}, \boldsymbol{v}))^{\frac{1}{2}}$ and the standard L^2-norm $\|p\|$, respectively.

LEMMA 2.1. *Let* $W = L^2(\Omega)$, $\boldsymbol{V} = \boldsymbol{H}_N(\mathrm{div})$, *and* $\|\boldsymbol{v}\|_{\Lambda_\alpha} := (\Lambda_\alpha(\boldsymbol{v},\boldsymbol{v}))^{\frac{1}{2}}$. *Assume also that the permeability coefficient* $K(x)$ *satisfies inequality* (2.7). *Then the following inequalities hold:*

(1) *For all* $\boldsymbol{u},\boldsymbol{v} \in \boldsymbol{V}$ *and for all* $p, q \in W$,

$$\mathcal{A}(\boldsymbol{u},p;\boldsymbol{v},q) \le (\|\boldsymbol{u}\|^2_{\Lambda_\alpha} + \|p\|^2)^{\frac{1}{2}} (\|\boldsymbol{v}\|^2_{\Lambda_\alpha} + \|q\|^2)^{\frac{1}{2}}. \tag{2.13}$$

(2) *There is a constant* $\alpha_0 > 0$ *independent of* α *such that*

$$\sup_{\boldsymbol{v}\in\boldsymbol{V},\, q\in W} \frac{\mathcal{A}(\boldsymbol{u},p;\boldsymbol{v},q)}{(\|\boldsymbol{v}\|^2_{\Lambda_\alpha} + \|q\|^2)^{\frac{1}{2}}} \ge \alpha_0 (\|\boldsymbol{u}\|^2_{\Lambda_\alpha} + \|p\|^2)^{\frac{1}{2}}. \tag{2.14}$$

Proof. The first inequality follows immediately by applying the Schwarz inequality to all three terms and keeping in mind that α is a positive function. Proving the inf-sup condition (2.14) is equivalent to proving the following inequality (see [9]):

$$\inf_{q\in W} \sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{(\nabla\cdot\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\Lambda_\alpha}\|q\|} \ge \gamma > 0 \qquad \text{for all} \quad \boldsymbol{v}\in\boldsymbol{V}, \quad \text{for all} \quad q \in W. \tag{2.15}$$

As is well known, if γ is independent of the contrast κ, then so is α_0. For more details on the relation between the constants γ and α_0, we refer the reader to [34]; see also Remark 2.1. Furthermore, due to assumption (2.7) we have that

$$\|\boldsymbol{v}\|_{\Lambda_\alpha} \le \|\boldsymbol{v}\|_{\boldsymbol{H}(\mathrm{div})} \quad \text{so that} \quad \inf_{q\in W} \sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{(\nabla\cdot\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\Lambda_\alpha}\|q\|} \ge \inf_{q\in W} \sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{(\nabla\cdot\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\boldsymbol{H}(\mathrm{div})}\|q\|} \ge \gamma > 0.$$

To find a computable bound for the constant γ, we can use the standard construction [6] for the case $K(x) = 1$ in Ω. For $q \in W$ we take $\boldsymbol{w} = \nabla\varphi \in \boldsymbol{V}$, where $\varphi \in H^1_D(\Omega)$ is the solution to the variational problem $(\nabla\varphi, \nabla\chi) = (q,\chi)$ for all $\chi \in H^1_D(\Omega)$. Then $\operatorname{div}\boldsymbol{w} = -q$ in $L^2(\Omega)$ by construction, and using the above Poincaré inequality we get $\|\boldsymbol{w}\| \le \sqrt{C_P}\|q\|$ so that

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{(q,\operatorname{div}\boldsymbol{v})}{\|\boldsymbol{v}\|_{\boldsymbol{H}(\mathrm{div})}} \ge \frac{(q,\operatorname{div}\boldsymbol{w})}{\|\boldsymbol{w}\|_{\boldsymbol{H}(\mathrm{div})}} = \frac{\|q\|^2}{(\|\boldsymbol{w}\|^2 + \|\operatorname{div}\boldsymbol{w}\|^2)^{\frac{1}{2}}} \ge \frac{\|q\|}{\sqrt{C_P+1}}.$$

This shows (2.15) with $\gamma = 1/\sqrt{C_P+1}$, where C_P is the constant in the Poincaré inequality (2.12). Then using the results of [34, 20] and inequalities (2.13) and (2.15), we deduce that the constant α_0 in (2.14) is positive. A sharp lower bound for α_0 can be obtained using the best known results of [20, Theorem 1] to get $\alpha_0 \ge 1/(2+C_P)$, which completes the proof. ∎

Remark 2.1. As mentioned above, the case of scalar permeability $K(x)$, $0 < K(x) < \infty$, needs a different computational approach. First, one establishes a special (weighted) Poincaré inequality (involving the weighted L^2-norm) with a constant $C_P > 0$:

$$\|q\|^2_{0,K} := \int_\Omega K(x) q^2\, dx \le C_P \|\nabla q\|^2_{0,K}, \quad \text{where} \quad \|\nabla q\|^2_{0,K} = \int_\Omega K(x)|\nabla q|^2\, dx. \tag{2.16}$$

This type of inequality plays a role in domain decomposition methods, multiscale FEMs, and multigrid preconditioners, and has been studied in, e.g., [7, 10, 22, 23, 27]. Particularly relevant to our work is the study conducted in [22, 23, 27], where, under

certain restrictions on the distribution of the permeability $K(x)$, the constant C_P in (2.16) is shown to be independent of the contrast κ. Then using (2.16) one can prove the following inf-sup condition:

(2.17)
$$\inf_{q\in W}\sup_{\boldsymbol{v}\in\boldsymbol{V}}\frac{(\nabla\cdot\boldsymbol{v},q)}{(\|\boldsymbol{v}\|_{0,\alpha}^2+\|\nabla\cdot\boldsymbol{v}\|_{0,\alpha}^2)^{1/2}\|q\|_{0,K}}\geq\frac{1}{\sqrt{C_P+1}}\qquad\text{for all}\quad\boldsymbol{v}\in\boldsymbol{V},\quad q\in W.$$

However, this approach needs additional research for preconditioning the weighted $\boldsymbol{H}$(div)-norm $(\|\boldsymbol{v}\|_{0,\alpha}^2+\|\nabla\cdot\boldsymbol{v}\|_{0,\alpha}^2)^{1/2}$ and is left for future consideration.

3. FEM approximations.

3.1. FE partitioning and spaces. We assume that the domain Ω is connected and is triangulated with d-dimensional simplexes or bricks. The triangulation is denoted by $\mathcal{T}_h$, with the simplexes forming $\mathcal{T}_h$ assumed to be shape regular (the ratio between the diameter of a simplex and the inscribed ball is bounded above). We consider the FE approximation of problem (1.1) using the finite dimensional spaces $\boldsymbol{V}_h\subset\boldsymbol{V}$ and $W_h\subset W$ of piecewise polynomial functions.

It is well known that for the vector variable $\boldsymbol{u}$ we can use the $\boldsymbol{H}$(div)-conforming Raviart–Thomas space $\mathcal{RT}_k$ or Brezzi–Douglas–Marini $\mathcal{BDM}_{k+1}$ FEs. However, since the solution of the problem has low regularity, it is natural to use lowest-order FE spaces. For the vector variable $\boldsymbol{u}$ we use the standard Raviart–Thomas $\mathcal{RT}_0$ for simplexes and rectangles/bricks. In the case of simplexes we can also apply Brezzi–Douglas–Marini BDM_1 FEs. Since $W\equiv L^2(\Omega)$, for its FE counterpart we can use piecewise constant functions over the partition $\mathcal{T}_h$. We show that the FEM is uniformly stable with respect to the contrast κ.

3.2. Stability of the mixed FEM. We take

$$\boldsymbol{V}_h=\{\boldsymbol{v}\in\boldsymbol{V}:\ \boldsymbol{v}|_T\in\mathcal{RT}_0\ \text{ for } T\in\mathcal{T}_h\},\tag{3.1}$$

$$W_h=\{q\in L^2(\Omega):\ q|_T\in\mathcal{P}_0;\ \text{i.e., } q \text{ is a piecewise constant function on } \mathcal{T}_h\}.\tag{3.2}$$

The mixed FE approximation of problem (1.1) is as follows: find $\boldsymbol{u}_h\in\boldsymbol{V}_h$ and $p_h\in W_h$ such that

$$\mathcal{A}(\boldsymbol{u}_h,p_h;\boldsymbol{v},q)=-(f,q)\quad\text{for all}\quad(\boldsymbol{v},q)\in\boldsymbol{V}_h\times W_h,\tag{3.3}$$

where the bilinear form $\mathcal{A}(\boldsymbol{u}_h,p_h;\boldsymbol{v},q)$ is defined by (2.11). Our goal is to establish a discrete variant of the inf-sup condition.

LEMMA 3.1. *Let $\boldsymbol{V}_h$ be the space defined by* (3.1)*, and let W_h be the space defined by* (3.2)*. Assume also that the permeability coefficient $K(x)$ satisfies inequality* (2.7)*. Then, independent of the contrast κ and the mesh size h, the following inequality holds true:*

$$\inf_{q_h\in W_h}\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\frac{(\operatorname{div}\boldsymbol{v}_h,q_h)}{\|\boldsymbol{v}_h\|_{\Lambda_\alpha}\|q_h\|}\geq\gamma>0.\tag{3.4}$$

Proof. First we note that the inf-sup condition for the case $K(x)=1$ is well known [6, 9]. Then, using the same argument as in the proof of Lemma 2.1, we show the desired result. Note that the constant γ will depend on the constant C_P of the Poincaré inequality and the properties of the FE partitioning, but is not dependent on the contrast κ. ∎

As a consequence of Lemma 3.1 and (2.13) we have the following result.

THEOREM 3.1. *Assume that the permeability coefficient $K(x)$ satisfies inequality* (2.7)*. Then the following bounds are valid for all $\boldsymbol{u} \in \boldsymbol{V}_h$ and $p \in W_h$:*

$$\alpha_0(\|\boldsymbol{u}\|^2_{\Lambda_\alpha} + \|p\|^2)^{\frac{1}{2}} \le \sup_{\boldsymbol{v}\in\boldsymbol{V}_h, q\in W_h} \frac{\mathcal{A}(\boldsymbol{u}, p; \boldsymbol{v}, q)}{(\|\boldsymbol{v}\|^2_{\Lambda_\alpha} + \|q\|^2)^{\frac{1}{2}}} \le (\|\boldsymbol{u}\|^2_{\Lambda_\alpha} + \|p\|^2)^{\frac{1}{2}}. \tag{3.5}$$

The constant $\alpha_0 > 0$ may depend on the shape regularity of the mesh. However, it is independent of the contrast κ and the mesh size h. In fact, $\alpha_0 \ge 1/(1 + 1/\gamma^2)$, where γ is the constant in (3.4)*.*

4. Preconditioning. The goal of this section is to describe a uniform, with respect to mesh size h and coefficient variation (contrast) κ, preconditioner for the algebraic problem resulting from the Galerkin method (3.3). We write (3.3) as an operator equation in the space $X_h = \boldsymbol{V}_h \times W_h$ equipped with the norm $\|\boldsymbol{x}_h\|^2_{X_h} = \|\boldsymbol{u}_h\|^2_{\Lambda_\alpha} + \|p_h\|^2$ for $\boldsymbol{x}_h = (\boldsymbol{u}_h, p_h)$. We have

$$\mathcal{A}_h \boldsymbol{x}_h = \boldsymbol{f}_h \quad \text{for} \quad \boldsymbol{f}_h = (\boldsymbol{0}, -f_h) \in X_h, \tag{4.1}$$

where for all $\boldsymbol{y}_h = (\boldsymbol{v}_h, q_h) \in X_h$

$$\langle \mathcal{A}_h \boldsymbol{x}_h, \boldsymbol{y}_h \rangle = \mathcal{A}(\boldsymbol{u}_h, p_h; \boldsymbol{v}_h, q_h).$$

Here, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between X_h and $X_h^\star$. Clearly, the operator $\mathcal{A}_h : X_h \to X_h^\star$ is symmetric on $X_h = \boldsymbol{V}_h \times W_h$ and indefinite. Moreover, from (3.5) we have that $\|\mathcal{A}_h\|_{\mathcal{L}(X_h, X_h^\star)} \le c_0$ and $\|\mathcal{A}_h^{-1}\|_{\mathcal{L}(X_h^\star, X_h)} \le c_1 = \frac{1}{\alpha_0}$; i.e., $\mathcal{A}_h$ is a uniform isomorphism. By *uniform isomorphism*, here and in the following, we mean an operator which is bounded and has a bounded inverse, with the bounds independent of the mesh size and the coefficient variation.

Following the general framework on constructing preconditioners for discretized systems of PDEs developed in [21], as a preconditioner we may choose *any* SPD uniform isomorphism (in the sense defined above) $\mathcal{B}_h : X_h \to X_h^\star$. A standard choice for $\mathcal{B}_h$ is an operator (a diagonal preconditioner) of the form

$$\mathcal{B}_h := \begin{bmatrix} B_h & 0 \\ 0 & I_h \end{bmatrix}, \tag{4.2}$$

where I_h is the Riesz isomorphism $I_h : W_h \mapsto W_h^*$. The operator $B_h : \boldsymbol{V}_h \to \boldsymbol{V}_h^*$ is chosen so that

$$\langle B_h \boldsymbol{u}_h, \boldsymbol{u}_h \rangle \eqsim \|\boldsymbol{u}_h\|^2_{\Lambda_\alpha} \quad \text{or, equivalently,} \quad \langle \mathcal{B}_h \boldsymbol{x}_h, \boldsymbol{x}_h \rangle \eqsim \|\boldsymbol{x}_h\|^2_{X_h}, \tag{4.3}$$

where $\boldsymbol{x}_h = (\boldsymbol{u}_h, p_h) \in X_h$ and "$\eqsim$" stands for a norm equivalence, uniform with respect to h and κ. The theory in [21, section 2] shows that the norm equivalence (4.3) guarantees a uniform bound on the condition number of the preconditioned system $\|\mathcal{B}_h^{-1}\mathcal{A}_h\|_{\mathcal{L}(X,X)})\|(\mathcal{B}_h^{-1}\mathcal{A}_h)^{-1}\|_{\mathcal{L}(X,X)}$. Hence, a Krylov subspace method with $\mathcal{B}_h$ as a preconditioner has a convergence rate independent of h or the contrast κ (see also [25, 26, 30]).

Remark 4.1. We note that any successful development of a robust preconditioner for the bilinear form $\Lambda_\alpha(\cdot, \cdot)$ could also be used in the mixed-FE-least-squares approximation of the problem (1.1). In such approximation (see, e.g., [24]), one of the principal minors of the resulting matrix also corresponds to the bilinear form $\Lambda_\alpha(\cdot, \cdot)$.

In the rest of this section, we construct an operator B_h which, as shown later in section 5, leads to a uniform preconditioner $\mathcal{B}_h$ for $\mathcal{A}_h$.

4.1. FE problem and matrix notation. The derivation and the justification of the preconditioner are in the framework of algebraic multilevel/multigrid methods. As a first step we rewrite the operator equation (4.1) in a matrix form. Instead of the functions $\boldsymbol{x}_h = (\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times W_h$, we use vectors consisting of the degrees of freedom determining $\boldsymbol{x}_h$ through the nodal basis functions, namely,

$$\mathbf{x} = \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix}, \quad \text{where} \quad \mathbf{u} \in \mathbb{R}^{|\mathcal{E}_h|}, \quad \mathbf{p} \in \mathbb{R}^{|\mathcal{T}_h|}$$

are column vectors, $|\mathcal{E}_h|$ is the number of edges in $\mathcal{E}_h$, excluding those on Γ_N, and $|\mathcal{T}_h|$ is the number of rectangles of the partition $\mathcal{T}_h$. Then A, B_{div}, $\widetilde{A}$, and R will denote matrices being either square or rectangular. As a result of this convention, (4.1) can be written in the matrix form (1.2). Our aim now is to derive and study a preconditioner for algebraic systems of the form (1.2), which, due to the above considerations, reduces to the efficient preconditioning of the system

$$A\mathbf{u} = \mathbf{b}, \quad \mathbf{u}, \mathbf{b} \in \mathbb{R}^N, \quad N := |\mathcal{E}_h|. \tag{4.4}$$

4.2. Robust preconditioning of the weighted $\boldsymbol{H}$(div)-norm. In [14] additive Schur complement approximation (ASCA) has been introduced as a tool for constructing robust coarse spaces for high-frequency, high-contrast problems. Recently, this technique has also been utilized as a building block for a new class of multigrid methods in which a coarse-grid correction, as used in standard multigrid algorithms, is replaced by an auxiliary space correction [16]. Viewed as a block factorization algorithm, the major computations in this so-called auxiliary space multigrid (ASMG) method can be performed in parallel since they consist of a two-level block factorization of local FE stiffness matrices associated with a partitioning of the domain into overlapping or nonoverlapping subdomains. The analysis of the two-grid ASMG preconditioner relies on the fictitious space lemma; see [16]. However, the underlying construction is purely algebraic and thus essentially differs from the methodology in [12].

In this section we recall the basic construction of the ASMG method and specify modifications that allow its successful application to the linear systems arising from $\boldsymbol{H}$(div)-conforming discretizations of the subproblem involving the weighted $\boldsymbol{H}$(div) bilinear form (2.5). For details on ASCA we refer the reader to [14, 16].

4.2.1. Auxiliary space multigrid method. Let $k = 0, 1, \ldots, \ell-1$ be the index of mesh coarsening where $k = 0$ corresponds to the finest mesh; i.e., $A^{(0)} := A_h = A$ denotes the fine-grid matrix. Consider a sequence of auxiliary space matrices $\widetilde{A}^{(k)}$ in the 2×2 block factorized form

$$(\widetilde{A}^{(k)})^{-1} = (\widetilde{L}^{(k)})^T \widetilde{D}^{(k)} \widetilde{L}^{(k)},$$

$$\widetilde{L}^{(k)} = \begin{bmatrix} I & \\ -\widetilde{A}_{21}^{(k)} (\widetilde{A}_{11}^{(k)})^{-1} & I \end{bmatrix}, \quad \widetilde{D}^{(k)} = \begin{bmatrix} \widetilde{A}_{11}^{(k)} & \\ & Q^{(k)} \end{bmatrix}^{-1},$$

where the Schur complement $Q^{(k)} = \widetilde{A}_{22}^{(k)} - \widetilde{A}_{21}^{(k)} (\widetilde{A}_{11}^{(k)})^{-1} \widetilde{A}_{12}^{(k)}$ of $\widetilde{A}^{(k)}$ defines the ASCA of the Schur complement $S^{(k)} = A_{22}^{(k)} - A_{21}^{(k)} (A_{11}^{(k)})^{-1} A_{12}^{(k)}$ of $A^{(k)}$; $A^{(k)} \in \mathbb{R}^{N^{(k)} \times N^{(k)}}$ and $\widetilde{A}^{(k)} \in \mathbb{R}^{\widetilde{N}^{(k)} \times \widetilde{N}^{(k)}}$ are related via

$$A^{(k)} = R^{(k)} \widetilde{A}^{(k)} (R^{(k)})^T; \tag{4.5}$$

and $Q^{(k)}$ defines the next coarser level matrix, i.e.,

$$A^{(k+1)} := Q^{(k)}; \tag{4.6}$$

for details, see [16, 15].

Now define the (nonlinear) AMLI-cycle ASMG preconditioner $C^{(k)}$ at level k by

$$C^{(k)^{-1}} := \Pi^{(k)}(\widetilde{L}^{(k)})^T \begin{bmatrix} \widetilde{A}_{11}^{(k)} & \\ & C_\nu^{(k+1)} \end{bmatrix}^{-1} \widetilde{L}^{(k)}\Pi^{(k)^T}, \tag{4.7}$$

where $[C_\nu^{(k+1)}]^{-1}$ is an approximation of the inverse of the coarse-level matrix (4.6) for $k+1 < \ell$ and $[C_\nu^{(\ell)}]^{-1} := A^{(\ell)^{-1}}$ at the coarsest level ℓ. The operator $\Pi^{(k)} : \mathbb{R}^{\widetilde{N}^{(k)}} \to \mathbb{R}^{N^{(k)}}$ is a surjective map. Its construction is typically related to a proper approximation $\widetilde{D}^{(k)}$ of $\widetilde{A}^{(k)}$, e.g.,

$$\Pi^{(k)} = \Pi_{\widetilde{D}^{(k)}} := (R^{(k)}\widetilde{D}^{(k)}(R^{(k)})^T)^{-1}R^{(k)}\widetilde{D}^{(k)}, \tag{4.8}$$

where in what follows $\widetilde{D}^{(k)}$ will be of the particular form

$$\widetilde{D}^{(k)} = \begin{bmatrix} \widetilde{D}_{11}^{(k)} & 0 \\ 0 & I^{(k)} \end{bmatrix}, \tag{4.9}$$

and $\widetilde{D}_{11}^{(k)} = \widetilde{A}_{11}^{(k)}$ or $\widetilde{D}_{11}^{(k)} = \mathrm{diag}(\widetilde{A}_{11}^{(k)})$, for example.

In the linear AMLI cycle for $k < \ell - 1$ one employs a matrix polynomial of the form

$$\left[C_\nu^{(k+1)}\right]^{-1} := (I - p^{(k)}(C^{(k+1)^{-1}}A^{(k+1)}))A^{(k+1)^{-1}}. \tag{4.10}$$

If the polynomial $p^{(k)}(t)$ satisfies the condition $p^{(k)}(0) = 1$, we have the equivalent expression

$$\left[C_\nu^{(k+1)}\right]^{-1} = q^{(k)}(C^{(k+1)^{-1}}A^{(k+1)})C^{(k+1)^{-1}}, \tag{4.11}$$

where $q^{(k)}(t) = (1 - p^{(k)}(t))/t$. Then the application of $[C_\nu^{(k+1)}]^{-1}$ requires the action of the inverse of $C^{(k+1)}$ only. A classic choice for $p^{(k)}(t)$ is a scaled and shifted Chebyshev polynomial of degree ν. Other polynomials are possible, e.g., choosing $q^{(k)}(t)$ to be the polynomial of best approximation to $1/t$ in a uniform norm; see [19].

If we incorporate pre- and postsmoothing, the AMLI-cycle ASMG preconditioner $B^{(k)}$ at level k is given by

$$B^{(k)^{-1}} := \overline{M}^{(k)^{-1}} + (I - M^{(k)^{-T}}A^{(k)})\Pi^{(k)}(\widetilde{L}^{(k)})^T\overline{D}^{(k)^{-1}}\widetilde{L}^{(k)}\Pi^{(k)^T}(I - A^{(k)}M^{(k)^{-1}}), \tag{4.12}$$

where

$$\overline{D}^{(k)} := \begin{bmatrix} \widetilde{A}_{11}^{(k)} & \\ & B_\nu^{(k+1)} \end{bmatrix} \quad \text{and} \quad \left[B_\nu^{(k+1)}\right]^{-1} = q^{(k)}(B^{(k+1)^{-1}}A^{(k+1)})B^{(k+1)^{-1}}.$$

For the nonlinear AMLI-cycle ASMG method, $[B_\nu^{(k+1)}]^{-1} \equiv B_\nu^{(k+1)}[\cdot]$ (or $[C_\nu^{(k+1)}]^{-1} \equiv C_\nu^{(k+1)}[\cdot]$) is a nonlinear mapping whose action on a vector $\mathbf{d}$ is realized by ν iterations using a preconditioned Krylov subspace method. In the following the generalized conjugate gradient method serves this purpose, and hence we denote $B_\nu^{(k+1)}[\cdot] \equiv B_{\mathrm{GCG},\nu}^{(k+1)}[\cdot]$ (and $C_\nu^{(k+1)}[\cdot] \equiv C_{\mathrm{GCG},\nu}^{(k+1)}[\cdot]$). One can find more details in the longer version of this paper, the arXiv preprint [15].

4.2.2. Nonlinear ASMG algorithm for the weighted $\boldsymbol{H}$(div) bilinear form. In the remainder of this section we present the nonlinear ASMG algorithm for preconditioning the SPD matrices arising from discretization of the weighted bilinear form (2.5), and comment on some details of their implementation when specifically using lowest-order Raviart–Thomas elements on rectangles.

Consider a *partitioning of the set* $\mathcal{D}$ of degrees of freedom (DOF). We illustrate the construction for the case of two grids, a coarse grid $\mathcal{T}_H$ and a fine grid $\mathcal{T}_h$, where $H = 2h$. Then the corresponding sets of edges are $\mathcal{E}_H$ and $\mathcal{E}_h$ with the following relations being obvious: $4|\mathcal{T}_H| = |\mathcal{T}_h|$ and $2|\mathcal{E}_H| + 4|\mathcal{T}_H| = |\mathcal{E}_h|$. Since for lowest-order Raviart–Thomas FEs it is not immediately clear how to partition $\mathcal{D}$, we perform a preprocessing step which consists of a *compatible two-level basis transformation* (see also [14]). The global matrix A is transformed according to

$$\widehat{A} = J^T A J, \quad \widehat{A}, J, A \in \mathbb{R}^{|\mathcal{E}_h|\times|\mathcal{E}_h|}, \tag{4.13}$$

where the transformation matrix J is the product of a permutation matrix P and another transformation matrix $J_\pm$, i.e.,

$$J = P J_\pm, \quad P, J_\pm \in \mathbb{R}^{|\mathcal{E}_h|\times|\mathcal{E}_h|}. \tag{4.14}$$

The permutation P allows us to provide a two-level numbering of the DOF that splits them into two groups, the first consisting of DOF associated with fine-grid edges not part of any coarse-grid edge (interior DOF), and the second keeping all remaining DOF ordered such that any two that are on the same coarse edge have consecutive numbers. The transformation matrix $J_\pm$ in (4.14) is of the form

$$J_\pm = \begin{bmatrix} I & \\ & J_{22} \end{bmatrix}, \quad \text{where} \quad I \in \mathbb{R}^{(4|\mathcal{T}_H|)\times(4|\mathcal{T}_H|)}$$

and

$$J_{22} = \frac{1}{2}\begin{bmatrix} 1 & -1 & & & & & & & \\ & & 1 & -1 & & & & & \\ & & & & \ddots & \ddots & & & \\ & & & & & & -1 & 1 & \\ 1 & 1 & & & & & & & \\ & & 1 & 1 & & & & & \\ & & & & \ddots & \ddots & & & \\ & & & & & & 1 & 1 & \end{bmatrix} \quad \text{with} \quad J_{22} \in \mathbb{R}^{(2|\mathcal{E}_H|)\times(2|\mathcal{E}_H|)}.$$

The analogous transformation is performed at a local level on each subdomain Ω_i; i.e.,

$$\widehat{A}_i = J_i^T A_i J_i, \tag{4.15}$$

where again $J_i = P_i J_{\pm,i}$, with P_i the permutation as explained above but performed on the DOF in the subdomain Ω_i. $J_{\pm,i}$ has its usual meaning but is restricted to the subdomain Ω_i.

The introduced transformation matrix J defines a splitting of the DOF into fine (FDOF) and coarse (CDOF), where for every coarse edge one CDOF is generated by a half sum, and one FDOF by a half difference. The set of FDOF is then complemented by the DOF corresponding to all fine edges that do not form coarse edges.

Note that this transformation has been successfully used to identify CDOF in the context of nonconforming FEMs for scalar elliptic problems in [11, 18], and the ASMG method described in this section is straightforwardly applicable to the arising discrete problems. When constructed recursively on all coarse levels, the resulting two-level transformation matrices, referring to levels $k = 0, 1, \ldots, \ell - 1$, are denoted by $J^{(k)}$.

Finally, the nonlinear ASMG preconditioner employs the following two-level matrices:

$$\widehat{A}^{(k)} = {J^{(k)}}^T A^{(k)} J^{(k)}$$

for all $k < \ell$. Its application to a vector $\widehat{\mathbf{d}}$ for the two-level basis at level k can be formulated as follows.

ALGORITHM 4.1. Action of preconditioner (4.7) on a vector $\widehat{\mathbf{d}} = {J^{(k)}}^T \mathbf{d}$ at level k: $\widehat{C}^{(k)}[\widehat{\mathbf{d}}]$.

Auxiliary space correction:

$$\begin{cases} \begin{pmatrix} \tilde{\mathbf{q}}_1 \\ \tilde{\mathbf{q}}_2 \end{pmatrix} := \tilde{\mathbf{q}} = \Pi^T_{\widetilde{D}^{(k)}} \widehat{\mathbf{d}}, \\ \tilde{\mathbf{p}}_1 = (\widetilde{A}^{(k)}_{11})^{-1} \tilde{\mathbf{q}}_1, \\ \tilde{\mathbf{p}}_2 = J^{(k+1)} C^{(k+1)}_{\mathrm{GCG},\nu} [{J^{(k+1)}}^T (\tilde{\mathbf{q}}_2 - \widetilde{A}^{(k)}_{21} \tilde{\mathbf{p}}_1)], \\ \tilde{\mathbf{q}}_1 = \tilde{\mathbf{p}}_1 - (\widetilde{A}^{(k)}_{11})^{-1} \widetilde{A}^{(k)}_{12} \tilde{\mathbf{p}}_2, \\ \tilde{\mathbf{q}}_2 = \tilde{\mathbf{p}}_2, \\ \widehat{C}^{(k)}[\widehat{\mathbf{d}}] := \Pi_{\widetilde{D}^{(k)}} \tilde{\mathbf{q}}. \end{cases}$$

By incorporating pre- and postsmoothing, the realization of the preconditioner (4.12) takes the following form.

ALGORITHM 4.2. Action of preconditioner (4.12) on a vector $\widehat{\mathbf{d}}$ at level k: $\widehat{B}^{(k)}[\widehat{\mathbf{d}}]$.

Presmoothing:	$\widehat{\mathbf{u}} = (\widehat{M}^{(k)})^{-1} \widehat{\mathbf{d}}$,
Auxiliary space correction:	$\widehat{\mathbf{v}} = \widehat{\mathbf{u}} + \widehat{C}^{(k)} [\widehat{\mathbf{d}} - \widehat{A}^{(k)} \widehat{\mathbf{u}}]$,
Postsmoothing:	$\widehat{B}^{(k)}[\widehat{\mathbf{d}}] := \widehat{\mathbf{v}} + (\widehat{M}^{(k)})^{-T} (\widehat{\mathbf{d}} - \widehat{A}^{(k)} \widehat{\mathbf{v}})$.

4.3. On the complexity of the ASMG preconditioner. We now address the important topic of estimating the computational work required for performing the action of the ASMG preconditioner. Clearly, from the algorithm descriptions given earlier, the number of flops required to evaluate such an action is proportional to the *operator complexity* of the preconditioner, defined as the total number of nonzeros in the matrices on all levels. The most general algebraic multilevel preconditioners are usually constructed using the combinatorial graph structure of the underlying matrices (on the finest and coarser levels). Estimating the operator complexities in such cases is not only difficult, but in most cases impossible due to the fact that such estimates should hold for the set of *all* possible graphs. Reliable estimates are usually done for algorithms that construct coarse levels using at least some of the geometric information from the underlying problem. This is the case we consider here, and we also refer the reader to [5, 32] for more insight into how geometric information can be used to bound the operator complexity of a multilevel preconditioner.

Given a matrix $A \in \mathbb{R}^{N \times N}$, we characterize the nonzero structure of the ASMG coarse level matrix Q. To construct Q, recall that we first need to split the set of the DOF as a union of subsets, $\{1, \ldots, N\} = \cup_{i=1}^{n} \omega_i$. We assume that $\omega_i = \{\mathcal{F}_i, \mathcal{C}_i\}$, where $\mathcal{F}_i$ is a set of FDOF, and $\mathcal{C}_i$ is a set of CDOF, with $\mathcal{F}_i \cap \mathcal{C}_i = \emptyset$. The total number of

CDOF is $N_{\mathcal{C}} = \left|\cup_{j=1}^{n} \mathcal{C}_i\right|$. Since our considerations are permutation invariant, without loss of generality, we assume that globally we have numbered first the CDOF, and thus we have $\mathcal{C}_i \subset \{1, \ldots, N_{\mathcal{C}}\}$.

We also set $N_i = |\omega_i|$ and $n_i = |\mathcal{C}_i|$. We denote by $\mathbf{e}_k$ the kth Euclidean basis vector in $\mathbb{R}^N$; when we consider the canonical basis in $\mathbb{R}^m$, $m \neq N$, we use the notation $\mathbf{e}_{k(m)}$ for the kth basis vector. With each ω_i we associate a matrix $R_i \in \mathbb{R}^{N_i \times N}$, and, for $\omega_i = \{j_1, \ldots, j_{N_i}\}$, we set $R_i^T = [\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_{N_i}}]$. Next, we consider a fine-grid matrix A given by the identity

$$A = \sum_{i=1}^{n} R_i^T A_i R_i = \sum_{i=1}^{n} \left[R_{i,\mathcal{F}}^T, R_{i,\mathcal{C}}^T\right] \begin{bmatrix} A_{i,\mathcal{F}} & A_{i,\mathcal{F}\mathcal{C}} \\ A_{i,\mathcal{C}\mathcal{F}} & A_{i,\mathcal{C}} \end{bmatrix} \begin{bmatrix} R_{i,\mathcal{F}} \\ R_{i,\mathcal{C}} \end{bmatrix},$$

where we used a block form of the matrices corresponding to the splitting of ω_i into $\mathcal{F}$-ine level and $\mathcal{C}$-oarse level degrees of freedom. The Schur complements S_i used in the definition of the coarse-grid matrix Q are defined as $S_i = A_{i,\mathcal{C}} - A_{i,\mathcal{C}\mathcal{F}} A_{i,\mathcal{F}}^{-1} A_{i,\mathcal{F}\mathcal{C}}$. Recall that the coarse-grid matrix Q is then defined by $Q = \sum_{i=1}^{n} \widetilde{R}_{i,\mathcal{C}}^T S_i \widetilde{R}_{i,\mathcal{C}}$, and we have $Q \in \mathbb{R}^{N_{\mathcal{C}} \times N_{\mathcal{C}}}$. If we now use our assumption that the CDOF are numbered first, then $\widetilde{R}_{i,\mathcal{C}} \in \mathbb{R}^{n_i \times N_{\mathcal{C}}}$ is formed by the first $N_{\mathcal{C}}$ columns of $R_{i,\mathcal{C}} \in \mathbb{R}^{n_i \times N_{\mathcal{C}}}$. Next, we introduce the vectors

$$\mathbf{1}_i = (\underbrace{1, \ldots, 1}_{n_i})^T \quad \text{and} \quad \boldsymbol{\chi}_i = \sum_{j \in \mathcal{C}_i} \mathbf{e}_{j(N_{\mathcal{C}})}.$$

For a fixed i, the vector $\boldsymbol{\chi}_i \in \mathbb{R}^N$ is the indicator vector of the set $\mathcal{C}_i$ as a subset of $\{1, \ldots, N_{\mathcal{C}}\}$. Its components are equal to 1 for indices in $\mathcal{C}_i$, and equal to zero otherwise. We note that $\mathbf{1}_i \mathbf{1}_i^T$ is the $n_i \times n_i$ matrix of all ones, and we encourage the reader to check the identity $\boldsymbol{\chi}_i = \widetilde{R}_{i,\mathcal{C}}^T \mathbf{1}_i$.

To describe the nonzero structure of Q, we introduce the set $\mathcal{B}_m$ of Boolean $(m \times m)$ matrices whose entries are from the set $\{0, 1\}$. We introduce a mapping $\mathrm{nz} : \mathbb{R}^{m \times m} \mapsto \mathcal{B}_m$ such that $[\mathrm{nz}(A)]_{ij} = 0$ if and only if $A_{ij} = 0$, and $[\mathrm{nz}(A)]_{ij} = 1$ otherwise. We say that $X \preceq Y$ if $[\mathrm{nz}(Y) - \mathrm{nz}(X)]$ is a matrix with nonnegative entries. This is a formal way to state that the nonzero structure of Y contains the nonzero structure of X or, equivalently, that every zero in Y is also a zero in X. Clearly, $S_i \preceq \mathbf{1}_i \mathbf{1}_i^T$, and, as a consequence, we have the following relation characterizing the sparsity of Q:

$$Q \preceq \sum_{i=1}^{n} \sum_{i=1}^{n} \widetilde{R}_{i,\mathcal{C}}^T \mathbf{1}_i \mathbf{1}_i^T \widetilde{R}_{i,\mathcal{C}} = \sum_{i=1}^{n} \boldsymbol{\chi}_i \boldsymbol{\chi}_i^T =: X. \tag{4.16}$$

Note that from the right-hand side of (4.16) we can conclude that Q_{km} may be nonzero only in the case when there exists i such that $k \in \mathcal{C}_i$ and $m \in \mathcal{C}_i$. Using (4.16) it is easy to compute a bound on the number of nonzeros $n_{z,j}$ for fixed column j in Q. We have

$$n_{z,j} \leq \|X \mathbf{e}_{j(N_{\mathcal{C}})}\|_{\ell^1} = \sum_{i : j \in \mathcal{C}_i} |\mathcal{C}_i|.$$

As is immediately seen, the number of nonzeros per column in Q is bounded by a constant independent of N if the following two conditions are satisfied: (i) the number of CDOF in each $\mathcal{C}_i$ is bounded, and (ii) every CDOF lies in a bounded number of subsets $\mathcal{C}_i$.

As a simple, but instructive, example of how conditions (i) and (ii) can be satisfied, let us consider a PDE discretized by an FEM on a quasi-uniform grid with characteristic mesh size h in two dimensions. The considerations are independent of the PDE or the order of the FE spaces (but the constants hidden in "$\lesssim$" below may depend on the FE spaces and the order of polynomials). To define the sets ω_i on such a grid, we proceed as follows: (1) place a regular (square) auxiliary grid of size γh, $\gamma \geq 2$, that contains Ω, (2) set n to be the number of vertices on the auxiliary grid lying in Ω, and (3) choose ω_i to be the set of DOF lying in the support of the bilinear basis function corresponding to the ith vertex. Then we have that $|\mathcal{C}_i| < N_i \lesssim 4\gamma$ and every CDOF lies in at most four such subdomains. The constant hidden in "$\lesssim$" is a bound on the number of DOF lying in a square of size $2h$. The fact that this bound depends only on the polynomial order and type of FE spaces follows from the assumption that the mesh is quasi-uniform. For efficient and more sophisticated techniques using regular, but adaptively refined, auxiliary grids in coarsening algorithms for unstructured problems, we refer the reader to [5, 32]. Such techniques may be directly applied to yield optimal operator complexities for the ASMG preconditioner in the general case of shape-regular grids, although the details are beyond the scope of our consideration here.

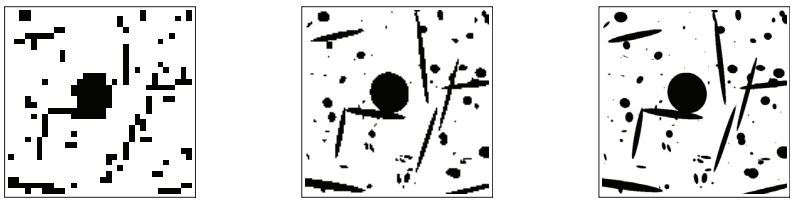
5. Numerical experiments.

5.1. Description of the parameters and the numerical test examples.

Subject to numerical testing are three representative cases characterized by a highly varying coefficient $\alpha(x) = K^{-1}(x)$:

(a) A binary distribution of the coefficient described by islands on which $\alpha = 1.0$ against a background where $\alpha = 10^{-q}$ (see Figure 1);
(b) Inclusions with $\alpha = 1.0$ and a background with a coefficient $\alpha = \alpha_T = 10^{-q_{rand}}$ that is constant on each element $\tau \in \mathcal{T}_h$, where the random integer exponent $q_{rand} \in \{0, 1, 2, \ldots, q\}$ is uniformly distributed (see Figure 2);
(c) Three 2D slices of the SPE10 benchmark problem, where the contrast κ is 10^7 for slices 44 and 74, and 10^6 for slice 54 (see Figure 3).

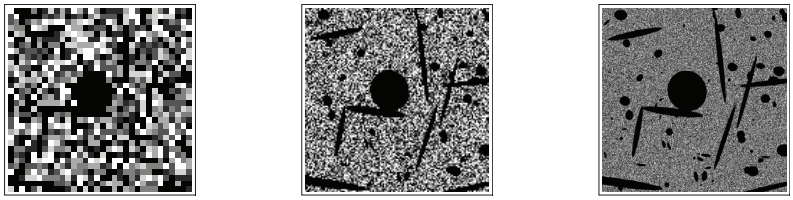
Test problems (a) and (b) are similar to those in other works, e.g., [8, 16, 33]. Example (c) consists of 2D slices of three-dimensional data of the SPE10 benchmark; see [28].



(a) 32×32 mesh (b) 128×128 mesh (c) 512×512 mesh

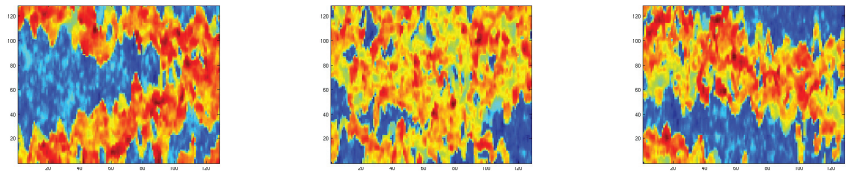
FIG. 1. *Binary distribution of the permeability $K(x)$ corresponding to test case* (a).

Numerical experiments were performed over a uniform mesh consisting of $N \times N$ square elements in $\Omega = (0, 1) \times (0, 1)$, where $N = 4, 8, \ldots, 512$, i.e., up to 525312 velocity DOF and 262144 pressure DOF. We have used a direct method to solve the problems on the coarsest grid. The iterative process has been initialized with a random vector. Its convergence has been tested for linear systems with the right-hand side zero except in the last example, where we have solved the mixed system (1.1) for



(a) 32×32 mesh (b) 128×128 mesh (c) 512×512 mesh

FIG. 2. *Random distribution of* $\alpha = K^{-1}(x)$ *corresponding to test case* (b).



(a) Slice 44 (b) Slice 54 (c) Slice 74

FIG. 3. *Distributions of the permeability* $K(x)$ *along planes* $x_3 = 44, 54, 74$ *from the benchmark SPE*10 *on a* 128×128 *mesh.*

slice 44 of the SPE10 problem with the right-hand side

$$f = \begin{cases} c & \text{for} \quad (x,y) \in \Omega_+ = [0.2, 0.3] \times [0.7, 0.8], \\ -c & \text{for} \quad (x,y) \in \Omega_- = [0.7, 0.8] \times [0.2, 0.3], \\ 0 & \text{for} \quad (x,y) \in \Omega \setminus (\Omega_+ \cup \Omega_-). \end{cases} \tag{5.1}$$

We have used overlapping coverings of the domain where the subdomains are composed of 8×8 elements and overlap with half their width or height. When presenting results, we use the following notation:

- ℓ denotes the number of levels.
- $q = \log \kappa / \log 10$ is the logarithm of the contrast κ.
- n_{ASMG} is the number of ASMG iterations.
- $m \geq 0$ is the number of point Gauss–Seidel pre- and postsmoothing steps.
- ρ_r is the average residual reduction factor defined by

$$\rho_r = \left(\|\mathbf{r}_{n_{ASMG}}\| / \|\mathbf{r}_0\| \right)^{1/n_{ASMG}}, \tag{5.2}$$

  where $\mathbf{u}_{n_{ASMG}}$ is the first iterate (approximate solution of (4.4)) for which the initial residual has decreased by a factor of at least 10^8.
- $\rho_e := \|I - {C^{(0)}}^{-1} A^{(0)}\|_{A^{(0)}}$ is the norm of the error propagation matrix of the linear V-cycle preconditioner (4.7), which is obtained by choosing the polynomial $p_\nu(t) = 1 - t$ in (4.10).

The matrix $\widetilde{D}$ is as defined in (4.9), where $\widetilde{D}_{11} = \widetilde{A}_{11}$. This choice of $\widetilde{D}$ requires an additional preconditioner for the iterative solution of linear systems with the matrix $D = R\widetilde{D}R^T$ a part of the efficient application of the operator $\Pi_{\widetilde{D}}$. Systems with D are solved using the preconditioned conjugate gradient (PCG) method. The stopping criterion for this inner iterative process is a residual reduction by a factor 10^6; the number of PCG iterations required to reach it—where reported—is denoted by n_i.

The preconditioner B_{ILUE} for D is constructed using incomplete factorization with exact local factorization (ILUE). The definition of B_{ILUE} is as follows:

$$B_{ILUE} := LU, \quad U := \sum_{i=1}^{n} R_i^T U_i R_i, \quad L := U^T \mathrm{diag}(U)^{-1},$$

where $D_i = L_i U_i$, $D = \sum_{i=1}^{n} R_i^T D_i R_i$, $\mathrm{diag}(L_i) = I$; for details see [17]. Note that as D_i are the local contributions to D related to the subdomains Ω_i, $i = 1, \ldots, n$, they are all nonsingular.

The following two sections are devoted to the presentation of numerical results. Experiments fall into two categories.

The first category, presented in section 5.2, addresses the evaluation of the performance of the ASMG method on linear systems arising from discretization of the weighted $\boldsymbol{H}(\mathrm{div})$ bilinear form (2.5). All three test cases, (a), (b), and (c), are considered, testing V- and W-cycle methods with and without smoothing. Additionally, we evaluate as a robustness indicator the quantity

$$c_\Pi := \|\pi_{\widetilde{D}}\|^2_{\widetilde{A}} := \|R^T \Pi_{\widetilde{D}}\|^2_{\widetilde{A}}, \tag{5.3}$$

which appears in the condition number estimate $\kappa(C^{-1}A) \leq c_\Pi$ for the two-level preconditioner C which is obtained from $C^{(0)}$ defined in (4.7) by replacing $C_\nu^{(1)}$ with $Q^{(0)}$; see [16].

The second category of experiments, discussed in section 5.3, addresses the solution of the indefinite linear system (1.2) arising from problem (3.3) by a preconditioned MinRes method. The main aims are, on the one hand, to confirm the robustness of the block-diagonal preconditioner (4.2) with respect to arbitrary multiscale coefficient variations, and, on the other hand, to demonstrate its numerical scalability. Furthermore, we include a test problem with a nonzero right-hand side.

5.2. Numerical tests for solving the system (4.4). An ASMG preconditioner was tested for solving the system (4.4) with a matrix corresponding to discretization of the form $\Lambda_\alpha(\boldsymbol{u}, \boldsymbol{v})$.

Example 5.1. First we compute $\|\pi_{\widetilde{D}}\|^2_{\widetilde{A}}$ for up to seven refinement levels of the mesh, and for increasing permeability contrast from 1 to 10^6 in the configurations described in cases (a) and (b). This quantity provides an upper bound for the condition number $\kappa(C^{-1}A)$. The results in Table 1 clearly demonstrate that $\kappa(C^{-1}A)$ is robust with respect to the variation of the contrast.

TABLE 1
Value of $\|\pi_{\widetilde{D}}\|^2_{\widetilde{A}}$ *(defined by* (5.3)*) for the bilinear form* (2.5).

	Case (a)					Case (b)				
$\kappa = 10^q$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$	$\ell = 7$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$	$\ell = 7$
$q = 0$	1.122	1.137	1.148	1.150	1.149	1.122	1.137	1.148	1.150	1.149
$q = 1$	1.148	1.169	1.149	1.152	1.138	1.115	1.126	1.169	1.167	1.123
$q = 2$	1.286	1.338	1.360	1.287	1.126	1.126	1.208	1.119	1.146	1.112
$q = 3$	1.336	1.389	1.418	1.326	1.132	1.014	1.261	1.338	1.334	1.110
$q = 4$	1.343	1.396	1.426	1.334	1.133	1.260	1.295	1.371	1.434	1.110
$q = 5$	1.343	1.397	1.426	1.333	1.369	1.268	1.329	1.394	1.493	1.145
$q = 6$	1.343	1.397	1.426	1.333	1.369	1.255	1.374	1.412	1.139	1.113

Example 5.2. Next we are interested in the convergence factor in the A-norm of the linear V-cycle method; that is, we evaluate the quantity $\rho_e := \|I - {C^{(0)}}^{-1} A^{(0)}\|_{A^{(0)}}$. Moreover, we compare ρ_e to the corresponding value of the average residual reduction factor ρ_r defined according to (5.2). We also report the number of iterations it_e that reduce the initial error in the A-norm by a factor 10^8, and the number of iterations it_r that reduce the Euclidean norm of the initial residual by the same factor. The problem configuration is test case (c) for a zero right-hand side. The results for Example 5.2 are summarized in Table 2. Although the average residual reduction factor ρ_r is much smaller than ρ_e, the error reduction in the A-norm is also surprisingly good, especially in view of the linear V-cycle performing without any additional smoothing, i.e., implementing Algorithm 4.1.

TABLE 2
Example 5.2: *Case* (c), *slice* 44 *of the SPE*10 *benchmark.*

Linear V-cycle: Bilinear form (2.5)				
	ρ_e	it_e	ρ_r	it_r
$\ell = 4$	0.105	7	0.031	6
$\ell = 5$	0.289	9	0.095	8
$\ell = 6$	0.494	12	0.168	11
$\ell = 7$	0.642	14	0.215	12
$\ell = 8$	0.729	17	0.262	14

Example 5.3. Now we test the nonlinear V-cycle and the effect of additional smoothing. Again the problem configuration is test case (c) for a zero right-hand side. We report the number of nonlinear AMLI-cycle ASMG iterations with Algorithm 4.1, denoted by n_{ASMG} for a residual reduction by eight orders of magnitude along with ρ_r. Comparing the results for Example 5.3, listed in Table 3, with those in Table 2 shows that the nonlinear V-cycle typically also reduces the residual norm faster than the linear V-cycle—for the reduction of the A-norm of the error this is a known fact—and the additional incorporation of a point Gauss–Seidel relaxation further accelerates the convergence.

TABLE 3
Example 5.3: *Case* (c), *slice* 44 *of the SPE*10 *benchmark.*

Nonlinear V-cycle: Bilinear form (2.5)						
	$m = 0$		$m = 1$		$m = 2$	
	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r
$\ell = 4$	6	0.032	5	0.025	6	0.027
$\ell = 5$	8	0.093	7	0.062	6	0.045
$\ell = 6$	11	0.157	8	0.091	8	0.083
$\ell = 7$	12	0.202	9	0.123	8	0.094
$\ell = 8$	14	0.243	11	0.172	10	0.154

Example 5.4. The next example tests the dependency of the convergence rate on the contrast. The configuration is test case (a) containing a zero right-hand side, and the number of smoothing steps is $m = 2$. Results in Table 4 show a slight increase in ρ_r with increasing contrast.

Example 5.5. In the next set of numerical experiments we consider the same distribution of inclusions of low permeability as before but this time against a background of a randomly distributed piecewise constant permeability coefficient as shown in Figure 2. The results, presented in Tables 5–7, are even better than those obtained for

TABLE 4
Example 5.4: *Case* (a) *with* $K(x) = 10^q$ *and two smoothing steps* ($m = 2$).

ASMG V-cycle: Bilinear form (2.5), Algorithm 4.1										
	$\ell = 3$		$\ell = 4$		$\ell = 5$		$\ell = 6$		$\ell = 7$	
	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r
$q = 0$	4	0.005	5	0.024	6	0.043	8	0.083	8	0.093
$q = 1$	3	0.002	5	0.022	7	0.058	8	0.084	9	0.121
$q = 2$	3	0.002	5	0.019	7	0.068	8	0.091	9	0.121
$q = 3$	3	0.002	5	0.018	7	0.070	8	0.095	9	0.125
$q = 4$	3	0.002	5	0.017	7	0.069	8	0.098	10	0.142
$q = 5$	3	0.002	5	0.017	8	0.082	9	0.118	10	0.145
$q = 6$	4	0.005	4	0.010	8	0.092	9	0.125	11	0.181

TABLE 5
Example 5.5: *Case* (b)*, no additional smoothing* ($m = 0$).

ASMG V-cycle: Bilinear form (2.5), Algorithm 4.1										
	$\ell = 3$		$\ell = 4$		$\ell = 5$		$\ell = 6$		$\ell = 7$	
	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r
$q = 0$	4	0.007	6	0.027	9	0.102	10	0.156	12	0.210
$q = 1$	4	0.006	6	0.035	9	0.103	11	0.171	13	0.224
$q = 2$	4	0.005	6	0.032	9	0.102	11	0.159	13	0.222
$q = 3$	4	0.006	6	0.042	9	0.110	11	0.174	13	0.229
$q = 4$	4	0.006	7	0.043	9	0.127	11	0.183	13	0.233
$q = 5$	4	0.006	7	0.049	10	0.138	12	0.195	13	0.239
$q = 6$	4	0.006	7	0.056	10	0.149	12	0.207	14	0.252

TABLE 6
Example 5.5: *Case* (b) *with two smoothing steps* ($m = 2$).

ASMG V-cycle: Bilinear form (2.5), Algorithm 4.1										
	$\ell = 3$		$\ell = 4$		$\ell = 5$		$\ell = 6$		$\ell = 7$	
	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r
$q = 0$	4	0.005	5	0.024	6	0.046	8	0.083	8	0.091
$q = 1$	4	0.005	6	0.033	7	0.060	8	0.091	9	0.124
$q = 2$	3	0.002	5	0.023	6	0.045	7	0.069	9	0.121
$q = 3$	3	0.002	5	0.021	6	0.043	7	0.071	8	0.100
$q = 4$	4	0.005	5	0.023	6	0.044	8	0.089	9	0.122
$q = 5$	4	0.005	5	0.024	6	0.045	8	0.090	9	0.125
$q = 6$	4	0.005	6	0.034	6	0.045	8	0.091	10	0.142

TABLE 7
Example 5.5: *Case* (b) *with one smoothing step* ($m = 1$).

ASMG W-cycle: Bilinear form (2.5), Algorithm 4.1										
	$\ell = 3$		$\ell = 4$		$\ell = 5$		$\ell = 6$		$\ell = 7$	
	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r	n_{ASMG}	ρ_r
$q = 0$	4	0.005	4	0.007	4	0.006	4	0.005	4	0.005
$q = 1$	4	0.006	4	0.007	4	0.007	4	0.006	4	0.005
$q = 2$	4	0.004	4	0.009	5	0.016	4	0.007	4	0.006
$q = 3$	4	0.005	5	0.015	5	0.015	4	0.009	4	0.006
$q = 4$	4	0.005	5	0.016	5	0.016	4	0.009	4	0.008
$q = 5$	4	0.005	5	0.018	5	0.015	4	0.009	4	0.008
$q = 6$	4	0.005	5	0.019	5	0.015	4	0.008	4	0.007

the binary distribution in the sense that both the V- and W-cycles are robust with respect to the contrast.

Example 5.6. The last experimental configuration refers to test case (c). As for Example 5.2, we examine the performance of the preconditioner for the bilinear form (2.5). We compare the ASMG preconditioners for three different coefficient

distributions, slices 44, 54, and 74 of the SPE10 benchmark problem. In this example the finest mesh is always composed of 256×256 cells, and changing the number of levels ℓ refers to a different size of the coarse-grid problem (for the coarsest grid). Tables 8–10 report the number of outer iterations n_{ASMG} and the maximum number of inner iterations n_i needed to reduce the residual with the matrix $R\widetilde{D}R^T$ by a factor of 10^6.

TABLE 8
Example 5.6: *Case* (c)*, slice* 44 *of the SPE*10 *benchmark.*

ASMG V-cycle and W-cycle: Bilinear form (2.5)												
	V-cycle						W-cycle					
	$m=0$			$m=1$			$m=0$			$m=1$		
	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i
$\ell=3$	8	0.080	4	7	0.064	5	5	0.019	6	5	0.014	5
$\ell=4$	10	0.157	6	9	0.122	6	5	0.019	6	5	0.014	5
$\ell=5$	12	0.209	6	10	0.154	6	5	0.019	6	5	0.014	5
$\ell=6$	13	0.239	6	11	0.179	6	5	0.019	6	5	0.014	5
$\ell=7$	13	0.239	6	11	0.179	6	5	0.019	6	5	0.014	5

TABLE 9
Example 5.6: *Case* (c)*, slice* 54 *of the SPE*10 *benchmark.*

ASMG V-cycle and W-cycle: Bilinear form (2.5)												
	V-cycle						W-cycle					
	$m=0$			$m=1$			$m=0$			$m=1$		
	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i
$\ell=3$	7	0.070	4	7	0.059	4	5	0.016	4	5	0.013	4
$\ell=4$	10	0.156	5	9	0.122	6	5	0.017	6	5	0.013	5
$\ell=5$	13	0.236	5	11	0.173	6	5	0.018	6	5	0.013	6
$\ell=6$	14	0.253	5	11	0.183	6	5	0.018	6	5	0.013	6
$\ell=7$	14	0.253	6	11	0.183	6	5	0.018	6	5	0.013	6

TABLE 10
Example 5.6: *Case* (c)*, slice* 74 *of the SPE*10 *benchmark.*

ASMG V-cycle and W-cycle: Bilinear form (2.5)												
	V-cycle						W-cycle					
	$m=0$			$m=1$			$m=0$			$m=1$		
	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i	n_{ASMG}	ρ_r	n_i
$\ell=3$	8	0.090	4	7	0.070	4	5	0.019	4	5	0.014	4
$\ell=4$	11	0.178	5	10	0.145	5	5	0.020	5	5	0.015	5
$\ell=5$	13	0.229	5	11	0.166	6	5	0.020	6	5	0.015	6
$\ell=6$	13	0.242	6	11	0.180	6	5	0.020	6	5	0.015	6
$\ell=7$	13	0.242	6	11	0.180	6	5	0.020	6	5	0.015	6

5.3. Testing of block-diagonal preconditioner for system (1.2) within MinRes iteration. Now we present a number of numerical experiments for solving the mixed FE system (1.2) with a preconditioned MinRes method. We consider two different examples: (1) Example 5.7, in which the performance of the block-diagonal preconditioner and its dependence on the accuracy of the inner solves with a W-cycle ASMG preconditioner is evaluated, and (2) Example 5.8, testing the scalability of the MinRes iteration, again using a W-cycle ASMG preconditioner with one smoothing step for the inner iterations.

Example 5.7. Here we apply the MinRes iteration to solve (1.2) for test case (c). The hierarchy of meshes is the same as in Example 5.3. An ASMG W-cycle based on Algorithm 4.1 with one smoothing step has been used as a preconditioner on the $\mathcal{RT}_0$ space. Table 11 shows the number of MinRes iterations denoted by n_{MinRes}, the

TABLE 11
Example 5.7: *Case* (c)*, slice* 44 *of the SPE*10 *benchmark. The hierarchy of meshes is the same as in Example* 5.6.

MinRes iteration: Saddle point system (1.2)						
	$\varpi = 10^6$		$\varpi = 10^8$		$\varpi = 10^{10}$	
	n_{MinRes}	n_{ASMG}	n_{MinRes}	n_{ASMG}	n_{MinRes}	n_{ASMG}
$\ell = 3$	24	4	17	6	15	8
$\ell = 4$	15	5	13	6	13	8
$\ell = 5$	21	5	17	6	15	8
$\ell = 6$	22	5	17	6	15	8
$\ell = 7$	22	5	17	6	15	8

maximum number of ASMG iterations n_{ASMG} needed to achieve an ASMG residual reduction by ϖ.

Example 5.8. Finally, the linear system (1.2) has been solved for test case (c) and the mesh hierarchy of Example 5.1. The right-hand side of (1.1) has been chosen according to (5.1). An ASMG W-cycle with one smoothing step has been used as a preconditioner on the $\mathcal{RT}_0$ space for a residual reduction by a factor 10^8. Table 12 shows the number of MinRes iterations n_{MinRes}, the maximum number of inner ASMG iterations n_{ASMG} per outer MinRes iteration, and the number of DOF. Note that as long as the product $n_{MinRes}n_{ASMG}$ is constant, the total number of arithmetic operations required to achieve any prescribed accuracy is proportional to the number of DOF.

TABLE 12
Example 5.8: *Case* (c)*, slice* 44 *of the SPE*10 *benchmark.*

MinRes iteration: Saddle point system (1.2)					
		zero r.h.s.		nonzero r.h.s.	
	DOF	n_{MinRes}	n_{ASMG}	n_{MinRes}	n_{ASMG}
$\ell = 4$	3136	13	5	13	5
$\ell = 5$	12416	13	6	14	6
$\ell = 6$	49408	15	6	17	6
$\ell = 7$	197120	17	6	17	6
$\ell = 8$	787456	17	6	18	6

5.4. Comments regarding the numerical experiments and some general conclusions. The presented numerical results clearly demonstrate the efficiency of the proposed algebraic multilevel iteration (AMLI)-cycle auxiliary space multigrid (ASMG) preconditioner for problems with highly varying coefficients as they typically arise in the mathematical modelling of physical processes in high-contrast and high-frequency media.

During the first tests, we evaluated the quantity $c_\Pi = \|\pi_{\widetilde{D}}\|^2_{\widetilde{A}}$, which provides an upper bound for the condition number $\kappa(C^{-1}A)$. Then the convergence factor in the A-norm of the linear V-cycle method was numerically studied. The results reported above show robustness with respect to a highly varying coefficient on multiple length scales. They also confirm that the nonlinear V-cycle reduces the residual norm faster than the linear V-cycle.

The next group of tests examines the convergence behavior of the nonlinear ASMG method for the weighted bilinear form (2.5). This is a key point in the presented study. Cases (a) and (b) are designed to represent a typical multiscale geometry with islands and channels. Although case (b), a background with a random coefficient, appears to

be more complicated, the impact of the multiscale heterogeneity seems to be stronger in the binary case (a), where the number of iterations is slightly larger. However, in both cases we observe a uniformly converging ASMG V-cycle with $m = 2$, and a W-cycle ($\nu = 2$) with $m = 1$. Case (c) (SPE10) is a benchmark problem in the petroleum engineering community. Here we observe robust and uniform convergence with respect to the number of levels ℓ, or, equivalently, mesh size h. Note that such uniform convergence is recorded for the ASMG V-cycle even without smoothing iterations (i.e., $m = 0$).

The results in Tables 11 and 12 confirm the expected optimal convergence rate of the block-diagonally preconditioned MinRes iteration applied to the coupled saddle point system (1.2). Table 11 indicates that the highest efficiency (in terms of the product $n_{MinRes}n_{ASMG}$) is achieved for the same relative accuracy (of 10^{-8}) for the inner (ASMG) solver as for the outer (MinRes) solver. Table 12 illustrates the scalability of the method by an almost constant number of MinRes and ASMG iterations since the total computational work in terms of fine-grid matrix vector multiplications is proportional to the product $n_{MinRes}n_{ASMG}$. Tests for nonhomogeneous right-hand side show promising robustness of the ASMG preconditioner beyond the presented theoretical framework.

Although it is beyond the scope of this study, we note that the proposed ASMG method would be suitable for implementation on distributed memory computer architectures.

Acknowledgments. The authors express their sincere thanks to the anonymous reviewers, who made a number of critical remarks and raised relevant questions that resulted in a revised and improved version of the paper which is now shorter, clearer, and more precise in the presentation of the main results. The authors sincerely thank Dr. Shaun Lymbery for his contribution to editing the paper.

REFERENCES

[1] D. N. Arnold, R. S. Falk, and R. Winther, *Preconditioning in H(div) and applications*, Math. Comp., 66 (1997), pp. 957–984.

[2] D. N. Arnold, R. S. Falk, and R. Winther, *Multigrid in H(div) and H(curl)*, Numer. Math., 85 (2000), pp. 197–217.

[3] D. Börm and R. Hiptmair, *Analysis of tensor multigrid*, Numer. Algorithms, 26 (2001), pp. 219–234.

[4] J. Bramble and X. Zhang, *Uniform convergence of the multigrid V-cycle for an anisotropic problem*, Math. Comp., 70 (1998), pp. 453–470.

[5] M. Brezina, P. Vaněk, and P. S. Vassilevski, *An improved convergence analysis of smoothed aggregation algebraic multigrid*, Numer. Linear Algebra Appl., 19 (2012), pp. 441–469.

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.

[7] M. Dryja, M. V. Sarkis, and O. B. Widlund, *Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions*, Numer. Math., 72 (1996), pp. 313–348.

[8] Y. Efendiev, J. Galvis, R. Lazarov, and J. Willems, *Robust domain decomposition preconditioners for abstract symmetric positive definite bilinear forms*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 1175–1199.

[9] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.

[10] J. Galvis and Y. Efendiev, *Domain decomposition preconditioners for multiscale flows in high-contrast media*, Multiscale Model. Simul., 8 (2010), pp. 1461–1483.

[11] I. Georgiev, J. Kraus, and S. Margenov, *Multilevel preconditioning of rotated bilinear nonconforming FEM problems*, Comput. Math. Appl., 55 (2008), pp. 2280–2294.

[12] R. Hiptmair and J. Xu, *Nodal auxiliary space preconditioning in* $\mathbf{H}(\mathbf{curl})$ *and* $\mathbf{H}$(div) *spaces*, SIAM J. Numer. Anal., 45 (2007), pp. 2483–2509.

[13] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.
[14] J. Kraus, *Additive Schur complement approximation and application to multilevel preconditioning*, SIAM J. Sci. Comput., 34 (2012), pp. A2872–A2895.
[15] J. Kraus, R. Lazarov, M. Lymbery, S. Margenov, and L. Zikatanov, *Preconditioning of Weighted $\boldsymbol{H}$(div)-Norm and Applications to Numerical Simulation of Highly Heterogeneous Media*, preprint, arXiv:1406.4455, 2014.
[16] J. Kraus, M. Lymbery, and S. Margenov, *Auxiliary space multigrid method based on additive Schur complement approximation*, Numer. Linear Algebra Appl., 22 (2015), pp. 965–986.
[17] J. Kraus and S. Margenov, *Robust Algebraic Multilevel Methods and Algorithms*, Walter de Gruyter, Berlin, 2009.
[18] J. Kraus, S. Margenov, and J. Synka, *On the multilevel preconditioning of Crouzeix-Raviart elliptic problems*, Numer. Linear Algebra Appl., 15 (2008), pp. 395–416.
[19] J. Kraus, P. Vassilevski, and L. Zikatanov, *Polynomial of best uniform approximation to $1/x$ and smoothing for two-level methods*, Comput. Methods Appl. Math., 12 (2012), pp. 448–468.
[20] W. Krendl, V. Simoncini, and W. Zulehner, *Stability estimates and structural spectral properties of saddle point problems*, Numer. Math., 124 (2013), pp. 183–213.
[21] K.-A. Mardal and R. Winther, *Preconditioning discretizations of systems of partial differential equations*, Numer. Linear Algebra Appl., 18 (2011), pp. 1–40.
[22] C. Pechstein and R. Scheichl, *Analysis of FETI methods for multiscale PDEs. Part* II: *Interface variation*, Numer. Math., 118 (2011), pp. 485–529.
[23] C. Pechstein and R. Scheichl, *Weighted Poincaré inequalities*, IMA J. Numer. Anal., 33 (2013), pp. 652–686.
[24] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.
[25] C. Powell, *Parameter-free H(div) preconditioning for a mixed finite element formulation of diffusion problems*, IMA J. Numer. Anal., 25 (2005), pp. 783–796.
[26] C. E. Powell and D. Silvester, *Optimal preconditioning for Raviart–Thomas mixed formulation of second-order elliptic problems*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 718–738.
[27] R. Scheichl, P. S. Vassilevski, and L. T. Zikatanov, *Multilevel methods for elliptic problems with highly varying coefficients on nonaligned coarse grids*, SIAM J. Numer. Anal., 50 (2012), pp. 1675–1694.
[28] SPE International, *SPE*10 *Comparative Solution Project: General Information and Description of Model* 2, http://www.spe.org/web/csp/#datasets (2000).
[29] P. Vassilevski, *Multilevel Block Factorization Preconditioners: Matrix-Based Analysis and Algorithms for Solving Finite Element Equations*, Springer, New York, 2008.
[30] P. Vassilevski and R. Lazarov, *Preconditioning mixed finite element saddle-point elliptic problems*, Numer. Linear Algebra Appl., 3 (1996), pp. 1–20.
[31] P. Vassilevski and J. Wang, *Multilevel iterative methods for mixed finite element discretizations of elliptic problems*, Numer. Math., 63 (1992), pp. 503–520.
[32] L. Wang, X. Hu, J. Cohen, and J. Xu, *A parallel auxiliary grid algebraic multigrid method for graphic processing units*, SIAM J. Sci. Comput., 35 (2013), pp. C263–C283.
[33] J. Willems, *Robust multilevel methods for general symmetric positive definite operators*, SIAM J. Numer. Anal., 52 (2014), pp. 103–124.
[34] J. Xu and L. Zikatanov, *Some observations on Babuska and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.

INCOMPLETE FACTORIZATION BY LOCAL EXACT FACTORIZATION (ILUE)

Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 145 (2018) 50–61
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

Incomplete factorization by local exact factorization (ILUE)

Johannes Kraus, Maria Lymbery*

Thea-Leymann Str.9, Department of Mathematics, University of Duisburg-Essen, Essen, Germany

Received 8 November 2014; received in revised form 11 October 2017; accepted 11 October 2017
Available online 27 October 2017

Abstract

This study proposes a new preconditioning strategy for symmetric positive (semi-)definite SP(S)D matrices referred to as incomplete factorization by local exact factorization (ILUE). The investigated technique is based on exact LU decomposition of small-sized local matrices associated with a splitting of the domain into overlapping or non-overlapping subdomains. The ILUE preconditioner is defined and its relative condition number estimated. Numerical tests on linear systems arising from the finite element (FE) discretization of a second order elliptic boundary value problem in mixed form demonstrate the advantage of the new algorithm, even for problems with highly oscillatory permeability coefficients, against the classical ILU(p) and ILUT(τ) incomplete factorization preconditioners.

Keywords: Incomplete LU factorization; Local exact factorization; Domain decomposition; Preconditioned Krylov subspace methods

1. Introduction

Consider the linear system of algebraic equations

$$A\,\mathbf{x} \;=\; \mathbf{f} \tag{1}$$

where $\mathbf{f}$ is given and $\mathbf{x}$ is an unknown vector. Further, let A be a sparse $N \times N$ matrix with elements a_{ij}, $i,\, j = 1, \ldots, N$. Such sparse systems commonly result, for example, from finite element (FE) discretizations of (partial) differential or integral equations used in the modeling of various processes in science and engineering.

While in the last years the involvement of software technologies in scientific research has led to new diverse applications of mathematical modeling, the dimensions of the studied discrete systems constantly increase due to the demand for more accurate numerical solutions. The performance enhancement of computer hardware changes the understanding of large dimension, but regardless of progress in this direction, crucial for the development of the mathematical and computer simulations are achievements in the field of numerical methods and algorithms along with their efficient implementation.

* Corresponding author.
E-mail addresses: johannes.kraus@uni-due.de (J. Kraus), maria.lymbery@uni-due.de (M. Lymbery).

https://doi.org/10.1016/j.matcom.2017.10.007

There are two distinct approaches one can apply to solve (1). The first one is by means of using direct methods. Although proven to be very robust and predictable in view of required computational resources, see [14,16], these methods scale unsatisfactorily with the problem size and for this reason iterative solvers are often advantageous over direct methods, especially when considering problems of sufficiently large dimension.

For symmetric positive definite problems the method of conjugate gradients (CG), first proposed in 1952 by Lanczos, [22], and Hestenes and Stiefel, [19], is a method of choice. It is based on the minimization of a quadratic functional over a sequence of Krylov subspaces and its effectiveness can be improved considerably by preconditioning, see, e.g., [5–7,18]. In recent years the development of optimal iterative methods such as the preconditioned CG (PCG) method with a uniform preconditioner that can be applied at the cost of $\mathcal{O}(N)$ arithmetic operations has been an area of active interest and research.

Different strategies can be implemented in order to construct a preconditioner to a given matrix A. One of them is to consider the incomplete LU (ILU) factorization

$$A = LU + E \tag{2}$$

of A where L is unit lower triangular (lower triangular and all entries on the main diagonal are one), U is upper triangular and E is some error matrix, see [26]. Then the matrix B defined by

$$B = LU$$

can be used as a preconditioner when solving system (1) with the PCG method. The classic algorithms for defining (2) are based on adopting fill-in criteria such as position, value, or a combination of the two, see e.g. [8].

Typically, the spectral condition number of the system matrix obtained from FE discretization of second-order, self-adjoint, elliptic partial differential equations (PDE) behaves as $\mathcal{O}(h^{-2})$ where h is the mesh parameter of the underlying partitioning and the number of CG iterations required to achieve a prescribed reduction of the error scales as h^{-1}. Allowing no fill-ins (or small levels of fill) in the incomplete LU factors does not significantly affect the spectral properties of the preconditioned system and the number of CG iterations is proportional to h^{-1} as in the absence of preconditioning only with a much smaller multiplicative constant.

In some special cases, when a technique known as modified incomplete Cholesky (MIC) factorization is applied the condition number of the preconditioned matrix grows as $O(h^{-1})$ and consequently the estimated number of preconditioned CG iterations becomes $O(h^{-1/2})$, see. e.g. [3,15,17,20]. Regardless of any big improvement, however, MIC preconditioners are not so popular as they are more likely to break down on non-model problems and exhibit higher vulnerability to rounding errors, see [29].

In this paper a new technique for constructing an incomplete LU factorization is proposed. It relates to domain decomposition (DD) methods, see e.g. [24,28], from the viewpoint of considering a splitting of the original problem into 'local' subproblems in order to define the LU factors. This algorithm is applicable to a wide range of discrete problems and is easy to implement.

The remainder of the paper is organized as follows. An overview of some popular ILU factorizations is presented in Section 2. The incomplete factorization by local exact factorization (ILUE) subject to the present study is introduced and analyzed in Section 3. In Section 4 a collection of numerical tests on linear systems arising from the FE discretization of a second order elliptic boundary value problem in mixed form is included. Finally, some conclusions are drawn in Section 5.

2. Some classic incomplete LU factorization algorithms

In ILU factorization without fill-in the L and U factors are allowed to have non-zero entries only at positions (i, j) in which the original matrix is non-zero, i.e. $a_{ij} \neq 0$. Such a restriction on the sparsity pattern makes the computation of L and U easy algorithmically and cheap in view of memory requirements, see [25].

The following pseudo-code demonstrates an implementation of the ILU factorization without fill-in. After program execution the entries of U are stored in the modified upper triangular part of A whereas L can be extracted from the strictly lower part and additionally one has to set all diagonal entries of L to be equal to 1.

Algorithm 1. ILU factorization without fill-in

For $k = 1, \ldots, N-1$
  $d = 1/a_{kk}$
  For $i = k+1, \ldots, N$
    If $a_{ik} \neq 0$
      $e = d * a_{ik}$
      $a_{ik} = e$
      For $j = k+1, \ldots, N$
        If $a_{ij} \neq 0$ (∗)
          $a_{ij} = a_{ij} - e * a_{kj}$

Although ILU factorization without fill-in may approximate A well in some particular cases, such as low-order discretizations of scalar elliptic PDE, this technique is not very accurate and therefore may not provide satisfactory preconditioners in other cases, such as those considered in this study.

One remedy for this is to allow additional p levels of fill-in in L and U. This idea of improving the accuracy of the ILU factorization, first presented in [17] and further developed and generalized in [30], is known as ILU(p) factorization.

The pseudo-code presented below shows an algorithm realizing the ILU(p) factorization. According to [26] the matrix storing the level of fill-in entries is initialized by

$$level(a_{ij}) = \begin{cases} 0 & \text{if} \quad a_{ij} \neq 0 \quad \text{or} \quad i = j, \\ \infty & \text{otherwise.} \end{cases}$$

Through the factorization procedure, the levels of the processed elements are constantly updated and whenever $level(a_{ij})$ becomes bigger than p the entry at position (i, j) is dropped. Finally, the entries of L and U overwrite the matrix in the same manner as in Algorithm 1.

Note that the case $p = 0$ results in ILU factorization without fill-in as before which is also commonly referred to as ILU(0).

Algorithm 2. ILU(p) ILUfactorization

For $i, j = 1, \ldots, N$
  If $a_{i,j} \neq 0$ **or** $i = j$
    $level(a_{ij}) = 0$
  Else
    $level(a_{ij}) = p + 1$
For $i = 2, \ldots, N$
  For $k = 1, \ldots, i-1$
    $a_{ik} = a_{ik}/a_{kk}$
    For $j = k+1, \ldots, N$
      $a_{ij} = a_{ij} - a_{ik} * a_{kj}$
      $level(a_{ij}) = \min(level(a_{ij}), level(a_{ik}) + level(a_{kj}) + 1)$
      If $level(a_{ij}) > p$
        $a_{ij} = 0$

The ILU(p) factorization is especially suitable for preconditioning diagonally dominant matrices since, in this case, the larger the level of fill-in, the smaller the fill-in tends to be in absolute value. However, if a matrix is far from being diagonally dominant, increasing p in the ILU(p) algorithm contributes little to improving the quality of the preconditioner as many elements that are small in absolute value have to be stored which further makes this approach expensive to use. Therefore, a different strategy is required to improve the accuracy of the ILU approximation.

Such a strategy can be to integrate a dropping criteria for the computed entries in L and U based on their value. A new fill-in is accepted only if it is greater than τ times a certain (scaled) norm of the current row where $\tau > 0$ is called the *dropping tolerance* or *threshold*.

This algorithm is known as ILUT(τ) factorization and can be realized via the pseudo-code given in Algorithm 3. Here $\|\mathbf{v}\|_1 := \sum_{i=1}^{N} |v_i|$ denotes the ℓ^1-norm of a vector $\mathbf{v} \in \mathbb{R}^N$ and nnz($\mathbf{v}$) its number of nonzero entries.

A drawback to this technique is the impossibility of predicting the number of non-zero elements created during the factorization process. For this reason one sometimes imposes an additional restriction on the number of entries in a row, that is, at most the p largest ones in magnitude to be finally stored in the ILUT preconditioner, see [27].

Algorithm 3. ILUT(τ) factorization

```
For i = 2, ..., N
    τ_i = τ * ||a_{i,1:N}||_1 / nnz(a_{i,1:N})
    For k = 1, ..., i − 1
        If a_ik ≠ 0
            a_ik = a_ik / a_kk
            For j = k + 1, ..., N
                a_ij = a_ij − a_ik * a_kj
                If |a_ij| < τ_i
                    a_ij = 0
```

Remark 1. In the presented algorithms it has been assumed that no breakdowns occur due to divisions by zero, which is guaranteed, for example, for M-matrices. In many cases, however, one can avoid such instances also for non-M matrices by applying proper pivoting strategies.

Remark 2. There are also block variants of the described incomplete factorizations known as BILU, for more details see e.g. [4,11–13].

3. Incomplete factorization by local exact factorization (ILUE)

The starting point of the ILUE factorization discussed in this paper is that A can be presented in the form

$$A = \sum_{k=1}^{n} R_k^T A_k R_k \tag{3}$$

where $A_k, k = 1, \ldots, n$, are small-sized (local) symmetric positive (semi-) definite (SP(S)D) subdomain matrices and R_k^T are the natural inclusions that extend any local vector defined on the degrees of freedom (DOF) of Ω_k by zeros to a global vector defined on the set of all DOF (of Ω). That is, $R_k = (r_{k,ij}) \in \mathbb{R}^{N_k \times N}$ is a {0, 1}-matrix with one nonzero entry, equal to one, in every row. The position (i, j) of each nonzero entry in R_k determines the global number, which equals the column index j, corresponding to its local number (within Ω_k), which equals the row index i. It follows that $R_k R_k^T = I_k$.

Consider now the exact LU factorizations of the matrices $A_k, k = 1, \ldots, n$

$$A_k = L_k U_k. \tag{4}$$

Definition 4. Incomplete factorization by local exact factorization (ILUE) is defined as

$$B_{\mathrm{ILUE}} := LU \tag{5}$$

where

$$U := \sum_{k=1}^{n} R_k^T U_k R_k, \qquad L := U^T \mathrm{diag}(U)^{-1}, \qquad \mathrm{diag}(L_k) = I_k \tag{6}$$

and the matrices L_k, U_k and R_k are as introduced in (3)–(4).

If one defines the auxiliary block diagonal matrix

$$\widetilde{A} = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_n \end{bmatrix} = \begin{bmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_n \end{bmatrix} \begin{bmatrix} U_1 & & & \\ & U_2 & & \\ & & \ddots & \\ & & & U_n \end{bmatrix}$$

$$= \widetilde{L}\widetilde{U}$$

then

$$A = R\widetilde{A}R^T,$$

where the matrix R is of size $N \times \widetilde{N} = \dim(A) \times \dim(\widetilde{A})$ is defined by $R = \left[R_1^T, R_2^T, \ldots, R_n^T\right]$. Then Eq. (6) implies that $U = R\widetilde{U}R^T$ and hence the construction of U is analogous to that of the approximate inverse B_{AS}^{-1} obtained from the one-level additive Schwarz method, which is given by $B_{\mathrm{AS}}^{-1} := \sum_{k=1}^{n} R_k^T A_k^{-1} R_k$, relating the ILUE factorization to domain decomposition methods.

Now let $N_k = \dim(A_k)$ and $N = \dim(A)$. Further, let $num[k, i]$ denote the column index of the nonzero entry in row i of R_k and $u_{k,ij}$ the entry in position (i, j) of U_k for $1 \le k \le n$ and $1 \le i, j \le N_k$. Then Algorithm 4. given below formally describes the implementation of the ILUE factorization.

Remark 3. To compute U_k one performs a complete LU factorization of the matrix A_k subject to the condition $\mathrm{diag}(L_k) = I$. For that purpose one can use Algorithm 1 thereby skipping line $(*)$. In the case of processing symmetric positive semidefinite matrices A_k, the occurrence of zero pivot elements, which would cause a breakdown of the algorithm due to division by zero, requires special treatment. Simply skipping all operations involving zero pivot elements is one possible strategy.

Remark 4. In the important case in which A is SPD and all the A_k's are SPSD with some (or maybe most) of them being singular, one has to use proper local numberings of the degrees of freedom (DOF) within the subdomains and local-to-global mappings to ensure that U and hence $\mathrm{diag}(U)$ and L will turn out to be regular. This, however, will always be possible if the subdomains have sufficiently large overlaps, i.e., share sufficiently many DOF.

Algorithm 4. ILUE factorization

For $k = 1, \ldots, n$
  factorize $A_k = L_k U_k$ subject to $\mathrm{diag}(U_k) = I_k$
  For $i = 1, \ldots, N_k$
    For $j = 1, \ldots, N_k$
      $u_{num[k,i]num[k,j]} += u_{k,ij}$
For $i = 1, \ldots, N$
  For $j = 1, \ldots, N$
    $l_{ij} = u_{ji}/u_{jj}$

The next lemma is useful in the analysis of the spectral condition number of the preconditioned matrix $(B_{\mathrm{ILUE}}^{-1}A)$.

Lemma 5. *Let X_i and Y_i, $i = 1, \ldots, n$, be real matrices of size $k \times l$ and $k \times m$ respectively and*

$$Z_{11} := \sum_{i=1}^{n} Y_i^T Y_i. \tag{7}$$

Then the following relation holds true

$$\sum_{i=1}^{n} X_i^T X_i - \sum_{i=1}^{n} X_i^T Y_i \left(\sum_{i=1}^{n} Y_i^T Y_i \right)^{-1} \sum_{i=1}^{n} Y_i^T X_i \ge 0. \tag{8}$$

Proof. Define

$$Z := \begin{bmatrix} \sum_{i=1}^{n} Y_i^T Y_i & \sum_{i=1}^{n} Y_i^T X_i \\ \sum_{i=1}^{n} X_i^T Y_i & \sum_{i=1}^{n} X_i^T X_i \end{bmatrix}.$$

Since

$$\begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}^T \begin{bmatrix} Y_i^T Y_i & Y_i^T X_i \\ X_i^T Y_i & X_i^T X_i \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \|Y_i \mathbf{v}_1 + X_i \mathbf{v}_2\|^2 \geq 0$$

$\forall\, \mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2]^T$ and $\forall\, i$, it follows that Z is an SPSD matrix and therefore

$$S_Z := \sum_{i=1}^{n} X_i^T X_i - \sum_{i=1}^{n} X_i^T Y_i \left(\sum_{i=1}^{n} Y_i^T Y_i \right)^{-1} \sum_{i=1}^{n} Y_i^T X_i \geq 0. \quad \square$$

Remark 5. When $\sum_{i=1}^{n} Y_i^T Y_i$ is a singular (SPSD) matrix its inverse can be understood as the generalized Moore–Penrose inverse. Note that like this the extremal property of the Schur complement is still valid.

Remark 6. Let $k = l = m = 1$ for $i = 1, \ldots, n$. Then X_i and Y_i are real numbers and consequently inequality (8) reduces to the classic discrete Cauchy–Schwarz inequality

$$\sum_{i=1}^{N} X_i^2 \sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} X_i Y_i \right)^2 \geq 0.$$

If the matrices X_i and Y_i are defined as

$$X_i := R_i^T \operatorname{diag}(U_i)^{-1/2} U_i R_i \qquad \text{and} \qquad Y_i := R_i^T \operatorname{diag}(U_i)^{1/2} R_i,$$

where R_i and U_i are defined according to (3) and (4) then $\sum_{i=1}^{N} Y_i^T Y_i$ is an invertible matrix (if U is regular) and Lemma 5 holds, i.e.,

$$\begin{aligned} &\sum_{i=1}^{N} X_i^T X_i - \sum_{i=1}^{N} X_i^T Y_i \left(\sum_{i=1}^{N} Y_i^T Y_i \right)^{-1} \sum_{i=1}^{N} Y_i^T X_i \\ &= \sum_{i=1}^{N} R_i^T U_i^T \operatorname{diag}(U_i)^{-1/2} R_i R_i^T \operatorname{diag}(U_i)^{-1/2} U_i R_i \\ &- \sum_{i=1}^{N} R_i^T U_i^T \operatorname{diag}(U_i)^{-1/2} R_i R_i^T \operatorname{diag}(U_i)^{1/2} R_i \\ &\times \Big(\sum_{i=1}^{N} R_i^T \operatorname{diag}(U_i)^{1/2} R_i R_i^T \operatorname{diag}(U_i)^{1/2} R_i \Big)^{-1} \\ &\times \sum_{i=1}^{N} R_i^T \operatorname{diag}(U_i)^{1/2} R_i R_i^T \operatorname{diag}(U_i)^{-1/2} U_i R_i \geq 0. \end{aligned}$$

Furthermore, from

$$A_i = U_i^T \operatorname{diag}(U_i)^{-1} U_i \qquad \text{and} \qquad R_i R_i^T = I_i$$

it follows that the last inequality is equivalent to

$$A - B_{\mathrm{ILUE}} \geq 0 \qquad \text{or} \qquad \mathbf{w}^T B_{\mathrm{ILUE}} \mathbf{w} \leq \mathbf{w}^T A \mathbf{w} \quad \forall \mathbf{w}. \tag{9}$$

On the other hand, to find an estimate of the form

$$\mathbf{w}^T A \mathbf{w} \leq \bar{c}\, \mathbf{w}^T B_{\mathrm{ILUE}} \mathbf{w} \quad \forall \mathbf{w}, \tag{10}$$

which equivalently reads

$$\mathbf{w}^T\Big(\sum_{i=1}^{n} R_i^T A_i R_i\Big)\mathbf{w} \le \bar{c}\,\mathbf{w}^T LU\mathbf{w} = \bar{c}\,\mathbf{w}^T U^T \mathrm{diag}(U)^{-1}U, \mathbf{w} \quad \forall \mathbf{w},$$

$\lambda_{\max}$ is defined as

$$\lambda_{\max} := \max_{1\le i\le n}\{\lambda_{i,\max}\} \tag{11}$$

where $\lambda_{i,\max}$ are the maximum eigenvalues of the corresponding low-rank generalized eigenvalue problems

$$R_i^T A_i R_i \mathbf{v} = \lambda_i U^T \mathrm{diag}(U)^{-1} U\mathbf{v}. \tag{12}$$

Then the number $c := \lambda_{\max} n_{\text{subdomain}}$ provides an upper bound for $\bar{c}$, i.e., (10) is satisfied with $\bar{c} = c$. Here $n_{\text{subdomain}}$ denotes the number of subdomains.

The generalized eigenvalue problems (12), although related to the local matrices A_i, are of the size of the original problem. However, the matrices $R_i^T A_i R_i$ are of low rank, which is bounded by $N_i = \dim(A_i)$. Using the decomposition $A_i = U_i^T \mathrm{diag}(U_i)^{-1/2}\mathrm{diag}(U_i)^{-1/2}U_i$ problem (12) can be rewritten as[1]

$$\mathrm{diag}(U)^{1/2}U^{-T}R_i^T U_i^T \mathrm{diag}(U_i)^{-1/2}\mathrm{diag}(U_i)^{-1/2}U_i R_i U^{-1}\mathrm{diag}(U)^{1/2}\mathbf{v} = \lambda_i \mathbf{v},$$

that is, in the form $W_i^T W_i \mathbf{v} = \lambda_i \mathbf{v}$ with $W_i = \mathrm{diag}(U_i)^{-1/2}U_i R_i U^{-1}\mathrm{diag}(U)^{1/2}$ where $W_i \in \mathbb{R}^{N_i\times N}$. In order to find the nonzero eigenvalues of (12) one can equivalently solve the eigenvalue problem $W_i W_i^T \mathbf{v}_i = \lambda_i \mathbf{v}_i$. The latter is of the size of the local matrices A_i and of the form

$$\mathrm{diag}(U_i)^{-1/2}U_i R_i U^{-1}\mathrm{diag}(U)U^{-T}R_i^T U_i^T \mathrm{diag}(U_i)^{-1/2}\mathbf{v}_i = \lambda_i \mathbf{v}_i. \tag{13}$$

Finally, since $n_{\text{subdomain}}$ is typically of the order $\mathcal{O}(h^{-2})$ for two-dimensional problems, it can be concluded that $\kappa(B_{\text{ILUE}}^{-1}A) = \lambda_{\max}\mathcal{O}(h^{-2})$.

For a d-dimensional unit cube which is partitioned by a uniform mesh of mesh size h into $(1/h)^d$ cubes with side h the number of congruent subdomains is given by

$$n_{\text{subdomain}} = \left(\frac{1-w}{H-w}\right)^d$$

where H is the size (side) of the subdomains and w the width of the overlap. Assuming that H and w are both multiples of h, i.e., $H = kh$ and $w = mh$, an obvious requirement for $n_{\text{subdomain}}$ to be bounded is that $m < k$. It is even reasonable to assume that $m \le k/2$ resulting in

$$n_{\text{subdomain}} = \left(\frac{1}{H/2}\right)^d = \mathcal{O}(H^{-d}).$$

The resulting condition number estimate in this case is $\lambda_{\max}\mathcal{O}(H^{-d})$. Note that in practice H can be significantly larger than h. To give an example, for $h = 1/256$ and $H = 8h$, as used in the numerical tests in the next section, one has $H^{-3} = h^{-2}/2$.

4. Numerical results

4.1. Model problem

The presented numerical tests refer to the second order elliptic boundary value problem in mixed form

$$\boldsymbol{u} + K(x)\nabla p = 0 \qquad \text{in } \Omega, \tag{14a}$$
$$\mathrm{div}\,\boldsymbol{u} = f \qquad \text{in } \Omega, \tag{14b}$$
$$p = 0 \qquad \text{on } \Gamma_D\,, \tag{14c}$$
$$\boldsymbol{u}\cdot\boldsymbol{n} = 0 \qquad \text{on } \Gamma_N. \tag{14d}$$

[1] Again inverses eventually have to be understood in the sense of the Moore–Penrose inverse.

Here Ω is a bounded polyhedral domain in $\mathbb{R}^2$ and $\boldsymbol{n}$ is the outward unit vector normal to the boundary $\partial\Omega$ that is split into two non-overlapping parts Γ_D and Γ_N. In order to guarantee that problem (14) has a unique solution $p \in H^1(\Omega)$ it is assumed that Γ_D is a non-empty set with strictly positive measure which is also closed with respect to $\partial\Omega$. The given forcing term f in (14b) is a function from $L^2(\Omega)$ whereas the coefficient $K \in \mathbb{R}^{2\times 2}_{SPD}$ for almost all $x \in \Omega$. In what follows we consider $K(x) = k(x)I$ where $k(x) > 0\ \forall x \in \Omega$.

The considered model problem is used to describe different processes, for example, diffusion of passive chemicals, heat and mass transfer, or electrostatics. In the terminology of flows in porous media the unknown variable $\boldsymbol{u}$ is called the *velocity* whereas p is referred to as the *fluid pressure*.

In order to discretize (14) a mixed FEM is applied. Due to the low regularity of the problem in the case of heterogeneous media (highly varying permeability $k(x)$) lowest order Raviart–Thomas–Nédélec functions are chosen for the discretization of the vector valued function $\boldsymbol{u}$ and piecewise constant functions are chosen for the unknown pressure variable. As a result the following saddle point system is derived

$$\begin{bmatrix} M_\alpha & B_{\mathrm{div}} \\ B_{\mathrm{div}}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f} \end{bmatrix}. \tag{15}$$

The matrices in (15) are determined by

$$\begin{aligned} \mathbf{v}^T M_\alpha \mathbf{u} &= (\alpha \boldsymbol{u}_h, \boldsymbol{v}_h), \\ \mathbf{v}^T B_{\mathrm{div}}^T \mathbf{p} &= (p_h, \operatorname{div} \boldsymbol{v}_h), \end{aligned}$$

where $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$ are lowest order Raviart–Thomas–Nédélec FE functions, p_h is piecewise constant and

$$\alpha = \alpha(x) = k^{-1}(x). \tag{16}$$

The stability of the considered mixed FE approximation of problem (14) has been proven in [21] from where it follows that the Arnold–Falk–Winther preconditioner, see [1,2],

$$\mathcal{B}_h = \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} \tag{17}$$

is uniform. Here I denotes the (matrix representation of the) Riesz isomorphism defined by the L^2 inner product while A is the operator (matrix) defined by the weighted $\boldsymbol{H}(\mathrm{div})$ bilinear form, i.e.,

$$\mathbf{u}^T A \mathbf{v} = (\alpha\, \boldsymbol{u}_h, \boldsymbol{v}_h) + (\nabla \cdot \boldsymbol{u}_h, \nabla \cdot \boldsymbol{v}_h) = \Lambda_\alpha(\boldsymbol{u}_h, \boldsymbol{v}_h). \tag{18}$$

Therefore, the application of the preconditioner (17) requires a preconditioner for the system

$$A\mathbf{u} = \mathbf{b}, \qquad \mathbf{u}, \mathbf{b} \in \mathbb{R}^N, \tag{19}$$

which is subject to the numerical study in this section. For the considered FE discretization N is the number of edges in the underlying partitioning of the domain, excluding edges on Γ_N.

When incomplete factorization preconditioners are considered the ordering of the DOF in the system can significantly affect their performance, see, e.g. [9,10,23,31] where PCG convergence has been mainly experimentally investigated. We test a lexicographical and a two-level ordering; the latter is combined with a two-level basis transformation as commonly used in the context of two- and multilevel methods.

4.2. Description of the parameters

The numerical experiments are executed over a uniform mesh composed of $\ell \times \ell$ elements (squares) where $\ell = 8, \ldots, 256$, i.e. up to 131 584 velocity DOF. An overlapping covering of the domain as shown in Fig. 1 is considered where the subdomains are composed of 8×8 elements and overlap with half of their width or height.

A random vector is chosen as an initial guess for the solution of (19) and the PCG algorithm is tested on problems with a zero right-hand side and a stopping criterion of residual reduction by a factor of 10^6 or exceeding 5000 PCG iterations.

The numerical results in Tables 1–7 are obtained for problems exhibiting high oscillations in the coefficient $\alpha(x) = K^{-1}(x)$, namely a randomly distributed coefficient $\alpha = 10^{-p_{\mathrm{rand}}}$, $p_{\mathrm{rand}} \in \{0, 1, 2, \ldots, q\}$ where α is constant for any given element.

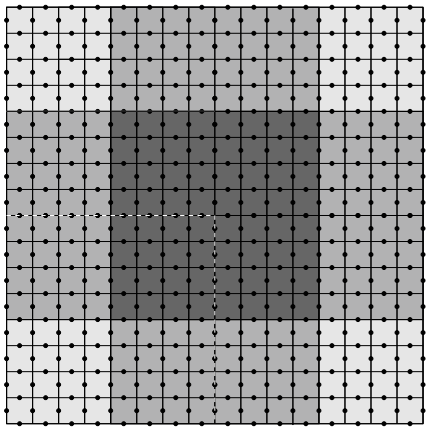


Fig. 1. Covering of a domain of 16×16 elements by nine overlapping subdomains.

Table 1
Values of the parameter τ for the ILUT preconditioner.

q	$\ell = 8$	$\ell = 16$	$\ell = 32$	$\ell = 64$	$\ell = 128$	$\ell = 256$
0	10^{-16}	10^{-16}	$3.2 * 10^{-3}$	$4.5 * 10^{-3}$	$3.3 * 10^{-3}$	$2.0 * 10^{-3}$
1	10^{-16}	10^{-16}	$1.8 * 10^{-3}$	$2.7 * 10^{-3}$	$2.0 * 10^{-3}$	$1.2 * 10^{-3}$
2	10^{-16}	10^{-16}	$1.2 * 10^{-3}$	$1.8 * 10^{-3}$	$1.4 * 10^{-3}$	$8.5 * 10^{-4}$
3	10^{-16}	10^{-16}	$9.1 * 10^{-4}$	$1.3 * 10^{-3}$	$1.1 * 10^{-3}$	$6.5 * 10^{-4}$
4	10^{-16}	10^{-16}	$8.1 * 10^{-4}$	$1.1 * 10^{-3}$	$8.3 * 10^{-4}$	$5.3 * 10^{-4}$
5	10^{-16}	10^{-16}	$7.8 * 10^{-4}$	$9.5 * 10^{-4}$	$7.1 * 10^{-4}$	$4.3 * 10^{-4}$
6	10^{-16}	10^{-16}	$3.8 * 10^{-4}$	$8.4 * 10^{-4}$	$6.1 * 10^{-4}$	$3.7 * 10^{-4}$

Table 2
Approximate number of non-zero entries in L for the ILU(4), ILUT and ILUE.

q	L	$\ell = 8$	$\ell = 16$	$\ell = 32$	$\ell = 64$	$\ell = 128$	$\ell = 256$
0–6	$L_{\mathrm{ILU(4)}}$	$1.4 * 10^4$	$6.5 * 10^4$	$2.7 * 10^5$	$1.1 * 10^6$	$4.5 * 10^6$	$1.8 * 10^7$
	L_{ILUT}	$1.7 * 10^4$	$1.2 * 10^5$	$6.8 * 10^5$	$2.8 * 10^6$	$1.1 * 10^7$	$4.6 * 10^7$
	L_{ILUE}	$1.7 * 10^4$	$7.9 * 10^4$	$3.4 * 10^5$	$1.4 * 10^6$	$5.7 * 10^6$	$2.3 * 10^7$

Table 3
Number of PCG iterations for residual reduction by 10^6, standard basis.

ILU(4) preconditioner							
	Contrast						
	0	1	2	3	4	5	6
$\ell = 8$	64	62	85	115	115	126	166
$\ell = 16$–256	>5000	>5000	>5000	>5000	>5000	>5000	>5000

The performance of the ILU(p) and ILUT(τ) preconditioners is tested for $p = 4$ and τ as in Table 1. In choosing these parameters as in Table 2 the number of non-zeros produced in the corresponding L factors is comparable to the number of non-zero entries in the studied ILUE factorization (for lexicographical ordering), i.e.

$$\frac{\mathrm{nnz}(L_{\mathrm{ILUE}})}{2} \leq \mathrm{nnz}(L_{\mathrm{ILU}(p)}),\ \mathrm{nnz}(L_{\mathrm{ILUT}}) \leq 2 * \mathrm{nnz}(L_{\mathrm{ILUE}}).$$

Note that the choice of the parameter τ in Table 1 for $\ell = 32, 64, 128, 256$ is already quite unfavorable for comparing our algorithm against ILUT since it results in $\mathrm{nnz}(L_{\mathrm{ILUT}}) \approx 2 * \mathrm{nnz}(L_{\mathrm{ILUE}})$.

As can be seen from Tables 3 and 5 the number of PCG iterations with the ILUE preconditioner is always less than that with the ILU(4) preconditioner. The ILUT preconditioner would appear to be robust with respect to the contrast for smaller-size problems, see Table 4 for $\ell = 8, 16$, whereas its performance deteriorates when more refined meshes

Table 4
Number of PCG iterations for residual reduction by 10^6, standard basis.

ILUT preconditioner							
	Contrast						
	0	1	2	3	4	5	6
$\ell = 8$	1	1	1	1	1	1	1
$\ell = 16$	1	1	1	1	1	1	1
$\ell = 32$–256	>5000	>5000	>5000	>5000	>5000	>5000	>5000

Table 5
Number of PCG iterations for residual reduction by 10^6, standard basis.

ILUE preconditioner							
	Contrast						
	0	1	2	3	4	5	6
$\ell = 8$	1	1	1	1	1	1	1
$\ell = 16$	17	21	24	26	31	37	42
$\ell = 32$	35	50	67	82	111	164	213
$\ell = 64$	66	90	130	210	347	577	999
$\ell = 128$	125	159	228	355	635	1135	2086
$\ell = 256$	220	260	350	541	903	1682	3142

are considered. However, as can be seen from Table 1 the choice of τ for $\ell = 8$ and $\ell = 16$ actually gives the exact factorization of A thereby explaining the observed robust behavior. It should also be mentioned that choosing a proper parameter τ can often be quite difficult.

Table 5 furthermore shows that the number of PCG iterations with ILUE preconditioning grows moderately for increasing contrast whereas for a fixed contrast it indicates a growth of the relative condition number as predicted by our estimate, i.e., $\kappa(B_{\mathrm{ILUE}}^{-1}A) = \mathcal{O}(H^{-2})$.

Table 3 suggests that ILU(p) factorization is a bad choice for the preconditioning of system (19). The results in Table 6 for a fixed contrast $q = 2$ and bigger levels of fill-in confirm this observation and show that allowing more entries is not a remedy. Note that $\mathrm{nnz}(L_{\mathrm{ILU}(p)})$ is always approximately or exactly the number of non-zeros in the corresponding L factors of the exact factorization whenever the number of PCG iterations is not >5000.

To test the effect of changing into a two-level hierarchical basis, A is transformed according to

$$\widehat{A} = J^T A J, \qquad J = P J_{\pm}, \tag{20}$$

where the permutation P splits the DOF into two groups. The first one contains all DOF associated with fine-grid edges not part of any coarse-grid edge (interior DOF) whereas the second contains the remaining DOF, both groups numbered consecutively along the grid lines. The transformation matrix $J_{\pm}$ has the form

$$J_{\pm} = \begin{bmatrix} I & \\ & J_{22} \end{bmatrix}, \quad \text{where} \quad J_{22} = \frac{1}{2}\begin{bmatrix} 1 & -1 & & & & & \\ & & 1 & -1 & & & \\ & & & & \ddots & \ddots & \\ & & & & & & 1 & -1 \\ 1 & 1 & & & & & \\ & & 1 & 1 & & & \\ & & & & \ddots & \ddots & \\ & & & & & & 1 & 1 \end{bmatrix}$$

and the identity matrix I corresponds to the interior DOF. In order to compute the ILUE preconditioner in a changed basis the analogous transformation is performed also at a local level, i.e., on each subdomain Ω_i.

Table 7 demonstrates an even better performance of the ILUE preconditioner for a two-level basis and shows only a mild dependence on the contrast.

Table 6
Number of PCG iterations for residual reduction by 10^6, standard basis.

Fixed contrast $q = 2$				
	ILU(4)	ILU(12)	ILU(29)	ILU(62)
$\ell = 8$	85	1	1	1
$\ell = 16$	>5000	185	1	1
$\ell = 32$	>5000	>5000	125	1
$\ell = 64$	>5000	>5000	>5000	35
$\ell = 128, 256$	>5000	>5000	>5000	>5000

Table 7
Number of PCG iterations for residual reduction by 10^6, two-level basis.

ILUE preconditioner							
	Contrast						
	0	1	2	3	4	5	6
$\ell = 8$	1	1	1	1	1	1	1
$\ell = 16$	12	12	13	14	14	15	18
$\ell = 32$	25	27	31	34	41	52	69
$\ell = 64$	44	50	59	82	115	163	256
$\ell = 128$	79	84	97	123	169	254	391
$\ell = 256$	127	122	128	151	194	273	427

Table 8
Estimation of the relative condition number.

	$\ell = 8$	$\ell = 16$	$\ell = 32$	$\ell = 64$
$\kappa(B_{\mathrm{ILUE}}^{-1}A)$	6.97	36.99	172.88	750.29
$\lambda_{\max} \cdot n_{\mathrm{subdomain}}$	$9 \cdot 4.21$	$49 \cdot 13.42$	$225 \cdot 32.20$	$961 \cdot 73.88$

Finally, the true condition number $\kappa(B_{\mathrm{ILUE}}^{-1}A)$ is compared against its upper bound $\lambda_{\max} n_{\mathrm{subdomain}}$, where $\lambda_{\max}$ is defined as in (11) with $\lambda_{i,\max}$ computed from (13), for a problem with constant coefficient ($\alpha = 1$) and subdomains of size 4×4 elements. As can be seen, the estimate is highly conservative. (See Table 8.)

5. Concluding remarks

A new type of preconditioner called ILUE from the class of approximate factorization methods has been constructed and analyzed. It utilizes exact LU factorizations of local (small-sized) matrices associated with a splitting of the domain into overlapping or non-overlapping subdomains. The principal design of the algorithm makes the ILUE preconditioner easy to implement and leads to a simple condition number estimate. The performed numerical results demonstrate monotone convergence for a class of problems where Krylov methods with classic incomplete factorization preconditioning suffer from stagnation.

Acknowledgments

This work has been partly supported by the Austrian Science Fund, Grant P22989, and the Project AComIn, Grant 316087, funded by the FP7 Capacity Programme.

The contribution of the authors of the open source program ITSOL should be acknowledged since their code allowed an independent assessment and verification of the performed numerical tests contained in this paper, http://www-users.cs.umn.edu/~saad/software/ITSOL/index.html.

References

[1] D. Arnold, R. Falk, R. Winther, Preconditioning in H(div) and applications, Math. Comp. 66 (1997) 957–984.
[2] D. Arnold, R. Falk, R. Winther, Multigrid in H(div) and H(curl), Numer. Math. 85 (2000) 197–217.

[3] O. Axelsson, A generalized SSOR method, BIT 12 (4) (1972) 443–467.
[4] O. Axelsson, A general incomplete block matrix factorization method, Linear Algebra Appl. 74 (1986) 179–190.
[5] O. Axelsson, I. Gustafsson, Preconditioning and two-level multigrid methods of arbitrary degree of approximations, Math. Comp. 40 (1983) 219–242.
[6] O. Axelsson, P. Vassilevski, Algebraic multilevel preconditioning methods I, Numer. Math. 56 (1989) 157–177.
[7] O. Axelsson, P. Vassilevski, Algebraic multilevel preconditioning methods II, SIAM J. Numer. Anal. 27 (1990) 1569–1590.
[8] M. Benzi, Preconditioning techniques for large linear systems: a survey, J. Comput. Phys. 182 (2002) 418–477.
[9] M. Benzi, J.C. Haws, M. Tuma, Preconditioning highly indefinite and nonsymmetric matrices, SIAM J. Sci. Comput. 22 (4) (2000) 1333–1353.
[10] M. Benzi, D.B. Szyld, A. van Duin, Orderings for incomplete factorisation preconditioning of nonsymmetric problems, SIAM J. Sci. Comput. 20 (5) (1999) 1652–1670.
[11] A. Chapman, Y. Saad, L. Wigton, High-order ILU preconditioners for CFD problems, Internat. J. Numer. Methods Fluids 33 (6) (2000) 767–788.
[12] E. Chow, M.A. Heroux, An object-oriented framework for block preconditioning, ACM Trans. Math. Software 24 (2) (1998) 159–183.
[13] E. Chow, Y. Saad, Approximate inverse techniques for block-partioned matrices, SIAM J. Sci. Comput. 18 (6) (1997) 1657–1675.
[14] I.S. Duff, A.M. Erisman, J.K. Reid, Direct Methods for Sparse Matrices, Oxford University Press, Inc, New York, 1986.
[15] T. Dupont, R.P. Kendall, H.H. Rachford, An approximate factorization procedure for solving self-adjoint elliptic difference equations, SIAM J. Numer. Anal. 5 (3) (1968) 559–573.
[16] A. George, J.W. Liu, Computer Solution of Large Sparse Positive Definite, Prentice Hall Professional Technical Reference, 1981.
[17] I. Gustafsson, A class of first order factorization methods, BIT 18 (2) (1978) 142–156.
[18] W. Hackbusch, Iterative Solution of Large Sparse Systems of Equations, Springer, New York, 1993.
[19] M.R. Hestenes, E.L. Stiefel, Methods of conjugate gradients for solving linear systems, J. Res. Natl. Bur. Stand. 49 (1952) 409–436.
[20] V.P. Il'in, Iterative Incomplete Factorization Methods, World Scientific, Singapore, 1992.
[21] J. Kraus, R. Lazarov, M. Lymbery, S. Margenov, L. Zikatanov, Preconditioning heterogeneous $H(\mathrm{div})$ problems by additive Schur complement approximation and applications, SIAM J. Sci. Comput. 38 (2) (2016) A875–A898.
[22] C. Lanczos, Solution of systems of linear equations by minimized iterations, J. Res. Natl. Bur. Stand. 49 (1952) 33–53.
[23] S. Le Borne, Ordering techniques for two- and three-dimensional convection-dominated elliptic boundary value problems, Computing 64 (2) (2000) 123–155.
[24] T.P.A. Mathew, Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations, in: LNCSE, vol. 61, Springer, Berlin Heidelberg, 2008.
[25] J.A. Meijerink, H.A. van der Vorst, An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, Math. Comp. 31 (1977) 148–162.
[26] Y. Saad, Iterative Methods for Sparse Linear Systems, second ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003.
[27] Y. Saad, ILUT: A dual threshold incomplete LU factorization, Numer. Linear Algebra Appl. 1 (4) (2005) 387–402.
[28] A. Toselli, O. Widlund, Domain Decomposition Methods-Algorithms and Theory, in: Springer Series in Computational Mathematics, vol. 34, Springer, Verlag Berlin Heidelberg, 2005.
[29] H.A. van der Vorst, The convergence behavior of preconditioned CG and CGS in the presence of rounding errors, in: Preconditioned Conjugate Gradient Methods, Vol. 1457, 1990, pp. 126–136.
[30] I.J. Watts, Conjugate gradient-truncated direct method for the iter-ative solution of the reservoir simulation pressure equation, Soc. Pet. Eng. J. 21 (1981) 345–353.
[31] J. Zhang, A multilevel dual reordering strategy for robust incomplete lu factorisation of indefinite matrices, SIAM J. Matrix Anal. Appl. 22 (3) (2001) 925–947.

CURRICULUM VITAE

CURRICULUM VITAE

Name Maria Lymbery
E-Mail maria.lymbery@uni-due.de

Higher Education:

Doctoral Degree in Mathematics 2013
Optimal Multilevel Methods for Conforming Quadratic, Biquadratic and Bicubic Finite
Elements
*B*AS, Bulgaria, Advisor: Prof. Dr. S. Margenov

Master Degree in Applied Mathematics 2008
Solving Numerically the Matrix Sturm-Liouville Problem for Multilayered Josephson Junctions
*S*ofia University, Bulgaria, Advisor: Prof. Dr. S. Dimova

Bachelor Degree in Applied Mathematics 2004
State exam
*S*ofia University, Bulgaria

Professional Appointments:

- **S**cientific Associate 2023–
  IKIM, University of Duisburg-Essen (Essen, Germany)
- **S**cientific Associate 2014–2022
  Faculty of Mathematics, University of Duisburg-Essen (Essen, Germany)
- **R**esearch Scientist 2014–2014
  IICT, Bulgarian Academy of Sciences (Sofia, Bulgaria)
- **P**rogram developer 2013–2014
  IICT, Bulgarian Academy of Sciences (Sofia, Bulgaria)
- **R**esearch Scientist 2012–2013
  IICT, Bulgarian Academy of Sciences (Sofia, Bulgaria)
- **P**rogram Developer 2011–2012
  IICT, Bulgarian Academy of Sciences (Sofia, Bulgaria)
- **I**nstructor in Numerical Methods 2006–2007
  Faculty of Mathematics, Sofia University (Sofia, Bulgaria)

Member of Scientific and Organizing Committees:

- Member of the Scientific Committee and organizer of a special session on "Robust Iterative Solution Methods for Coupled Problems in Poromechanics" at the "13th International Conference on Large-Scale Scientific Computations", Sozopol, Bulgaria, June 7–11, 2021.
- Member of the Scientific Committee and organizer of a special session on "Modelling of coupled processes in porous media and other multiphysics problems" at the "Modelling 2019" Conference, Olomouc, Czechia, September 16–20, 2019.
- Member of the Scientific Committee and organizer of a special session on "Recent Advances in Numerical Methods for Flow in Deformable Porous Media"

at the "12th International Conference on Large-Scale Scientific Computations", Sozopol, Bulgaria, June 10–14, 2019.

- Member of the Organizing Committee of the "8th International Conference on Large-Scale Scientific Computations", Sozopol, Bulgaria, June 15–18, 2011, "9th International Conference on Large-Scale Scientific Computations", Sozopol, Bulgaria, June 3–7, 2013, "10th International Conference on Large-Scale Scientific Computations", Sozopol, Bulgaria, June 8–12, 2015.

Reviews:

- Referee for SIAM Journal on Scientific Computing, Numerical Linear Algebra with Applications, Computational Geosciences, Computational and Mathematical Methods, Expert Systems with Applications, ETNA

Selected Seminar and Conference Talks:

- Integrated framework for the numerical solution of generalized Biot's systems and applications in biomedical sciences. Invited talk at 92nd annual meeting of the International Association of Applied Mathematics and Mechanics (GAMM), Aachen, Germany, August 15–19, 2022.
- Hybridized discontinuous Galerkin/hybrid mixed discretizations for multiple network poroelasticity. European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022), Oslo, Norway, June 5–9 2022.
- Performance Comparison of Discretizations and Solvers for Simulation of Fluid Transport in the Brain. SIAM Conference on Mathematical & Computational Issues in the Geosciences (GS21) Advanced Solvers for Poromechanics. Virtual conference, June 21–24 2021.
- Simulation of Fluid Transport in the Human Brain: a Comparison of Discretizations and Solvers. 13th International Conference on Large-Scale Scientific Computations (LSSC'21), Sozopol, Bulgaria, June 7–11, 2021.
- Parameter-robust convergence analysis of fixed-stress split iterative scheme for multiple-permeability models. European Numerical Mathematics and Advanced Applications Conference (ENUMATH 2019), Egmond aan Zee, The Netherlands, September 30-October 4, 2019.
- Parameter-robust fixed-stress split iterative scheme for multiple-permeability poroelasticity systems. Modelling 2019, Olomouc, Czechia, September 16–20, 2019.
- Fixed-stress split iterative scheme for multiple-permeability poroelasticity system: Parameter-robust convergence analysis. 12th International Conference on Large-Scale Scientific Computations (LSSC'19), Sozopol, Bulgaria, June 10–14, 2019.
- Multiple-network poroelastic systems: stable discretizations and applications. 8th International Conference on Computational Methods in Applied Mathematics (CMAM-8), Minsk, Belarus, July 2–6, 2018.
- Numerically Computed Estimates of the LBB Constant. 10th International Conference on Large-Scale Scientific Computations (LSSC'15), Sozopol, Bulgaria, June 8–12, 2015.

- MinRes Iteration with Auxiliary Space Multigrid Preconditioning for Darcy problem. Numerical Methods and Applications (NM&A'10), Borovets, Bulgaria, August 20 – 24, 2014.
- Incomplete Factorization by Local Exact Factorization (ILUE). Modelling 2014, Roznov Pod Radhostem, Czechia, June 2–6, 2014.
- ILUE preconditioning within auxiliary space multigrid methods. 6th International Conference on Computational Methods in Applied Mathematics (CMAM-6), Strobl, Austria, September 29–October 03, 2014.
- Auxiliary space multigrid method for flows in porous media. Preconditioning of Iterative Methods (PIM 2013), Prague, Czechia, July 1–5, 2013.
- On the robustness of multilevel preconditioners for quadratic FE discretizations of anisotropic elliptic problems. European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2012), Vienna, Austria, September 10–14, 2012.
- Robust balanced semi-coarsening AMLI preconditioning of biquadratic FEM systems. 11th Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences (AMiTaNS'11), Albena, Bulgaria, June 20–25, 2011.
- Analysis of the Constant in the Strengthened Cauchy-Bunyakowski-Schwarz Inequality for Quadratic Finite Elements. Numerical Methods and Applications (NM&A'10), Borovets, Bulgaria, August 20 – 24, 2010.

Publications in Peer-Reviewed Journals:

[1] Q. Hong, J. Kraus, M. Lymbery, F. Philo. A new practical framework for the stability analysis of perturbed saddle-point problems and applications. *Math. Comp.* 92, 607–634, 2023.

[2] Q. Hong, J. Kraus, M. Kuchta, M. Lymbery, K.A. Mardal, M.E. Rognes. Robust approximation of generalized Biot-Brinkman problems. *J. Sci. Comput.* 93(77), 2022.

[3] M. Bause, M. Lymbery, K. Osthues. C^1-conforming variational discretization of the biharmonic wave equation. *Comput. Math. with Appl.*, 119, 208–219, 2022.

[4] J. Kraus, P. Lederer, M. Lymbery, J. Schöberl. Uniformly well-posed hybridized discontinuous Galerkin/hybrid mixed discretizations for Biot's consolidation model. *Comput. Methods Appl. Mech. Eng.* 384, 113991, 2021.

[5] Q. Hong, J. Kraus, M. Lymbery, F. Philo. Parameter-robust Uzawa-type iterative methods for double saddle point problems arising in Biot's consolidation and multiple-network poroelasticity models. *Math. Models Methods Appl. Sci. (M3AS)* 30(13), 2523–2555, 2020.

[6] Q. Hong, J. Kraus, M. Lymbery, M.F. Wheeler. Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems. *SIAM Multiscale Model. Sim. (SIAM MMS)* 18(2), 916–941, 2020.

[7] Q. Hong, J. Kraus, M. Lymbery, F. Philo. Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models. *Numer. Linear Alg. Appl.* 2019;e2242, 2019.

[8] J. Kraus, M. Lymbery. Incomplete factorization by local exact factorization (ILUE). *Math. Comput. Simul.* 145:50-61, 2018.

[9] J. Kraus, R. Lazarov, M. Lymbery, S. Margenov, L. Zikatanov. Preconditioning heterogeneous $H(\mathrm{d}iv)$ problems by additive Schur complement approximation and applications. *SIAM J. Sci. Comput.* 38(2):A875-A898, 2016.

[10] J. Kraus, M. Lymbery, S. Margenov. Auxiliary space multigrid method based on additive Schur complement approximation. *Numer. Linear Algebra Appl.* 22(6):965-986, 2015.

[11] J. Kraus, M. Lymbery, S. Margenov. Robust multilevel methods for quadratic finite element anisotropic elliptic problems. *Numer. Linear Algebra Appl.* 21:375–398, 2014.

[12] M. Lymbery, S. Margenov. Robust Semi-Coarsening Multilevel Preconditioning of Biquadratic FEM Systems. *Cent. Eur. J. Math.*, 10(1), pp. 357–369, 2012.

Refereed Publications in Conference Proceedings:

[13] J. Kraus, M. Lymbery. Auxiliary space multigrid method for elliptic problems with highly varying coefficients. In *Domain Decomposition Methods in Science and Engineering XXII*, R. Bank et al., eds., Volume 104 of *Lecture Notes in Computational Science and Engineering*, pp. 29–40, 2016.

[14] M. Lymbery. Robust Balanced Semi-Coarsening Multilevel Preconditioning of Bicubic FEM Systems. In *Large-Scale Scientific Computing*, I. Lirkov, S. Margenov, and J. Wasniewski, eds., *LNCS* vol. 8353, pp. 628–635, Berlin Heidelberg, 2014.

[15] J. Kraus, M. Lymbery, S. Margenov. Semi-coarsening AMLI preconditioning of higher order elliptic problems. In *AIP Conf. Proc.* vol. 1487, pp. 30–41, 2013.

[16] J. Kraus, M. Lymbery, S. Margenov. On the robustness of two-level preconditioners for quadratic FE orthotropic elliptic problems. In *Large-Scale Scientific Computing*, I. Lirkov, S. Margenov, and J. Wasniewski, eds., *LNCS* vol. 7116, pp. 582–589, Springer, Berlin Heidelberg, 2012.

[17] M. Lymbery, S. Margenov. Robust Balanced Semi-Coarsening AMLI Preconditioning of Biquadratic FEM Systems. In *AIP Conf. Proc.*, vol. 1404, pp. 438–447, 2011.

Chapters in Books and Survey Articles:

[18] J. Kraus, M. Lymbery, S. Margenov. Robust algebraic multilevel preconditioners for anisotropic problems. In *Numerical Solution of Partial Differential Equations: Theory, Algorithms and their Applications*, Springer Proceedings in Mathematics & Statistics 45, pp. 217-245, Springer Science+Business Media New York 2013.

Submitted Manuscripts/Technical Reports:

[19] J. Kraus, P. Lederer, M. Lymbery, J. Schöberl, K. Osthues. Hybridized discontinuous Galerkin/hybrid mixed methods for a multiple network poroelasticity model with application in biomechanics. arXiv:2205.06732.

[20] J. Alff, O. Bakumenko, M. Birnbach, A. Dada, J. Fragemann, F. Jonske, J. Kleesiek, M. Kondratenko, J. Kraus, M. Lymbery, E. Nasca, K. Osthues, L. Reinike, B. Schulz, F. Siethoff, T. Simon, J. Wang, and N. Zhang. Designing and implementing an interactive cloud platform for teaching machine learning with medical data.

Teaching and Lecturing Experience:

- Seminar/Praktikum "Machine Learning in Medicine - Theory & Practice", in German, UDE, winter 2022/2023.
- Exercise "Numerical Analysis 2", in German, UDE, summer 2022.
- Seminar "Deep Neural Networks", in German, UDE, summer 2022.
- Exercise "Numerical Analysis 1", in German, UDE, winter 2021/2022.
- Seminar "Neural Networks and Deep Learning" exercise "Multigrid and Domain Decomposition Methods", in German, UDE, summer 2021.
- Lecture "Machine Learning", seminar "Approximations Theory and Praxis", in German, UDE, winter 2020/2021.
- Lecture "Machine Learning", master seminar "Discontinuous Galerkin Methods", in German, UDE, summer 2020.
- Master seminar "Machine Learning", exercise "Algebraic Multigrid Methods", in German, UDE, winter 2019/2020.
- Lecture "Programming of Numerical Algorithms in C++", bachelor and master seminar "Machine Learning", exercise "Multigrid and Domain Decomposition Methods", in German, UDE, summer 2019.
- Exercise "Numerical Methods for Partial Differential Equations", in German, UDE, winter 2018/2019.
- Lecture "Numerical Analysis 2", substitution of Prof. J. Kraus, in German, UDE, summer 2018.
- Exercise "Numerical Analysis 1", in German, winter 2017/2018.
- Bachelor and master seminar "Algebraic Multigrid Methods", combined course "Object-oriented Programming", exercise "Iterative Methods for Linear Systems of Equations", in German, UDE, summer 2017.
- Bachelor and master seminar "Algebraic Multigrid Methods", exercise "Partial Differential Equations", in German, UDE, winter 2016/2017.
- Exercise "Multigrid and Domain Decomposition Methods", practicum "Numerical Analysis", in German, UDE, summer 2016.

- Master seminar "Preconditioning and Krylov-Subspace Methods", exercise "Numerical Methods for Partial Differential Equations", in German, UDE, winter 2015/2016.
- Combined course "Object-oriented Programming", exercise "Numerical Analysis 2", practicum "Numerical Analysis", in German, UDE, summer 2015.
- Exercise "Numerical Analysis 1", in English, UDE, winter 2014/2015.
- Exercise "Finite Elements", in Bulgarian, Sofia University, summer 2007.

Supervision of Master Students:
Anna Hofmann (UDE), Kevin Osthues (UDE), Frederik Baucks (UDE), Herve Emedac (UDE), Jana Fragemann (UDE), Jan Meinen (UDE), Jan Ulmer (UDE), Sascha Beutler (UDE), Saskia Rogl (UDE)