Results in Applied Mathematics 21 (2024) 100438

Contents lists available at ScienceDirect

Results in Applied Mathematics

journal homepage: www.elsevier.com/locate/results-in-applied-mathematics

New twofold saddle-point formulations for Biot poroelasticity with porosity-dependent permeability ☆

Bishnu P. Lamichhane [a], Ricardo Ruiz-Baier [b,c,d,*], Segundo Villa-Fuentes [b]

[a] School of Mathematical & Physical Sciences, University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
[b] School of Mathematics, Monash University, 9 Rainforest Walk, Melbourne, VIC 3800, Australia
[c] World-Class Research Center "Digital biodesign and personalised healthcare", Sechenov First Moscow State Medical University, Moscow, Russia
[d] Universidad Adventista de Chile, Casilla 7-D, Chillán, Chile

ARTICLE INFO

MSC:
65N30
65N15
65J15
76S05
35Q74

Keywords:
Mixed finite element methods
Hu–Washizu formulation
Nonlinear poroelasticity
Twofold saddle-point problems
Fixed-point operators

ABSTRACT

We propose four-field and five-field Hu–Washizu-type mixed formulations for nonlinear poroelasticity – a coupled fluid diffusion and solid deformation process – considering that the permeability depends on a linear combination between fluid pressure and dilation. As the determination of the physical strains is necessary, the first formulation is written in terms of the primal unknowns of solid displacement and pore fluid pressure as well as the poroelastic stress and the infinitesimal strain, and it considers strongly symmetric Cauchy stresses. The second formulation imposes stress symmetry in a weak sense and it requires the additional unknown of solid rotation tensor. We study the unique solvability of the problem using the Banach fixed-point theory, properties of twofold saddle-point problems, and the Banach–Nečas–Babuška theory. We propose monolithic Galerkin discretisations based on conforming Arnold–Winther for poroelastic stress and displacement, and either PEERS or Arnold–Falk–Winther finite element families for the stress–displacement-rotation field variables. The wellposedness of the discrete problem is established as well, and we show a priori error estimates in the natural norms. Some numerical examples are provided to confirm the rates of convergence predicted by the theory, and we also illustrate the use of the formulation in some typical tests in Biot poroelasticity.

1. Introduction

1.1. Scope

The coupling of interstitial fluid flow and solid mechanics in a porous medium has an important role in a number of socially relevant applications [1]. In particular, nonlinear poroelasticity equations arise, for example, in models of geomechanics and in the study of deformable soft tissues (such as filtration of aqueous humor through cartilage-like structures in the eye and with application in glaucoma formation). In the present work we focus on the case of fully saturated deformable porous media (the

☆ **Funding:** This work has been partially supported by the Monash Mathematics Research Fund S05802-3951284; by the Australian Research Council through the Future Fellowship grant FT220100496 and Discovery Project grant DP22010316; by the Ministry of Science and Higher Education of the Russian Federation within the framework of state support for the creation and development of World-Class Research Centers Digital biodesign and personalized healthcare No. 075-15-2022-304; and by the National Research and Development Agency (ANID) of the Ministry of Science, Technology, Knowledge and Innovation of Chile through the postdoctoral program Becas Chile grant 74220026.

* Corresponding author at: School of Mathematics, Monash University, 9 Rainforest Walk, Melbourne, VIC 3800, Australia.
E-mail addresses: Bishnu.Lamichhane@newcastle.edu.au (B.P. Lamichhane), Ricardo.RuizBaier@monash.edu (R. Ruiz-Baier), Segundo.VillaFuentes@monash.edu (S. Villa-Fuentes).

https://doi.org/10.1016/j.rinam.2024.100438
Received 29 June 2023; Received in revised form 13 December 2023; Accepted 19 January 2024
Available online 29 January 2024

solid and fluid constituents of the mixture occupy a complementary fraction of volume in the macroscopic body) and in instances where the permeability coefficient depends on the porosity (which is in turn related to the total amount of fluid in the poroelastic mixture) [2]. In the classical form for this class of problems the momentum and mass balance equations for a solid–fluid mixture are written in terms of the solid displacement of the porous matrix and the averaged interstitial pressure. Examples of analysis of existence and uniqueness of solution can be found in [3–9]. These works expand the theory available for linear Biot consolidation problems using, for example, constructive Galerkin approximations together with Brouwer's fixed-point arguments with compactness and passage to the limit, the theory of monotone operators in Banach spaces and semigroups, abstract results on doubly nonlinear evolution equations, and the modification of the arguments to the case of pseudo-monotone nonlinear couplings using Brézis' theory. One of the goals of this paper is to extend the previous analysis to the case of mixed formulations for the solid phase. Rewriting the governing equations in mixed form using the Hellinger–Reissner principle and writing the total poroelastic stress as a new unknown, is an approach employed already in the analysis of a number of mixed models for linear poroelasticity [10–14] and in poroelasticity/free-fluid couplings [15–17]. These formulations may incorporate also the tensor of rotations to impose in a weak manner the symmetry of the poroelastic stress tensor. There, in deriving the weak forms, one tests the constitutive equation for stress against a test function for stress. In contrast, in the present treatment we also require the tensor of infinitesimal strains as an unknown since it is an important field acting in the coupling with the fluid phase mechanics via the nonlinear permeability (regarded as a function of the interstitial fluid pressure and the trace of the strain tensor). In the context of elasticity problems, the popular Hu–Washizu formulation [18,19] has displacement, stress and strain tensors as three unknowns, and we note that the Hellinger–Reissner formulation mentioned above is a special case of the Hu–Washizu formulation (it is obtained after applying the Fenchel–Legendre transformation eliminating the strain from the latter formulation [20]).

Apart from the application of Hu–Washizu formulations in many works for linear elasticity (see, e.g., [21–25] and the references therein), the solvability analysis of the continuous and discrete twofold saddle-point mixed problems (including also error estimates) has been carried out in [26,27] for Hencky-strain nonlinear elasticity, as well as for more recent models for stress-assisted diffusion coupled with poroelasticity [28]. There, one tests the constitutive equation for stress against the test function associated with the space of infinitesimal strains. In that setting, a key ingredient in the analysis is the assumption that the nonlinearity in the weak forms induces a Lipschitz continuous and strongly monotone operator (this last condition being required in a suitable kernel).

In the present scenario the analysis requires to define a nonlinear operator $\boldsymbol{A} : \mathbf{X} \to \mathbf{X}'$ where $\mathbf{X}$ consists of square integrable and symmetric tensors and scalar functions in H^1. The nonlinearity is inherited from the nonlinear dependence of permeability on fluid pressure and on skeleton strains, and for some constitutive forms, it does not necessarily imply that $\boldsymbol{A}$ is monotone. In this work we present two formulations, in the first we impose the symmetry of the stress tensor in a strong way, while in the second we impose the symmetry in a weak sense, using the rotation tensor as a further unknown. Then, similarly to [29] (see also [30]) we have the nonlinear term inside the saddle-point structure, unlike [31,32] where the nonlinear term is associated with a perturbation of the saddle-point problem. Therefore we proceed by a fixed-point argument and consider a linear twofold saddle-point formulation that suggests the structure of a fixed-point operator. This map is shown to be well-defined (for this we use an appropriate adaptation of the theory from, e.g., [33]), to map a conveniently chosen ball into itself, and to be Lipschitz continuous. Then, by establishing a contracting property the unique solvability will be a consequence of Banach fixed-point theorem. Such an analysis hinges on data smallness assumptions, which involve boundary data, source terms, and permeability bounds.

For the associated Galerkin schemes we employ, on the one hand, for the strong symmetry formulation, Arnold–Winther finite elements of degree $k \geq 1$ [34] to approximate the strain tensor, poroelastic stress tensor and displacement, and continuous piecewise polynomials of degree $k+1$ for the pore pressure; and we note that for this finite element family, one can also employ piecewise polynomials of degree $k+2$ for the symmetric strain tensor to maintain inf-sup stability. On the other hand, for the weak symmetry formulation, we use the classical PEERS elements [35] to approximate the strain tensor, poroelastic stress tensor, displacement and the rotation tensor, and continuous piecewise polynomials of degree $k+1$ for the interstitial pressure (we also use a family based on Arnold–Falk–Winther elements [36]). Next we apply the same arguments utilised for the continuous problem to prove unique solvability. In addition, using standard tools and techniques for the error decomposition, and approximation properties of mentioned finite element spaces, we obtain the corresponding Céa estimate and rates of convergence.

1.2. Outline

The content of this paper has been laid out as follows. In the remainder of this section we include notation conventions and preliminary results that are used throughout the manuscript. Section 1.4 provides details of the model, describing the components of the balance equations and stating boundary conditions. One weak formulation (with strong stress symmetry) and its properties are collected in Section 2. Section 3 is devoted to the analysis of solvability of this weak form, using arguments from the twofold saddle-point variant of the Babuška–Brezzi theory from [30]. Next we address in Section 4 the solvability and stability analysis of the discrete problem, where similar arguments are employed, following [27]. A priori error estimates are derived in Section 5. In section Section 6 we reformulate the equations to include weak imposition of stress, and in Section 7 we collect computational results, consisting in verification of convergence and simulation of different cases on simple geometries.

1.3. Notation and preliminaries

Let $L^2(\Omega)$ be the set of all square-integrable functions in $\Omega \subset \mathbb{R}^d$ where $d \in \{2,3\}$ is the spatial dimension, and denote by $\mathbf{L}^2(\Omega) = L^2(\Omega)^d$ its vector-valued counterpart and by $\mathbb{L}^2(\Omega) = L^2(\Omega)^{d\times d}$ its tensor-valued counterpart. We also write

$$\mathbb{L}^2_{\rm sym}(\Omega) := \{\boldsymbol{\tau} \in \mathbb{L}^2(\Omega) : \ \boldsymbol{\tau} = \boldsymbol{\tau}^{\rm t}\}, \quad \mathbb{L}^2_{\rm skew}(\Omega) := \{\boldsymbol{\tau} \in \mathbb{L}^2(\Omega) : \ \boldsymbol{\tau} = -\boldsymbol{\tau}^{\rm t}\},$$

to represent the symmetric and skew-symmetric tensors in Ω with each component being square-integrable. Standard notation will be employed for Sobolev spaces $\mathrm{H}^m(\Omega)$ with $m \geq 0$ (and we note that $\mathrm{H}^0(\Omega) = L^2(\Omega)$). Their norms and seminorms are denoted as $\|\cdot\|_{m,\Omega}$ and $|\cdot|_{m,\Omega}$, respectively (as well as for their vector and tensor-valued counterparts $\mathbf{H}^m(\Omega)$, $\mathbb{H}^m(\Omega)$) see, e.g., [37].

As usual $\mathbb{I}$ stands for the identity tensor in $\mathbb{R}^{d\times d}$, and $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^d$. Also, for any vector fields $\boldsymbol{v} = (v_i)_{i=1,d}$ we set the gradient and divergence operators as

$$\nabla \boldsymbol{v} := \left(\frac{\partial v_i}{\partial x_j}\right)_{i,j=1,d} \quad \text{and} \quad \mathrm{div}\ \boldsymbol{v} := \sum_{j=1}^{d} \frac{\partial v_j}{\partial x_j}.$$

In addition, for any tensor fields $\boldsymbol{\tau} = (\tau_{ij})_{i,j=1,d}$ and $\boldsymbol{\zeta} = (\zeta_{ij})_{i,j=1,d}$, we let $\mathbf{div}\,\boldsymbol{\tau}$ be the divergence operator div acting along the rows of $\boldsymbol{\tau}$, and define the transpose, the trace and the tensor inner product, respectively, as

$$\boldsymbol{\tau}^{\rm t} := (\tau_{ji})_{i,j=1,d}, \quad \mathrm{tr}(\boldsymbol{\tau}) := \sum_{i=1}^{d} \tau_{ii}, \quad \text{and} \quad \boldsymbol{\tau} : \boldsymbol{\zeta} := \sum_{i,j=1}^{n} \tau_{ij}\zeta_{ij}.$$

We also recall the Hilbert space

$$\mathbf{H}(\mathrm{div};\Omega) := \left\{\boldsymbol{z} \in \mathbf{L}^2(\Omega) : \ \mathrm{div}\ \boldsymbol{z} \in \mathrm{L}^2(\Omega)\right\},$$

with norm $\|\boldsymbol{z}\|^2_{\mathrm{div};\Omega} := \|\boldsymbol{z}\|^2_{0,\Omega} + \|\mathrm{div}\ \boldsymbol{z}\|^2_{0,\Omega}$, and introduce the tensor version of $\mathbf{H}(\mathrm{div};\Omega)$ given by

$$\mathbb{H}(\mathbf{div};\Omega) := \left\{\boldsymbol{\tau} \in \mathbb{L}^2(\Omega) : \quad \mathbf{div}\,\boldsymbol{\tau} \in \mathbf{L}^2(\Omega)\right\},$$

whose norm will be denoted by $\|\cdot\|_{\mathbf{div};\Omega}$.

1.4. Governing equations

Let us consider a fully-saturated poroelastic medium (consisting of a mechanically isotropic and homogeneous fluid–solid mixture) occupying the open and bounded domain Ω in $\mathbb{R}^d$, with Lipschitz boundary Γ. The symbol $\boldsymbol{n}$ will stand for the unit outward normal vector on the boundary. Let $\boldsymbol{f} \in \mathbf{L}^2(\Omega)$ be a prescribed body force per unit of volume (acting on the fluid–structure mixture) and let $g \in L^2(\Omega)$ be a net volumetric fluid production rate.

Under the assumption of negligible gravitational effects as well as material deformations being sufficiently small, and varying sufficiently slowly so that inertial effects are considered negligible (see further details in [38]), we have that the balance of linear momentum for the solid–fluid mixture is written as

$$-\mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f} \qquad \text{in } \Omega, \tag{1.1}$$

with $\boldsymbol{\sigma}$ being the total Cauchy stress tensor of the mixture (conformed by the effective solid stress and effective fluid stress), whose dependence on strain and on fluid pressure is given by the constitutive assumption (or effective stress principle)

$$\boldsymbol{\sigma} = C\boldsymbol{d} - \alpha p\mathbb{I} \qquad \text{in } \Omega. \tag{1.2}$$

Here the skeleton displacement vector $\boldsymbol{u}$ from the position $\boldsymbol{x} \in \Omega$ is an unknown, the tensor $\boldsymbol{d} = \boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + [\nabla\boldsymbol{u}]^{\rm t})$ is the infinitesimal strain, by C we denote the fourth-order elasticity tensor, also known as Hooke's tensor (symmetric and positive definite and characterised by $C\boldsymbol{d} := \lambda(\mathrm{tr}\,\boldsymbol{d})\mathbb{I} + 2\mu\,\boldsymbol{d}$), $\mathbb{I}$ is the identity second-order tensor, λ and μ are the Lamé parameters (assumed constant and positive), $0 \leq \alpha \leq 1$ is the Biot–Willis parameter, and p denotes the Darcy fluid pressure (positive in compression), which is an unknown in the system.

We also consider the balance of angular momentum, which in this context states that the total poroelastic stress is a symmetric tensor

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\rm t}. \tag{1.3}$$

The fluid content (due to both fluid saturation and local volume dilation) is given by

$$\zeta = c_0 p + \alpha\,\mathrm{div}\,\boldsymbol{u} = c_0 p + \alpha\,\mathrm{tr}\,\boldsymbol{d}, \tag{1.4}$$

where c_0 is the constrained specific storage (or storativity) coefficient. Using Darcy's law to describe the discharge velocity in terms of the fluid pressure gradient, we can write the balance of mass for the total amount of fluid in the mixture as $\partial_t\zeta - \mathrm{div}(\kappa\nabla p) = g$ in $\Omega\times(0,t_{\rm end})$, where κ is the intrinsic permeability (divided by the fluid viscosity) of the laminar flow in the medium, a nonlinear function of the porosity. In turn, in the small strains limit the porosity can be approximated by a linear function of the fluid content ζ (see for example [9, Section 2.1]), and so, thanks to (1.4), we can simply write $\kappa(\boldsymbol{d}, p)$. Furthermore, after a backward Euler

semi-discretisation in time with a constant time step and rescaling appropriately, we only consider the type of equations needed to solve at each time step and therefore we will concentrate on the form

$$c_0 p + \alpha \operatorname{tr} \boldsymbol{d} - \operatorname{div}(\kappa(\boldsymbol{d}, p)\nabla p) = g \qquad \text{in } \Omega. \tag{1.5}$$

Typical constitutive relations for permeability are, for example, of exponential or Kozeny–Carman type (see, e.g., [39])

$$\kappa(\boldsymbol{d}, p) = \frac{k_0}{\mu_f}\mathbb{I} + \frac{k_1}{\mu_f}\exp(k_2(c_0 p + \alpha \operatorname{tr} \boldsymbol{d}))\mathbb{I}, \quad \kappa(\boldsymbol{d}, p) = \frac{k_0}{\mu_f}\mathbb{I} + \frac{k_1(c_0 p + \alpha \operatorname{tr} \boldsymbol{d})^3}{\mu_f(1-(c_0 p + \alpha \operatorname{tr} \boldsymbol{d}))^2}\mathbb{I}, \tag{1.6}$$

where μ_f denotes the viscosity of the interstitial fluid and k_0, k_1, k_2 are model constants. We note that in the case of incompressible constituents one has $c_0 = 0$ and $\alpha = 1$, indicating that permeability depends only on the dilation $\operatorname{tr} \boldsymbol{d} = \operatorname{div} \boldsymbol{u}$ (see, e.g., [3]). We also note that even in such a scenario (of incompressible phases) the overall mixture is not necessarily incompressible itself. More precise assumptions on the behaviour of the permeability are postponed to Section 3.

To close the system of equations, we consider non-homogeneous displacement boundary conditions for the momentum balance and non-homogeneous flux boundary conditions on the mass balance equation. For prescribed $\boldsymbol{u}_\Gamma \in \mathbf{H}^{1/2}(\Gamma)$ and $r_\Gamma \in \mathrm{H}^{-1/2}(\Gamma)$ we set

$$\boldsymbol{u} = \boldsymbol{u}_\Gamma \quad \text{and} \quad \kappa(\boldsymbol{d}, p)\nabla p \cdot \boldsymbol{n} = r_\Gamma \qquad \text{on } \Gamma. \tag{1.7}$$

Remark 1.1. Note that the boundary conditions (1.7) are considered to simplify the exposition. While there is no theoretical issue in imposing Dirichlet condition for p on a subset of Γ, our analysis covers the case if Dirichlet condition is imposed for $\boldsymbol{u}$ on only a subset of Γ, which has non-zero $(d-1)$ measure. Part of the required modifications are in the proof of unique solvability of auxiliary problems required in the verification of inf-sup conditions, which will still hold in the case of mixed boundary conditions provided that suitable assumptions are made on the domain boundary.

We stress that the wellposedness of the time-dependent variational problem in two-field formulation of the same underlying continuous problem (physical model of nonlinear poroelasticity) has recently been presented in [40]. The analysis in that reference is conducted for the case of incompressible constituents (taking $c_0 = 0$). The authors also assume that the permeability is uniformly bounded away from zero and define a fixed-point map in terms of the amount of fluid (which only involves the dilation term). In our case, the starting model problem is steady, the weak formulations are in mixed and mixed-primal form, and the analysis follows a different fixed-point argument.

2. A four-field formulation and preliminary properties

2.1. Derivation of weak forms

We proceed to test Eq. (1.1) against $\boldsymbol{v} \in \mathbf{L}^2(\Omega)$, to test the constitutive equation for strain $\boldsymbol{d} = \boldsymbol{\varepsilon}(\boldsymbol{u})$ against $\boldsymbol{\tau} \in \mathbb{H}_{\mathrm{sym}}(\mathbf{div}; \Omega) := \{\boldsymbol{\tau} \in \mathbb{L}^2_{\mathrm{sym}}(\Omega) : \ \mathbf{div}\, \boldsymbol{\tau} \in \mathbf{L}^2(\Omega)\}$, the Eqs. (1.2) and (1.5), by $\boldsymbol{e} \in \mathbb{L}^2_{\mathrm{sym}}(\Omega)$ and $q \in \mathrm{H}^1(\Omega)$, respectively, integrate by parts and using the boundary conditions (1.7) naturally, we finally arrive at

$$\begin{aligned}
-\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\sigma} &= \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} && \forall \boldsymbol{v} \in \mathbf{L}^2(\Omega),\\
-\int_\Omega \boldsymbol{\tau} : \boldsymbol{d} - \int_\Omega \boldsymbol{u} \cdot \mathbf{div}\, \boldsymbol{\tau} &= -\langle \boldsymbol{\tau}\boldsymbol{n}, \boldsymbol{u}_\Gamma\rangle_\Gamma && \forall \boldsymbol{\tau} \in \mathbb{H}_{\mathrm{sym}}(\mathbf{div}; \Omega),\\
\int_\Omega C\boldsymbol{d} : \boldsymbol{e} - \alpha\int_\Omega p \operatorname{tr} \boldsymbol{e} - \int_\Omega \boldsymbol{\sigma} : \boldsymbol{e} &= 0 && \forall \boldsymbol{e} \in \mathbb{L}^2_{\mathrm{sym}}(\Omega),\\
\int_\Omega \kappa(\boldsymbol{d}, p)\nabla p \cdot \nabla q + c_0\int_\Omega p\, q + \alpha\int_\Omega q \operatorname{tr} \boldsymbol{d} &= \int_\Omega g\, q + \langle r_\Gamma, q\rangle_\Gamma && \forall q \in \mathrm{H}^1(\Omega),
\end{aligned} \tag{2.1}$$

where $\langle\cdot,\cdot\rangle_\Gamma$ denotes the duality pairing between $\mathrm{H}^{-1/2}(\Gamma)$ and its dual $\mathrm{H}^{1/2}(\Gamma)$ with respect to the inner product in $L^2(\Gamma)$ (and we use the same notation in the vector-valued case). Note also that the balance of angular momentum (1.3) has been enforced as an essential condition in the functional space for poroelastic stress.

Next, we notice that (2.1) can be regarded as a twofold saddle-point structure. In fact, let us adopt the following notation for the Hilbert spaces for the strain–pressure pair, the poroelastic stress, and the displacement:

$$\mathbf{X} := \mathbb{L}^2_{\mathrm{sym}}(\Omega) \times \mathrm{H}^1(\Omega), \qquad \mathbf{Y} := \mathbb{H}_{\mathrm{sym}}(\mathbf{div}; \Omega) \quad \text{and} \quad \mathbf{Z} := \mathbf{L}^2(\Omega),$$

respectively. In addition, we group and order the trial and test functions as follows:

$$\underline{\boldsymbol{d}} := (\boldsymbol{d}, p) \in \mathbf{X}, \quad \boldsymbol{\sigma} \in \mathbf{Y}, \quad \boldsymbol{u} \in \mathbf{Z},$$

$$\underline{\boldsymbol{e}} := (\boldsymbol{e}, q) \in \mathbf{X}, \quad \boldsymbol{\tau} \in \mathbf{Y}, \quad \boldsymbol{v} \in \mathbf{Z},$$

where $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X}\times\mathbf{Y}$ and $\mathbf{Z}$ are endowed with the norms

$$\|\underline{\boldsymbol{e}}\|^2_{\mathbf{X}} := \|\boldsymbol{e}\|^2_{0,\Omega} + \|q\|^2_{1,\Omega}, \quad \|\boldsymbol{\tau}\|_{\mathbf{Y}} := \|\boldsymbol{\tau}\|_{\mathbf{div},\Omega}, \quad \|(\underline{\boldsymbol{e}}, \boldsymbol{\tau})\|^2_{\mathbf{X}\times\mathbf{Y}} := \|\underline{\boldsymbol{e}}\|^2_{\mathbf{X}} + \|\boldsymbol{\tau}\|^2_{\mathbf{Y}}$$

$$\|\boldsymbol{v}\|_{\mathbf{Z}} := \|\boldsymbol{v}\|_{0,\Omega}, \quad \|((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{v})\|^2 := \|(\underline{\boldsymbol{e}},\boldsymbol{\tau})\|^2_{\mathbf{X}\times\mathbf{Y}} + \|\boldsymbol{v}\|^2_{\mathbf{Z}}.$$

Introducing the nonlinear and bilinear weak forms $a : \mathbf{X}\times\mathbf{X} \to \mathbb{R}$, $b_1 : \mathbf{X}\times\mathbf{Y} \to \mathbb{R}$ and $b_2 : (\mathbf{X}\times\mathbf{Y})\times\mathbf{Z} \to \mathbb{R}$ defined by

$$\begin{aligned} a(\underline{\boldsymbol{d}},\underline{\boldsymbol{e}}) &:= \int_\Omega C\boldsymbol{d} : \boldsymbol{e} + \int_\Omega \kappa(\underline{\boldsymbol{d}})\nabla p\cdot\nabla q + c_0\int_\Omega p\,q + \alpha\int_\Omega q\ \mathrm{tr}\,\boldsymbol{d} - \alpha\int_\Omega p\ \mathrm{tr}\,\boldsymbol{e},\\ b_1(\underline{\boldsymbol{e}},\boldsymbol{\tau}) &:= -\int_\Omega \boldsymbol{\tau} : \boldsymbol{e},\\ b_2((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{v}) &:= -\int_\Omega \boldsymbol{v}\cdot\mathbf{div}\,\boldsymbol{\tau}, \end{aligned} \tag{2.2}$$

respectively; and the linear functionals $F \in \mathbf{Z}'$, $H \in \mathbf{Y}'$, $G \in \mathbf{X}'$ by

$$F(\boldsymbol{v}) := \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}, \quad H(\boldsymbol{\tau}) := -\langle\boldsymbol{\tau}\boldsymbol{n},\boldsymbol{u}_\Gamma\rangle_\Gamma, \quad G(\underline{\boldsymbol{e}}) := \int_\Omega g\,q + \langle r_\Gamma, q\rangle_\Gamma,$$

we can write the weak form (2.1) as follows: Find $((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\boldsymbol{u}) \in (\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$ such that

$$\begin{aligned} a(\underline{\boldsymbol{d}},\underline{\boldsymbol{e}}) \;+\; b_1(\underline{\boldsymbol{e}},\boldsymbol{\sigma}) &= G(\underline{\boldsymbol{e}}),\\ b_1(\underline{\boldsymbol{d}},\boldsymbol{\tau}) \;+\; b_2((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{u}) &= H(\boldsymbol{\tau}),\\ b_2((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\boldsymbol{v}) &= F(\boldsymbol{v}), \end{aligned} \tag{2.3}$$

for all $((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{v}) \in (\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$.

Remark 2.1. Note that when c_0 approaches zero, our control over the L^2-part of the fluid pressure norm diminishes. Consequently, the uniqueness of fluid pressure cannot be guaranteed unless we search for it within a space such as $\mathrm{H}^1(\Omega)\cap L^2_0(\Omega)$, due to the pure flux boundary conditions imposed on the mass balance equation. Without this consideration, the lack of uniqueness would also extend to stress, as implied by (1.2). In such cases, it becomes necessary to restrict tensors in the spaces $\mathbf{Y}$ to those with a zero mean value (and similarly for the space $\widetilde{\mathbf{Y}}$ to be introduced in section Section 6). A similar scenario arises when α tends to zero: the poroelastic stress loses its unique definition (even though the fluid pressure retains it), requiring the adoption of the zero mean condition within the stress space.

2.2. Stability properties and suitable inf-sup conditions

We start by establishing the boundedness of the bilinear forms b_1 and b_2:

$$\left|b_1(\underline{\boldsymbol{e}},\boldsymbol{\tau})\right| \le \|\underline{\boldsymbol{e}}\|_{\mathbf{X}}\|\boldsymbol{\tau}\|_{\mathbf{Y}}, \qquad \left|b_2((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{v})\right| \le \|(\underline{\boldsymbol{e}},\boldsymbol{\tau})\|_{\mathbf{X}\times\mathbf{Y}}\|\boldsymbol{v}\|_{\mathbf{Z}}. \tag{2.4a}$$

On the other hand, using Hölder and trace inequalities we can readily observe that the right-hand side functionals are all bounded

$$\begin{aligned} &\left|G(\underline{\boldsymbol{e}})\right| \lesssim (\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma})\|\underline{\boldsymbol{e}}\|_{\mathbf{X}} \quad \forall\underline{\boldsymbol{e}}\in\mathbf{X}, \qquad \left|H(\boldsymbol{\tau})\right| \lesssim \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma}\|\boldsymbol{\tau}\|_{\mathbf{Y}} \quad \forall\boldsymbol{\tau}\in\mathbf{Y},\\ &\left|F(\boldsymbol{v})\right| \le \|\boldsymbol{f}\|_{0,\Omega}\,\|\boldsymbol{v}\|_{\mathbf{Z}} \quad \forall\boldsymbol{v}\in\mathbf{Z}. \end{aligned} \tag{2.5}$$

Finally, it is straightforward to see that the kernel of the bilinear form b_2 is a closed subspace of $\mathbf{X}\times\mathbf{Y}$. It is denoted as

$$\mathbf{X}\times\mathbf{Y}_0, \tag{2.6}$$

and the second component admits the characterisation

$$\mathbf{Y}_0 := \{\boldsymbol{\tau}\in\mathbf{Y} : \quad \mathbf{div}\,\boldsymbol{\tau} = \mathbf{0}\} \tag{2.7}$$

On the other hand, we note that b_2 satisfies the inf-sup condition

$$\sup_{\mathbf{0}\neq(\underline{\boldsymbol{e}},\boldsymbol{\tau})\in\mathbf{X}\times\mathbf{Y}} \frac{b_2((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{v})}{\|(\underline{\boldsymbol{e}},\boldsymbol{\tau})\|_{\mathbf{X}\times\mathbf{Y}}} \ge \beta_{b_2}\|\boldsymbol{v}\|_{\mathbf{Z}} \quad \forall\,\boldsymbol{v}\in\mathbf{Z}. \tag{2.8}$$

This is a well-known result stating the surjectivity of the divergence operator (see, e.g., [41]) extended to the tensor case and considering symmetric stresses. It is easily proven by means of the wellposed auxiliary problem of finding, for a given $\boldsymbol{v}\in\mathbf{Z}$, the unique $\boldsymbol{y}\in\mathbf{H}^1_0(\Omega)$ such that

$$-\mathbf{div}[\boldsymbol{\epsilon}(\boldsymbol{y})] = \boldsymbol{v} \quad \text{in } \Omega; \qquad \boldsymbol{y} = \mathbf{0} \quad \text{on } \Gamma,$$

and then constructing $\hat{\boldsymbol{\tau}} = \boldsymbol{\epsilon}(\boldsymbol{y})$ which clearly belongs to $\mathbf{Y}$ and, moreover, it satisfies $\|\hat{\boldsymbol{\tau}}\|_{\mathbf{Y}} \le C\|\boldsymbol{v}\|_{\mathbf{Z}}$. In addition, we note that for all $\boldsymbol{\tau}\in\mathbf{Y}_0$, it suffices to take $\boldsymbol{e} = \boldsymbol{\tau}$ to easily arrive at

$$\sup_{\mathbf{0}\neq\underline{\boldsymbol{e}}\in\mathbf{X}} \frac{b_1(\underline{\boldsymbol{e}},\boldsymbol{\tau})}{\|\underline{\boldsymbol{e}}\|_{\mathbf{X}}} \ge \|\boldsymbol{\tau}\|_{\mathbf{Y}} \quad \forall\,\boldsymbol{\tau}\in\mathbf{Y}_0. \tag{2.9}$$

3. Existence and uniqueness of weak solution

3.1. Preliminaries

We stress that if the permeability κ is a positive constant $\kappa = \kappa_0$ or a space-dependent uniformly bounded scalar field $\kappa(\boldsymbol{x})$ in $L^\infty(\Omega)$, or a positive definite matrix $\kappa = \mathbb{K}(\boldsymbol{x})$, then the variational form $a(\cdot,\cdot)$ is bounded and coercive in $\mathbf{X}$. In this case the system (2.3) is a linear twofold saddle-point problem which is uniquely solvable, thanks to the properties of the bilinear forms $b_1(\cdot,\cdot)$, $b_2(\cdot,\cdot)$ and owing to, e.g., [33, Theorem 3.1].

On the other hand, if the permeability in the variational form $a(\cdot,\cdot)$ induces monotone and Lipschitz-continuous nonlinear operator, i.e.,

$$\boldsymbol{A} : \mathbf{X} \to \mathbf{X}', \quad \underline{\boldsymbol{d}} \mapsto \boldsymbol{A}(\underline{\boldsymbol{d}}), \quad \langle \boldsymbol{A}(\underline{\boldsymbol{d}}), \underline{\boldsymbol{e}} \rangle := a(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}}),$$

with

$$|\langle \boldsymbol{A}(\underline{\boldsymbol{d}}_1) - \boldsymbol{A}(\underline{\boldsymbol{d}}_2), \underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2 \rangle| \gtrsim \|\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2\|_{\mathbf{X}}^2, \qquad \|\boldsymbol{A}(\underline{\boldsymbol{d}}_1) - \boldsymbol{A}(\underline{\boldsymbol{d}}_2)\|_{\mathbf{X}'} \lesssim \|\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2\|_{\mathbf{X}},$$

then the system (2.3) is a nonlinear twofold saddle-point problem, which is uniquely solvable thanks to the properties of the bilinear forms $b_1(\cdot,\cdot)$, $b_2(\cdot,\cdot)$ and a direct application of [30, Lemma 2.1].

However, and as discussed in [4,9,40], some of the typical nonlinearities assumed by κ (1.6) do not guarantee monotonicity of the nonlinear operator $\boldsymbol{A}$.

Note, for example, that in [9] the authors ask that κ (they only consider it a function of the dilation $\operatorname{tr} \boldsymbol{d}$) is such that

$$\kappa \in C^1(\Omega), \quad \kappa(0) > 0, \quad \kappa' > 0,$$

in [6] the permeability κ depends only on the fluid pressure p and it is assumed that

$$0 < k_0 \le \kappa(s) \le k_1 \qquad \forall s \in \mathbb{R}^+,$$

and in [42] a similar uniform boundedness is assumed even if the permeability depends on both pore pressure and the symmetric strain. In our case, for sake of the analysis in this section, we allow the permeability $\kappa(\underline{\boldsymbol{d}}) = \kappa(\boldsymbol{d}, p)$ to be anisotropic but still require that it is a uniformly positive definite second-order tensor in $\mathbb{L}^\infty(\Omega)$, and Lipschitz continuous in $p \in \mathrm{H}^1(\Omega)$. That is, there exist positive constants κ_1, κ_2 such that

$$\kappa_1 |\boldsymbol{v}|^2 \le \boldsymbol{v}^{\mathrm{t}} \kappa(\cdot,\cdot) \boldsymbol{v}, \qquad \|\kappa(\cdot, p_1) - \kappa(\cdot, p_2)\|_{\mathbb{L}^\infty(\Omega)} \le \kappa_2 \|p_1 - p_2\|_{1,\Omega}, \tag{3.1}$$

for all $\boldsymbol{v} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, and for all $p_1, p_2 \in \mathrm{H}^1(\Omega)$.

3.2. Definition of a fixed-point operator

In view of the discussion in Section 3.1, if $\boldsymbol{A}$ (the operators induced by the nonlinear weak form $a(\cdot,\cdot)$) is not monotone, then we proceed to define, for a given $r > 0$, the following set

$$\mathbf{W} := \left\{ \underline{\boldsymbol{w}} := (\boldsymbol{w}, s) \in \mathbf{X} : \quad \|\underline{\boldsymbol{w}}\|_{\mathbf{X}} \le r \right\}, \tag{3.2}$$

which is a closed ball of $\mathbf{X}$, with centre at the origin and radius r. Next, for a fixed $\underline{\boldsymbol{w}} := (\boldsymbol{w}, s)$ in $\mathbf{W}$, we define the bilinear form $a_{\underline{\boldsymbol{w}}} : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ as follows

$$a_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}}) := \int_\Omega C\boldsymbol{d} : \boldsymbol{e} + \int_\Omega \kappa(\underline{\boldsymbol{w}}) \nabla p \cdot \nabla q + c_0 \int_\Omega p\, q + \alpha \int_\Omega q \operatorname{tr} \boldsymbol{d} - \alpha \int_\Omega p \operatorname{tr} \boldsymbol{e}, \quad \forall\, \underline{\boldsymbol{d}}, \underline{\boldsymbol{e}} \in \mathbf{X}. \tag{3.3}$$

Thanks to the assumptions on the nonlinear permeability, we can infer that this form is continuous, as well as coercive over all of $\mathbf{X}$

$$\left| a_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}}) \right| \le C_a \|\underline{\boldsymbol{d}}\|_{\mathbf{X}} \|\underline{\boldsymbol{e}}\|_{\mathbf{X}} \qquad \forall\, \underline{\boldsymbol{d}}, \underline{\boldsymbol{e}} \in \mathbf{X}, \tag{3.4}$$

$$a_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{e}}, \underline{\boldsymbol{e}}) \ge c_a \|\underline{\boldsymbol{e}}\|_{\mathbf{X}}^2 \qquad \forall\, \underline{\boldsymbol{e}} \in \mathbf{X}, \tag{3.5}$$

with

$$C_a := \max\{2\mu + d\,\lambda,\, c_0,\, d\,\alpha,\, \kappa_2\, r\} > 0 \quad \text{and} \quad c_a := \min\{\kappa_1,\, 2\mu,\, c_0\} > 0 \tag{3.6}$$

Then we define the following fixed-point operator

$$\mathcal{J} : \mathbf{W} \subseteq \mathbf{X} \to \mathbf{X}, \quad \underline{\boldsymbol{w}} \mapsto \mathcal{J}(\underline{\boldsymbol{w}}) := \underline{\boldsymbol{d}}, \tag{3.7}$$

where given $\underline{\boldsymbol{w}} = (\boldsymbol{w}, s) \in \mathbf{W}$, $\mathcal{J}(\underline{\boldsymbol{w}}) = \underline{\boldsymbol{d}} = (\boldsymbol{d}, p) \in \mathbf{X}$ is the first component of the solution of the linearised version of problem (2.3): Find $((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \boldsymbol{u}) \in \mathbf{X} \times \mathbf{Y} \times \mathbf{Z}$ such that

$$\begin{aligned} a_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}}) \quad &+ \quad b_1(\underline{\boldsymbol{e}}, \boldsymbol{\sigma}) \qquad\qquad\qquad\quad &&= G(\underline{\boldsymbol{e}}), \\ b_1(\underline{\boldsymbol{d}}, \boldsymbol{\tau}) \qquad\qquad &+ \quad b_2((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \boldsymbol{u}) \quad &&= H(\boldsymbol{\tau}), \\ b_2((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \boldsymbol{v}) \qquad\qquad\quad &\qquad\qquad\qquad\quad &&= F(\boldsymbol{v}), \end{aligned} \tag{3.8}$$

for all $((\underline{e},\tau),v)\in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$.

It is clear that $((\underline{d},\sigma),u)$ is a solution to (2.3) if and only if $\underline{d}$ satisfies $\mathcal{J}(\underline{d})=\underline{d}$, and consequently, the wellposedness of (2.3) is equivalent to the unique solvability of the fixed-point problem: Find $\underline{d}\in\mathbf{W}$ such that

$$\mathcal{J}(\underline{d})=\underline{d}. \tag{3.9}$$

In this way, in what follows we focus on proving the unique solvability of (3.9). According to the definition of $\mathcal{J}$ (cf. (3.7)), it is clear that proving that this operator is well-defined amounts to prove that problem (3.8) is wellposed.

With that in mind, let us define the bilinear form $A_{\underline{w}}:(\mathbf{X}\times\mathbf{Y}_0)\times(\mathbf{X}\times\mathbf{Y}_0)\to\mathbb{R}$ as

$$A_{\underline{w}}((\underline{d},\sigma),(\underline{e},\tau)):=a_{\underline{w}}(\underline{d},\underline{e})+b_1(\underline{e},\sigma)+b_1(\underline{d},\tau), \tag{3.10}$$

and we state the unique solvability of the linearised problem (3.8), depending on a smallness of data assumption, as follows.

Lemma 3.1. *Given $r>0$, let us assume that*

$$\frac{\gamma_1}{r}\left(\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|u_\Gamma\|_{1/2,\Gamma}+\|f\|_{0,\Omega}\right)\le 1, \tag{3.11}$$

where

$$\gamma_1:=\frac{(C_a+1+\beta_2+\gamma_2)^2}{\beta_2^2\,\gamma_2}\quad\text{and}\quad\gamma_2:=\frac{(C_a+2)(c_a+1+C_a)}{c_a}. \tag{3.12}$$

Then, for a given $\underline{w}\in\mathbf{W}$ (cf. (3.2)), there exists a unique $\underline{d}\in\mathbf{W}$ such that $\mathcal{J}(\underline{w})=\underline{d}$.

Proof. From the properties of $a_{\underline{w}}$ and b_1, (3.4), (3.5), (2.9) and (2.4a), we have that the bilinear form $A_{\underline{w}}$ induces an invertible operator on the kernel of the bilinear form b_2, $\mathbf{X}\times\mathbf{Y}_0$ (cf. (2.6)) (see also, e.g., [43,44]). Then, from the inf-sup condition of b_2 (2.8), and a straightforward application of the Babuška–Brezzi theory we have that there exists a unique $((\underline{d},\sigma),u)\in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$ solution to (3.8), or equivalently, the existence of a unique $\underline{d}\in\mathbf{X}$ such that $\mathcal{J}(\underline{w})=\underline{d}$. Finally, from [44, Proposition 2.36], together with (2.5), we readily obtain that

$$\begin{aligned}&\|((\underline{d},\sigma),u\|_{(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}}\\&\quad\le\gamma_1\sup_{\mathbf{0}\neq((\underline{e},\tau),v)\in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}}\frac{A_{\underline{w}}((\underline{d},\sigma),(\underline{e},\tau))+b_2((\underline{d},\sigma),v)+b_2((\underline{e},\tau),u)}{\|((\underline{e},\tau),v\|}\\&\quad\le\gamma_1\left(\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|u_\Gamma\|_{1/2,\Gamma}+\|f\|_{0,\Omega}\right),\end{aligned} \tag{3.13}$$

and after invoking assumption (3.11), the bounds above imply that $\underline{d}$ belongs to $\mathbf{W}$, therefore completing the proof. □

3.3. Wellposedness of the continuous problem

Here, we provide the main result of this section, namely, the existence and uniqueness of solution of the nonlinear problem (2.3). This result is established in the following theorem.

Theorem 3.1. *Let $f\in\mathbf{L}^2(\Omega)$, $g\in L^2(\Omega)$, $u_\Gamma\in\mathbf{H}^{1/2}(\Gamma)$ and $r_\Gamma\in\mathrm{H}^{-1/2}(\Gamma)$ such that*

$$\frac{\gamma_1}{r}\max\{\gamma_1\,\kappa_2\,r,1\}\left(\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|u_\Gamma\|_{1/2,\Gamma}+\|f\|_{0,\Omega}\right)<1, \tag{3.14}$$

where γ_1 is defined in (3.12). Then, the operator $\mathcal{J}$ (cf. (3.7)) has a unique fixed point $\underline{d}\in\mathbf{W}$. Equivalently, the problem (2.3) has a unique solution $((\underline{d},\sigma),u)\in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$ with $\underline{d}\in\mathbf{W}$. In addition, we have the following continuous dependence on data

$$\|((\underline{d},\sigma),u\|\lesssim\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|u_\Gamma\|_{1/2,\Gamma}+\|f\|_{0,\Omega}. \tag{3.15}$$

Proof. We begin by recalling from the previous analysis that assumption (3.14) ensures the well-definiteness of $\mathcal{J}$. Now, let $\underline{w}_1=(w_1,s_1)$, $\underline{w}_2=(w_2,s_2)$, $\underline{d}_1=(d_1,p_1)$, $\underline{d}_2=(d_2,p_2)\in\mathbf{W}$, be such that $\mathcal{J}(\underline{w}_1)=\underline{d}_1$ and $\mathcal{J}(\underline{w}_2)=\underline{d}_2$. According to the definition of $\mathcal{J}$ (cf. (3.8)), it follows that there exist (σ_1,u_1), $(\sigma_2,u_2)\in\mathbf{Y}\times\mathbf{Z}$, such that for all $((\underline{e},\tau),v)\in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$, there hold

$$\begin{aligned}A_{\underline{w}_1}((\underline{d}_1,\sigma_1),(\underline{e},\tau))+b_2((\underline{e},\tau),u_1)+b_2((\underline{d}_1,\sigma_1),v)&=G(\underline{e})+H(\tau)+F(v),\\A_{\underline{w}_2}((\underline{d}_2,\sigma_2),(\underline{e},\tau))+b_2((\underline{e},\tau),u_2)+b_2((\underline{d}_2,\sigma_2),v)&=G(\underline{e})+H(\tau)+F(v).\end{aligned}$$

Then, subtracting both equations, adding and subtracting suitable terms, we easily arrive at

$$\begin{aligned}&A_{\underline{w}_1}((\underline{d}_1-\underline{d}_2,\sigma_1-\sigma_2),(\underline{e},\tau))+b_2((\underline{e},\tau),u_1-u_2)+b_2((\underline{d}_1-\underline{d}_2,\sigma_1-\sigma_2),v)\\&\quad=\int_\Omega(\kappa(\underline{w}_2)-\kappa(\underline{w}_1))\nabla p_2\cdot\nabla q.\end{aligned} \tag{3.16}$$

Therefore, recalling that $\underline{\boldsymbol{w}}_1 \in \mathbf{W}$, we can use the latter identity, the bound (3.13), and the assumptions of κ (cf. (3.1)), to obtain

$$\begin{aligned}
\|\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2\|_{\mathbf{X}} &\leq \|(\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2, \boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2\|_{(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}} \\
&\leq \gamma_1 \sup_{\substack{\mathbf{0}\neq((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{v}) \\ \in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}}} \frac{A_{\underline{\boldsymbol{w}}_1}((\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2, \boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2), (\underline{\boldsymbol{e}}, \boldsymbol{\tau})) + b_2((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \boldsymbol{u}_1 - \boldsymbol{u}_2) + b_2((\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2, \boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2), \boldsymbol{v})}{\|(\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \boldsymbol{v}\|} \\
&= \gamma_1 \sup_{\substack{\mathbf{0}\neq((\underline{\boldsymbol{e}},\boldsymbol{\tau}),\boldsymbol{v}) \\ \in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}}} \frac{\displaystyle\int_\Omega (\kappa(\underline{\boldsymbol{w}}_2) - \kappa(\underline{\boldsymbol{w}}_1))\nabla p_2 \cdot \nabla q}{\|(\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \boldsymbol{v}\|} \\
&\leq \gamma_1 \, \|\kappa(\underline{\boldsymbol{w}}_2) - \kappa(\underline{\boldsymbol{w}}_1)\|_{\mathbb{L}^\infty(\Omega)} \|\nabla p_2\|_{0,\Omega},
\end{aligned}$$

then, recalling that $\underline{\boldsymbol{w}}_1 = (\boldsymbol{w}_1, s_1)$ and $\underline{\boldsymbol{w}}_2 = (\boldsymbol{w}_2, s_2)$, and using the Lipschitz continuity of κ (cf. (3.1)), along with the fact that $\mathcal{J}(\underline{\boldsymbol{w}}_2) = \underline{\boldsymbol{d}}_2 = (\boldsymbol{d}_2, p_2) \in \mathbf{W}$, we have that the estimate (3.13) is satisfied. This implies that

$$\begin{aligned}
\|\mathcal{J}(\underline{\boldsymbol{w}}_1) - \mathcal{J}(\underline{\boldsymbol{w}}_2)\|_{\mathbf{X}} &= \|\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2\|_{\mathbf{X}} \\
&\leq \gamma_1 \, \|\kappa(\underline{\boldsymbol{w}}_2) - \kappa(\underline{\boldsymbol{w}}_1)\|_{\mathbb{L}^\infty(\Omega)} \|\nabla p_2\|_{0,\Omega} \\
&\leq \gamma_1 \, \kappa_2 \, \|s_2 - s_1\|_{1,\Omega} \, \gamma_1 \left(\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega} \right) \\
&\leq \gamma_1^2 \, \kappa_2 \left(\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega} \right) \|\underline{\boldsymbol{w}}_1 - \underline{\boldsymbol{w}}_2\|_{\mathbf{X}}.
\end{aligned}$$

The latter bound, in combination with the assumption (3.14) and the Banach fixed-point theorem, implies that $\mathcal{J}$ has a unique fixed point in $\mathbf{W}$. Equivalently, this result yields that there exists a unique $((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \boldsymbol{u}) \in (\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$ solution to (2.3). Finally, estimate (3.15) is obtained analogously to (3.13), which completes the proof. □

Remark 3.1. In [40], the authors introduce a technique similar to the one employed in Section 3 for a variational formulation of the time-dependent problem. They define a fixed-point operator associated with the linearised version of the problem to establish the existence of a solution. This is accomplished under the assumption of uniform boundedness of the permeability, instead of relying on the assumption (3.1). The proof makes use of the Schauder fixed-point theorem and the Lions–Aubin compactness theorem, diverging from the application of Banach's fixed-point theorem.

4. Finite element discretisation

Let us consider a regular partition $\mathcal{T}_h$ of $\bar{\Omega}$ made up of triangles K (in $\mathbb{R}^2$) or tetrahedra K (in $\mathbb{R}^3$) of diameter h_K, and denote the mesh size by $h := \max\{h_K : K \in \mathcal{T}_h\}$. We will start by defining finite-dimensional subspaces $\mathbf{X}_h$, $\mathbf{Y}_h$, $\mathbf{Z}_h$, of the functional spaces encountered before.

Given an integer $\ell \geq 0$ and $K \in \mathcal{T}_h$, we first let $\mathrm{P}_\ell(K)$ be the space of polynomials of degree $\leq \ell$ defined on K, whose vector and tensor versions are denoted $\boldsymbol{P}_\ell(K) := [\mathrm{P}_\ell(K)]^d$ and $\mathbb{P}_\ell(K) = [\mathrm{P}_\ell(K)]^{d\times d}$, respectively. Also, we let $\mathbf{RT}_\ell(K) := \boldsymbol{P}_\ell(K) \oplus \mathrm{P}_\ell(K)\boldsymbol{x}$ be the local Raviart–Thomas space of order ℓ defined on K, where $\boldsymbol{x}$ stands for a generic vector in $\mathbb{R}^d$.

4.1. Finite element spaces and definition of the Galerkin scheme

First, for fluid pressure we take Lagrangian elements as follows

$$\mathbf{X}_{2,h} := \left\{ q_h \in \mathrm{C}(\overline{\Omega}) : \quad q_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall K \in \mathcal{T}_h \right\}. \tag{4.1}$$

Arnold–Winther finite elements are defined in [34] for $k \geq 1$ and for the 2D case. The lowest-order conforming space for poroelastic stress (and here also for strain) consists of piecewise $\mathbb{P}_2$ tensors enriched with cubic shape functions, and piecewise $\mathbb{P}_1$ vectors for displacement:

$$\begin{aligned}
\mathbf{Y}_h &:= \left\{ \boldsymbol{\tau}_h \in \mathbb{H}_{\mathrm{sym}}(\mathbf{div}; \Omega) : \ \boldsymbol{\tau}_h|_K \in \mathbb{P}_{k+2}(K) \text{ and } \mathbf{div}\, \boldsymbol{\tau}_h|_K \in \mathbf{P}_k(K) \quad \forall K \in \mathcal{T}_h \right\}, \\
\mathbf{Z}_h &:= \left\{ \boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \ \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall K \in \mathcal{T}_h \right\}.
\end{aligned} \tag{4.2}$$

Note that a non-conforming version is also given in [34] but it gives an unbalanced approximation error for displacement and stress and we therefore keep only the conforming version. An appropriate interpolation operator (bounded, with suitable approximability and commutation properties) is constructed in [34], which thanks to Fortin's Lemma (cf. [45, Lemma 2.6]), imply a discrete inf-sup condition for $b_2(\cdot,\cdot)$ (see also [33]).

As announced, the following space for discrete strains is considered

$$\mathbf{X}_{1,h} := \left\{ \boldsymbol{\tau}_h \in \mathbb{L}^2_{\mathrm{sym}}(\Omega) : \ \boldsymbol{\tau}_h|_K \in \mathbb{P}_{k+2}(K) \text{ and } \mathbf{div}\, \boldsymbol{\tau}_h|_K \in \mathbf{P}_k(K) \quad \forall K \in \mathcal{T}_h \right\}. \tag{4.3}$$

Then, defining the product space $\mathbf{X}_h := \mathbf{X}_{1,h} \times \mathbf{X}_{2,h}$, we note that the finite element subspaces $(\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$ are inf-sup stable for the bilinear form b_2 (cf. [34])

$$\sup_{\mathbf{0}\neq(\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h)\in\mathbf{X}_h\times\mathbf{Y}_h} \frac{b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{v}_h)}{\|(\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h)\|_{\mathbf{X}\times\mathbf{Y}}} \geq \beta^*_{b_2} \, \|\boldsymbol{v}_h\|_{\mathbf{Z}} \quad \forall \boldsymbol{v}_h \in \mathbf{Z}_h. \tag{4.4}$$

In addition, it is straightforward to see that the kernel of the bilinear form b_2 can be characterised by

$$\mathbf{X}_h \times \mathbf{Y}_{h,0}, \quad \text{with} \quad \mathbf{Y}_{h,0} = \{\boldsymbol{\tau}_h \in \mathbf{Y}_h \,:\, \mathbf{div}\,\boldsymbol{\tau}_h = \mathbf{0}\}, \tag{4.5}$$

and, for all $\boldsymbol{\tau}_h \in \mathbf{Y}_{h,0}$, it is clear that $\boldsymbol{\tau}_h \in \mathbf{X}_{1,h}$, then we can take $\boldsymbol{e}_h = \boldsymbol{\tau}_h$, and thus b_1 satisfies the inf-sup condition

$$\sup_{\mathbf{0}\neq\underline{\boldsymbol{e}}_h\in\mathbf{X}_h} \frac{b_1(\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h)}{\|\underline{\boldsymbol{e}}_h\|_{\mathbf{X}}} \geq \|\boldsymbol{\tau}_h\|_{\mathbf{Y}} \quad \forall\, \boldsymbol{\tau}_h \in \mathbf{Y}_{0,h}. \tag{4.6}$$

The Galerkin scheme associated with the weak formulation (2.3) consists in finding $((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$ such that

$$\begin{aligned} a(\underline{\boldsymbol{d}}_h, \underline{\boldsymbol{e}}_h) \quad + \quad b_1(\underline{\boldsymbol{e}}_h, \boldsymbol{\sigma}_h) \qquad\qquad\qquad\qquad &= G(\underline{\boldsymbol{e}}_h), \\ b_1(\underline{\boldsymbol{d}}_h, \boldsymbol{\tau}_h) \qquad\qquad + \quad b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{u}_h) \quad &= H(\boldsymbol{\tau}_h), \\ b_2((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{v}_h) \qquad\qquad\qquad\qquad &= F(\boldsymbol{v}_h), \end{aligned} \tag{4.7}$$

for all $((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{v}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$, with $\underline{\boldsymbol{d}}_h = (\boldsymbol{d}_h, p_h)$ and $\underline{\boldsymbol{e}}_h = (\boldsymbol{e}_h, q_h)$.

Remark 4.1. There are several non-conforming finite element methods for the considered mixed formulation of the elasticity problem (see, e.g., [46–48]). For the simplicity of presentation and implementation, we only consider conforming finite element methods in the present study.

4.2. Unique solvability of the discrete problem

In this section we analyse the Galerkin scheme (4.7). We want to emphasise that the analysis of wellposedness can be easily accomplished by applying the results obtained for the continuous problem to the discrete scenario, which is why most of the specific details can be excluded.

Firstly, and similarly to the continuous case, we define the following set

$$\mathbf{W}_h := \left\{ \underline{\boldsymbol{w}}_h := (\boldsymbol{w}_h, s_h) \in \mathbf{X}_h \;:\; \|\underline{\boldsymbol{w}}_h\|_{\mathbf{X}} \leq r \right\}. \tag{4.8}$$

Next, for a fixed $\underline{\boldsymbol{w}}_h := (\boldsymbol{w}_h, s_h)$ in $\mathbf{W}_h$, we have that the bilinear form $a_{\underline{\boldsymbol{w}}}$ defined in (3.3), satisfies:

$$a_{\underline{\boldsymbol{w}}_h}(\underline{\boldsymbol{e}}_h, \underline{\boldsymbol{e}}_h) \geq c_a \|\underline{\boldsymbol{e}}_h\|^2_{\mathbf{X}} \qquad \forall\, \underline{\boldsymbol{e}}_h \in \mathbf{X}_h. \tag{4.9}$$

Then, and again analogously to the continuous case, we define the following fixed-point operator

$$\mathcal{J}_h : \mathbf{W}_h \subseteq \mathbf{X}_h \to \mathbf{X}_h, \quad \underline{\boldsymbol{w}}_h \mapsto \mathcal{J}_h(\underline{\boldsymbol{w}}_h) := \underline{\boldsymbol{d}}_h, \tag{4.10}$$

where, given $\underline{\boldsymbol{w}}_h = (\boldsymbol{w}_h, s_h) \in \mathbf{W}_h$, $\mathcal{J}_h(\underline{\boldsymbol{w}}_h) = \underline{\boldsymbol{d}}_h = (\boldsymbol{d}_h, p_h) \in \mathbf{X}_h$ is the first component of the solution of the linearised version of problem (4.7): Find $((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h) \in \mathbf{X}_h \times \mathbf{Y}_h \times \mathbf{Z}_h$ such that

$$\begin{aligned} a_{\underline{\boldsymbol{w}}_h}(\underline{\boldsymbol{d}}_h, \underline{\boldsymbol{e}}_h) \quad + \quad b_1(\underline{\boldsymbol{e}}_h, \boldsymbol{\sigma}_h) \qquad\qquad\qquad\qquad &= G(\underline{\boldsymbol{e}}_h), \\ b_1(\underline{\boldsymbol{d}}_h, \boldsymbol{\tau}_h) \qquad\qquad + \quad b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{u}_h) \quad &= H(\boldsymbol{\tau}_h), \\ b_2((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{v}_h) \qquad\qquad\qquad\qquad &= F(\boldsymbol{v}_h), \end{aligned} \tag{4.11}$$

for all $((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{v}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$.

It is clear that $((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h)$ is a solution to (4.7) if and only if $\underline{\boldsymbol{d}}_h$ satisfies $\mathcal{J}_h(\underline{\boldsymbol{d}}_h) = \underline{\boldsymbol{d}}_h$, and consequently, the wellposedness of (4.7) is equivalent to the unique solvability of the fixed-point problem: Find $\underline{\boldsymbol{d}}_h \in \mathbf{W}_h$ such that

$$\mathcal{J}_h(\underline{\boldsymbol{d}}_h) = \underline{\boldsymbol{d}}_h. \tag{4.12}$$

In this way, in what follows we focus on proving the unique solvability of (4.12).

Lemma 4.1. *Given $r > 0$, assume that*

$$\frac{\gamma_1^*}{r}\left(\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega}\right) \leq 1, \tag{4.13}$$

where γ_1^* is the discrete version of γ_1 (cf. (3.12)), defined by

$$\gamma_1^* := \frac{(C_a + 1 + \beta_2^* + \gamma_2)^2}{(\beta_2^*)^2\,\gamma_2}, \tag{4.14}$$

with γ_2 defined in (3.12). Then, given $\underline{\boldsymbol{w}}_h \in \mathbf{W}_h$ (cf. (4.8)) there exists a unique $\underline{\boldsymbol{d}}_h \in \mathbf{W}_h$ such that $\mathcal{J}_h(\underline{\boldsymbol{w}}_h) = \underline{\boldsymbol{d}}_h$.

Proof. Given $\underline{\boldsymbol{w}}_h = (\boldsymbol{e}_h, q_h) \in \mathbf{W}_h$, we proceed analogously to the proof of Lemma 3.1 and utilise (2.4a), (3.4), (4.4), (4.6), (4.9) and [44, Proposition 2.36] to deduce the discrete inf-sup condition

$$\begin{aligned} &\|((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h\| \\ &\quad \leq \gamma_1^* \sup_{\mathbf{0}\neq((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\boldsymbol{v}_h)\in(\mathbf{X}_h\times\mathbf{Y}_h)\times\mathbf{Z}_h} \frac{A_{\underline{\boldsymbol{w}}_h}((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), (\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h)) + b_2((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{v}_h) + b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{u}_h)}{\|(\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{v}_h\|}. \end{aligned} \tag{4.15}$$

Therefore, owing to the fact that for finite dimensional linear problems surjectivity and injectivity are equivalent, from (4.15) and the Banach–Nečas–Babuška theorem we obtain that there exists a unique $((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h) \in \mathbf{X}_h \times \mathbf{Y}_h \times \mathbf{Z}_h$ satisfying (4.11), with $\underline{\boldsymbol{d}}_h \in \mathbf{W}_h$, which concludes the proof. □

The following theorem provides the main result of this section, namely, existence and uniqueness of solution to the fixed-point problem (4.12), or equivalently, the wellposedness of problem (4.7).

Theorem 4.1. *Let $\boldsymbol{f} \in \mathbf{L}^2(\Omega)$, $g \in L^2(\Omega)$, $\boldsymbol{u}_\Gamma \in \mathbf{H}^{1/2}(\Gamma)$ and $r_\Gamma \in \mathrm{H}^{-1/2}(\Gamma)$ such that*

$$\frac{\gamma_1^*}{r} \max\{\gamma_1^* \kappa_2 r, 1\} \left(\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega} \right) < 1, \tag{4.16}$$

where γ_1^* is defined in (4.14). Then, the operator $\mathcal{J}_h$ (cf. (4.10)) has a unique fixed point $\underline{\boldsymbol{d}}_h \in \mathbf{W}_h$. Equivalently, the problem (4.7) has a unique solution $((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$ with $\underline{\boldsymbol{d}}_h \in \mathbf{W}_h$. In addition, there exists $C^* > 0$ such that

$$\|(\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h\| \leq C^* \left(\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega} \right). \tag{4.17}$$

Proof. First we observe that, as for the continuous case, assumption (4.16) ensures the well-definiteness of the operator $\mathcal{J}_h$. Now, adapting the arguments utilised in Theorem 3.1 one can obtain the following estimate

$$\begin{aligned} \|\mathcal{J}(\underline{\boldsymbol{w}}_1) - \mathcal{J}(\underline{\boldsymbol{w}}_2)\|_{\mathbf{X}} &= \|\underline{\boldsymbol{d}}_1 - \underline{\boldsymbol{d}}_2\|_{\mathbf{X}} \\ &\leq (\gamma_1^*)^2 \kappa_2 \left(\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega} \right) \|\underline{\boldsymbol{w}}_1 - \underline{\boldsymbol{w}}_2\|_{\mathbf{X}}. \end{aligned}$$

for all $\underline{\boldsymbol{w}}_1, \underline{\boldsymbol{w}}_2 \in \mathbf{W}_h$. In this way, using estimate (4.16) we obtain that $\mathcal{J}_h$ is a contraction mapping on $\mathbf{W}_h$, thus problem (4.12), or equivalently (4.7) is wellposed. Finally, estimate (4.17) is obtained analogously to (3.13), which completes the proof. □

5. A priori error estimates

In this section, we aim to provide the convergence of the Galerkin scheme (4.7) and derive the corresponding rate of convergence.

5.1. Preliminaries

From now on we assume that the hypotheses of Theorem 3.1 and Theorem 4.1 hold and let $((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \boldsymbol{u}) \in (\mathbf{X} \times \mathbf{Y}) \times \mathbf{Z}$ and $((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{u}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$ be the unique solutions of (2.3) and (4.7), respectively.

Then, similarly to [31], in order to simplify the subsequent analysis, we write $\mathrm{e}_{\underline{\boldsymbol{d}}} = \underline{\boldsymbol{d}} - \underline{\boldsymbol{d}}_h$, $\mathrm{e}_{\boldsymbol{\sigma}} = \boldsymbol{\sigma} - \boldsymbol{\sigma}_h$ and $\mathrm{e}_{\boldsymbol{u}} = \boldsymbol{u} - \boldsymbol{u}_h$. As usual, for a given $(\widehat{\underline{\boldsymbol{e}}}_h, \widehat{\boldsymbol{\sigma}}_h) \in \mathbf{X}_h \times \mathbf{Y}_h$ and $\widehat{\boldsymbol{v}}_h \in \mathbf{Z}_h$, we shall then decompose these errors into

$$\mathrm{e}_{\underline{\boldsymbol{d}}} = \boldsymbol{\xi}_{\underline{\boldsymbol{d}}} + \boldsymbol{\chi}_{\underline{\boldsymbol{d}}}, \qquad \mathrm{e}_{\boldsymbol{\sigma}} = \boldsymbol{\xi}_{\boldsymbol{\sigma}} + \boldsymbol{\chi}_{\boldsymbol{\sigma}}, \qquad \mathrm{e}_{\boldsymbol{u}} = \boldsymbol{\xi}_{\boldsymbol{u}} + \boldsymbol{\chi}_{\boldsymbol{u}}, \tag{5.1}$$

with

$$\boldsymbol{\xi}_{\underline{\boldsymbol{d}}} := \underline{\boldsymbol{d}} - \widehat{\underline{\boldsymbol{e}}}_h, \quad \boldsymbol{\chi}_{\underline{\boldsymbol{d}}} := \widehat{\underline{\boldsymbol{e}}}_h - \underline{\boldsymbol{d}}_h, \quad \boldsymbol{\xi}_{\boldsymbol{\sigma}} := \boldsymbol{\sigma} - \widehat{\boldsymbol{\tau}}_h, \quad \boldsymbol{\chi}_{\boldsymbol{\sigma}} := \widehat{\boldsymbol{\tau}}_h - \boldsymbol{\sigma}_h, \quad \boldsymbol{\xi}_{\boldsymbol{u}} := \boldsymbol{u} - \widehat{\boldsymbol{v}}_h, \quad \boldsymbol{\chi}_{\boldsymbol{u}} := \widehat{\boldsymbol{v}}_h - \boldsymbol{u}_h.$$

Recalling the definition of the bilinear form $A_{\underline{\boldsymbol{w}}}$ in (3.10), from (2.3) and (4.7) we have that the following identities hold

$$A_{\underline{\boldsymbol{d}}}((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), (\underline{\boldsymbol{e}}, \boldsymbol{\tau})) + b_2((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \boldsymbol{u}) + b_2((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \boldsymbol{v}) = G(\underline{\boldsymbol{e}}) + H(\boldsymbol{\tau}) + F(\boldsymbol{v}),$$

for all $((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \boldsymbol{v}) \in (\mathbf{X} \times \mathbf{Y}) \times \mathbf{Z}$, and

$$A_{\underline{\boldsymbol{d}}_h}((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), (\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h)) + b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{u}_h) + b_2((\underline{\boldsymbol{d}}_h, \boldsymbol{\sigma}_h), \boldsymbol{v}_h) = G(\underline{\boldsymbol{e}}_h) + H(\boldsymbol{\tau}_h) + F(\boldsymbol{v}_h).$$

for all $((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{v}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$. From these relations, and similarly to (3.16), we can obtain that for all $((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{v}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$, there holds

$$A_{\underline{\boldsymbol{d}}_h}((\mathrm{e}_{\underline{\boldsymbol{d}}}, \mathrm{e}_{\boldsymbol{\sigma}}), (\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h)) + b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \mathrm{e}_{\boldsymbol{u}}) + b_2((\mathrm{e}_{\underline{\boldsymbol{d}}}, \mathrm{e}_{\boldsymbol{\sigma}}), \boldsymbol{v}_h) = \int_\Omega (\kappa(\underline{\boldsymbol{d}}_h) - \kappa(\underline{\boldsymbol{d}})) \nabla p \cdot \nabla q_h,$$

which together with the definition of the errors in (5.1), implies that

$$\begin{aligned} &A_{\underline{\boldsymbol{d}}_h}((\boldsymbol{\chi}_{\underline{\boldsymbol{d}}}, \boldsymbol{\chi}_{\boldsymbol{\sigma}}), (\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h)) + b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{\chi}_{\boldsymbol{u}}) + b_2((\boldsymbol{\chi}_{\underline{\boldsymbol{d}}}, \boldsymbol{\chi}_{\boldsymbol{\sigma}}), \boldsymbol{v}_h) \\ &\quad = -A_{\underline{\boldsymbol{d}}_h}((\boldsymbol{\xi}_{\underline{\boldsymbol{d}}}, \boldsymbol{\xi}_{\boldsymbol{\sigma}}), (\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h)) - b_2((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{\xi}_{\boldsymbol{u}}) - b_2((\boldsymbol{\xi}_{\underline{\boldsymbol{d}}}, \boldsymbol{\xi}_{\boldsymbol{\sigma}}), \boldsymbol{v}_h) + \int_\Omega (\kappa(\underline{\boldsymbol{d}}_h) - \kappa(\underline{\boldsymbol{d}})) \nabla p \cdot \nabla q_h, \end{aligned} \tag{5.2}$$

for all $((\underline{\boldsymbol{e}}_h, \boldsymbol{\tau}_h), \boldsymbol{v}_h) \in (\mathbf{X}_h \times \mathbf{Y}_h) \times \mathbf{Z}_h$. Then, since $\underline{\boldsymbol{d}}_h \in \mathbf{W}_h$, we apply the discrete inf-sup condition (4.15) at the left-hand side of (5.2) followed by the continuity properties of a, b_1 and b_2 (cf. (3.4) and (2.4a)) on the right-hand side of (5.2), to obtain

$$\begin{aligned} &\|\boldsymbol{\chi}_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}} + \|\boldsymbol{\chi}_{\boldsymbol{\sigma}}\|_{\mathbf{Y}} + \|\boldsymbol{\chi}_{\boldsymbol{u}}\|_{\mathbf{Z}} \\ &\quad \leq \gamma_1^* \left((C_a + 2)\left(\|\boldsymbol{\xi}_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}} + \|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|_{\mathbf{Y}} + \|\boldsymbol{\xi}_{\boldsymbol{u}}\|_{\mathbf{Z}} \right) + \kappa_2 \|p_h - p\|_{1,\Omega} \|\nabla p\|_{0,\Omega} \right) \\ &\quad \leq \gamma_1^* \left((C_a + 2)\left(\|\boldsymbol{\xi}_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}} + \|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|_{\mathbf{Y}} + \|\boldsymbol{\xi}_{\boldsymbol{u}}\|_{\mathbf{Z}} \right) + \kappa_2 (\|\boldsymbol{\xi}_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}} + \|\boldsymbol{\chi}_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}}) \|\underline{\boldsymbol{d}}\|_{\mathbf{X}} \right). \end{aligned} \tag{5.3}$$

5.2. Derivation of Céa estimates

Now we turn to providing a best approximation estimate corresponding with the Galerkin scheme (4.7).

Theorem 5.1. *Assume that*

$$\kappa_2\gamma_1^*\gamma_1\left(\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma}+\|\boldsymbol{f}\|_{0,\Omega}\right)\leq\frac{1}{2}, \tag{5.4}$$

with γ_1 and γ_1^* being the constants in (3.12) and (4.14). Furthermore, assume that the hypotheses of Theorem 3.1 and Theorem 4.1 hold. Let $((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\boldsymbol{u})\in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$ and $((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\boldsymbol{u}_h)\in(\mathbf{X}_h\times\mathbf{Y}_h)\times\mathbf{Z}_h$ be the unique solutions of (2.3) and (4.7), respectively. Then, there exists $C_{Céa}>0$, such that

$$\|((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\boldsymbol{u})-((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\boldsymbol{u}_h)\|\leq C_{Céa}\inf_{((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\boldsymbol{v}_h)\in(\mathbf{X}_h\times\mathbf{Y}_h)\times\mathbf{Z}_h}\|((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\boldsymbol{u})-((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\boldsymbol{v}_h)\|. \tag{5.5}$$

Proof. From (5.3) we have

$$\begin{aligned}&\|\chi_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}}\left(1-\kappa_2\,\gamma_1^*\,\|\underline{\boldsymbol{d}}\|_{\mathbf{X}}\right)+\|\chi_{\boldsymbol{\sigma}}\|_{\mathbf{Y}}+\|\chi_{\boldsymbol{u}}\|_{\mathbf{Z}}\\&\quad\leq\gamma_1^*\left((C_a+2)\left(\|\xi_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}}+\|\xi_{\boldsymbol{\sigma}}\|_{\mathbf{Y}}+\|\xi_{\boldsymbol{u}}\|_{\mathbf{Z}}\right)+\kappa_2\|\xi_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}}\|\underline{\boldsymbol{d}}\|_{\mathbf{X}}\right).\end{aligned} \tag{5.6}$$

Hence, using the fact that $\underline{\boldsymbol{d}}$ satisfies (3.13), from assumption (5.4) and the latter inequality, we obtain

$$\|\chi_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}}+\|\chi_{\boldsymbol{\sigma}}\|_{\mathbf{Y}}+\|\chi_{\boldsymbol{u}}\|_{\mathbf{Z}}\leq C\left(\|\xi_{\underline{\boldsymbol{d}}}\|_{\mathbf{X}}+\|\xi_{\boldsymbol{\sigma}}\|_{\mathbf{Y}}+\|\xi_{\boldsymbol{\sigma}}\|_{\mathbf{Z}}\right), \tag{5.7}$$

with $C>0$ independent of h. In this way, from (5.1), (5.7) and the triangle inequality we obtain

$$\|((\mathrm{e}_{\underline{\boldsymbol{d}}},\mathrm{e}_{\boldsymbol{\sigma}}),\mathrm{e}_{\boldsymbol{u}})\|\leq\|((\chi_{\underline{\boldsymbol{d}}},\chi_{\boldsymbol{\sigma}}),\chi_{\boldsymbol{u}})\|+\|((\xi_{\underline{\boldsymbol{d}}},\xi_{\boldsymbol{\sigma}}),\xi_{\boldsymbol{u}})\|\leq(C+1)\|((\xi_{\underline{\boldsymbol{d}}},\xi_{\boldsymbol{\sigma}}),\xi_{\boldsymbol{u}})\|,$$

which combined with the fact that $(\widehat{\underline{\boldsymbol{e}}}_h,\widehat{\boldsymbol{\sigma}}_h)\in\mathbf{X}_h\times\mathbf{Y}_h$ and $\widehat{\boldsymbol{v}}_h\in\mathbf{Z}_h$ are arbitrary, concludes the proof. □

Remark 5.1. We note that, alternatively to the development in Theorem 5.1 above, we could proceed as in the proof of [44, Lemma 2.25] and apply a Strang-type argument to obtain Céa's estimate. In addition, using the estimates (3.13) and (5.6) together with the assumption (5.4), we can obtain the following estimate for the Céa's constant $C_{\text{Céa}}=2(\gamma_1^*(C_a+2)+1)$.

5.3. Rates of convergence

In order to establish the rate of convergence of the Galerkin scheme (4.7), we first recall the following approximation properties **AP** (interpolation estimates of Sobolev spaces for the twofold saddle-points for the poroelastic stress symmetry imposed strongly) associated with the finite element spaces specified in Section 4.1.

($\mathbf{AP}_h^{\boldsymbol{d}}$) For each $1\leq m\leq k+2$ and for each $\boldsymbol{e}\in\mathbb{H}^m(\Omega)\cap\mathbb{L}^2_{\text{sym}}(\Omega)$, there holds

$$\operatorname{dist}\left(\boldsymbol{e},\mathbf{X}_{1,h}\right):=\inf_{\boldsymbol{e}_h\in\mathbf{X}_{1,h}}\|\boldsymbol{e}-\boldsymbol{e}_h\|_{0,\Omega}\lesssim h^m\,\|\boldsymbol{e}\|_{m,\Omega}. \tag{5.8a}$$

($\mathbf{AP}_h^{p}$) For each $0\leq m\leq k+1$ and for each $q\in\mathrm{H}^{m+1}(\Omega)$, there holds

$$\operatorname{dist}\left(q,\mathbf{X}_{2,h}\right):=\inf_{q_h\in\mathbf{X}_{2,h}}\|q-q_h\|_{1,\Omega}\lesssim h^m\,\|q\|_{m+1,\Omega}. \tag{5.8b}$$

($\mathbf{AP}_h^{\boldsymbol{\sigma}}$) For each $1\leq m\leq k+1$ and for each $\boldsymbol{\tau}\in\mathbb{H}^m(\Omega)\cap\mathbb{H}_{\text{sym}}(\mathbf{div};\Omega)$ with $\mathbf{div}\,\boldsymbol{\tau}\in\mathbf{H}^m(\Omega)$, there holds

$$\operatorname{dist}\left(\boldsymbol{\tau},\mathbf{Y}_h\right):=\inf_{\boldsymbol{\tau}_h\in\mathbf{Y}_h}\|\boldsymbol{\tau}-\boldsymbol{\tau}_h\|_{\mathbf{div},\Omega}\lesssim h^m\left\{\|\boldsymbol{\tau}\|_{m,\Omega}+\|\mathbf{div}\,\boldsymbol{\tau}\|_{m,\Omega}\right\}. \tag{5.8c}$$

($\mathbf{AP}_h^{\boldsymbol{u}}$) For each $1\leq m\leq k+1$ and for each $\boldsymbol{v}\in\mathbf{H}^{m+1}(\Omega)$, there holds

$$\operatorname{dist}\left(\boldsymbol{v},\mathbf{Z}_h\right):=\inf_{\boldsymbol{v}_h\in\mathbf{Z}_h}\|\boldsymbol{v}-\boldsymbol{v}_h\|_{0,\Omega}\lesssim h^m\,\|\boldsymbol{v}\|_{m+1,\Omega}. \tag{5.8d}$$

For (5.8a), (5.8c) and (5.8d) we refer to [34, Theorem 6.1], whereas (5.8b) can be found in [44, Corollary 1.128].

With these steps we are now in position to state the rates of convergence associated with the Galerkin scheme (4.7).

Theorem 5.2. *Assume that the hypotheses of Theorem 5.1 hold and let $((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\boldsymbol{u})\in(\mathbf{X}\times\mathbf{Y})\times\mathbf{Z}$ and $((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\boldsymbol{u}_h)\in(\mathbf{X}_h\times\mathbf{Y}_h)\times\mathbf{Z}_h$ be the unique solutions of the continuous and discrete problems (2.3) and (4.7), respectively. Assume further that $\boldsymbol{d}\in\mathbb{H}^m(\Omega)$, $p\in\mathrm{H}^{m+1}(\Omega)$, $\boldsymbol{\sigma}\in\mathbb{H}^m(\Omega)$, $\mathbf{div}\,\boldsymbol{\sigma}\in\mathbf{H}^m(\Omega)$ and $\boldsymbol{u}\in\mathbf{H}^{m+1}(\Omega)$, for $1\leq m\leq k+1$. Then there exists $C_{rate}>0$, independent of h, such that*

$$\|((\mathrm{e}_{\underline{\boldsymbol{d}}},\mathrm{e}_{\boldsymbol{\sigma}}),\mathrm{e}_{\boldsymbol{u}})\|\leq C_{rate}\,h^m\left\{\|\boldsymbol{d}\|_{m,\Omega}+\|p\|_{m+1,\Omega}+\|\boldsymbol{\sigma}\|_{m,\Omega}+\|\mathbf{div}\,\boldsymbol{\sigma}\|_{m,\Omega}+\|\boldsymbol{u}\|_{m+1,\Omega}\right\}.$$

Proof. The result is a straightforward application of Theorem 5.1 and the approximation properties ($\mathbf{AP}_h^{\boldsymbol{d}}$), ($\mathbf{AP}_h^{p}$), ($\mathbf{AP}_h^{\boldsymbol{\sigma}}$), and ($\mathbf{AP}_h^{\boldsymbol{u}}$). □

6. A five-field mixed formulation

In this section, we present a second formulation, the weak formulation we treat here results from imposing the symmetry of the poroelastic stress in a weak manner (see, e.g., [45] for the general idea and [13] for the application in the context of poroelasticity but leading to a different formulation). We mention in advance that the well-posedness analysis, for both the continuous and discrete problems, follows straightforwardly by adapting the results derived in Sections 3 and 4, reason why most of the details are omitted.

To weakly impose the symmetry of stress, it is customary to introduce the rotation tensor

$$\boldsymbol{\gamma} = \frac{1}{2}(\nabla \boldsymbol{u} - [\nabla \boldsymbol{u}]^{\mathrm{t}}) = \nabla \boldsymbol{u} - \boldsymbol{d}, \tag{6.1}$$

and we can then rewrite the strong form of the coupled PDE system in mixed form as

$$\begin{aligned} &-\mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f} \quad \text{in} \quad \Omega, \qquad \boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathrm{t}} \quad \text{in} \quad \Omega, \qquad \boldsymbol{\gamma} = \nabla \boldsymbol{u} - \boldsymbol{d} \quad \text{in} \quad \Omega, \\ &\boldsymbol{\sigma} = C\boldsymbol{d} - \alpha p \mathbb{I}, \quad \text{in} \quad \Omega, \qquad c_0 p + \alpha \operatorname{tr} \boldsymbol{d} - \operatorname{div}(\kappa(\boldsymbol{d}, p)\nabla p) = g, \quad \text{in} \quad \Omega, \\ &\qquad \boldsymbol{u} = \boldsymbol{u}_\Gamma \quad \text{on} \quad \Gamma, \qquad \kappa(\boldsymbol{d}, p)\nabla p \cdot \boldsymbol{n} = r_\Gamma \quad \text{on} \quad \Gamma. \end{aligned} \tag{6.2}$$

After testing these equations by $\boldsymbol{v} \in \mathbf{L}^2(\Omega)$, $\boldsymbol{\eta} \in \mathbb{L}^2_{\mathrm{skew}}(\Omega)$, $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega)$, $\boldsymbol{e} \in \mathbb{L}^2(\Omega)$, and $q \in \mathrm{H}^1(\Omega)$, respectively; we integrate by parts and use (1.7) as natural boundary conditions to obtain the system

$$\begin{aligned} -\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\,\boldsymbol{\sigma} &= \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} && \forall \boldsymbol{v} \in \mathbf{L}^2(\Omega), \\ -\int_\Omega \boldsymbol{\tau} : \boldsymbol{d} - \int_\Omega \boldsymbol{u} \cdot \mathbf{div}\,\boldsymbol{\tau} - \int_\Omega \boldsymbol{\tau} : \boldsymbol{\gamma} &= -\langle \boldsymbol{\tau}\boldsymbol{n}, \boldsymbol{u}_\Gamma \rangle_\Gamma && \forall \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega), \\ \int_\Omega C\boldsymbol{d} : \boldsymbol{e} - \alpha \int_\Omega p \operatorname{tr} \boldsymbol{e} - \int_\Omega \boldsymbol{\sigma} : \boldsymbol{e} &= 0 && \forall \boldsymbol{e} \in \mathbb{L}^2(\Omega), \\ \int_\Omega \kappa(\boldsymbol{d}, p)\nabla p \cdot \nabla q + c_0 \int_\Omega p\, q + \alpha \int_\Omega q \operatorname{tr} \boldsymbol{d} &= \int_\Omega g\, q + \langle r_\Gamma, q \rangle_\Gamma && \forall q \in \mathrm{H}^1(\Omega), \\ -\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\eta} &= 0 && \forall \boldsymbol{\eta} \in \mathbb{L}^2_{\mathrm{skew}}(\Omega). \end{aligned}$$

Proceeding similarly as in the derivation of (2.3), we group spaces, unknowns and test functions as follows:

$$\widetilde{\mathbf{X}} := \mathbb{L}^2(\Omega) \times \mathrm{H}^1(\Omega), \quad \widetilde{\mathbf{Y}} := \mathbb{H}(\mathbf{div}; \Omega), \quad \widetilde{\mathbf{Z}} := \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega),$$

$$\underline{\boldsymbol{d}} := (\boldsymbol{d}, p) \in \widetilde{\mathbf{X}}, \quad \boldsymbol{\sigma} \in \widetilde{\mathbf{Y}}, \quad \underline{\boldsymbol{u}} := (\boldsymbol{u}, \boldsymbol{\gamma}) \in \widetilde{\mathbf{Z}},$$

$$\underline{\boldsymbol{e}} := (\boldsymbol{e}, q) \in \widetilde{\mathbf{X}}, \quad \boldsymbol{\tau} \in \widetilde{\mathbf{Y}}, \quad \underline{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\eta}) \in \widetilde{\mathbf{Z}},$$

where $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{Y}}$, $\widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{Z}}$ are endowed with the norms

$$\|\underline{\boldsymbol{e}}\|^2_{\widetilde{\mathbf{X}}} := \|\boldsymbol{e}\|^2_{0,\Omega} + \|q\|^2_{1,\Omega}, \quad \|\boldsymbol{\tau}\|_{\widetilde{\mathbf{Y}}} := \|\boldsymbol{\tau}\|_{\mathbf{div},\Omega}, \quad \|(\underline{\boldsymbol{e}}, \boldsymbol{\tau})\|^2_{\widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}} := \|\underline{\boldsymbol{e}}\|^2_{\widetilde{\mathbf{X}}} + \|\boldsymbol{\tau}\|^2_{\widetilde{\mathbf{Y}}},$$

$$\|\underline{\boldsymbol{v}}\|^2_{\widetilde{\mathbf{Z}}} := \|\boldsymbol{v}\|^2_{0,\Omega} + \|\boldsymbol{\eta}\|^2_{0,\Omega}, \quad \|((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \underline{\boldsymbol{v}})\|^2 := \|(\underline{\boldsymbol{e}}, \boldsymbol{\tau})\|^2_{\widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}} + \|\underline{\boldsymbol{v}}\|^2_{\widetilde{\mathbf{Z}}}.$$

Next, we define the weak forms $\tilde{a} : \widetilde{\mathbf{X}} \times \widetilde{\mathbf{X}} \to \mathbb{R}$, $\tilde{b}_1 : \widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}} \to \mathbb{R}$ and $\tilde{b}_2 : (\widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}) \times \widetilde{\mathbf{Z}} \to \mathbb{R}$ from the expressions

$$\begin{aligned} \tilde{a}(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}}) &:= \int_\Omega C\boldsymbol{d} : \boldsymbol{e} + \int_\Omega \kappa(\underline{\boldsymbol{d}})\nabla p \cdot \nabla q + c_0 \int_\Omega p\, q + \alpha \int_\Omega q \operatorname{tr} \boldsymbol{d} - \alpha \int_\Omega p \operatorname{tr} \boldsymbol{e}, \\ \tilde{b}_1(\underline{\boldsymbol{e}}, \boldsymbol{\tau}) &:= -\int_\Omega \boldsymbol{\tau} : \boldsymbol{e}, \\ \tilde{b}_2((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \underline{\boldsymbol{v}}) &:= -\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\,\boldsymbol{\tau} - \int_\Omega \boldsymbol{\tau} : \boldsymbol{\eta}, \end{aligned} \tag{6.3}$$

respectively, and the linear functionals $\widetilde{G} \in \widetilde{\mathbf{X}}'$, $\widetilde{H} \in \widetilde{\mathbf{Y}}'$ and $\widetilde{F} \in \widetilde{\mathbf{Z}}'$ as

$$\widetilde{G}(\underline{\boldsymbol{e}}) := \int_\Omega g\, q + \langle r_\Gamma, q \rangle_\Gamma, \quad \widetilde{H}(\boldsymbol{\tau}) := -\langle \boldsymbol{\tau}\boldsymbol{n}, \boldsymbol{u}_\Gamma \rangle_\Gamma, \quad \widetilde{F}(\underline{\boldsymbol{v}}) := \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v},$$

respectively, so that the weak formulation of the nonlinear coupled system (6.2) reads: Find $((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \underline{\boldsymbol{u}}) \in (\widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}) \times \widetilde{\mathbf{Z}}$ such that

$$\begin{aligned} \tilde{a}(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}}) \quad + \quad \tilde{b}_1(\underline{\boldsymbol{e}}, \boldsymbol{\sigma}) \qquad\qquad\qquad\quad &= \widetilde{G}(\underline{\boldsymbol{e}}), \\ \tilde{b}_1(\underline{\boldsymbol{d}}, \boldsymbol{\tau}) \qquad\qquad + \quad \tilde{b}_2((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \underline{\boldsymbol{u}}) \quad &= \widetilde{H}(\boldsymbol{\tau}), \\ \tilde{b}_2((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \underline{\boldsymbol{v}}) \qquad\qquad\qquad\quad &= \widetilde{F}(\underline{\boldsymbol{v}}), \end{aligned} \tag{6.4}$$

for all $((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \underline{\boldsymbol{v}}) \in (\widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}) \times \widetilde{\mathbf{Z}}$.

Similarly to Section 2.2, we have that the bilinear forms $\tilde{b}_1$, $\tilde{b}_2$, and the functionals $\widetilde{G}$, $\widetilde{H}$ and $\widetilde{F}$ are bounded. In addition, the kernel of the bilinear form $\tilde{b}_2$ can be characterised by

$$\widetilde{\mathbf{X}} \times \mathbf{Y}_0, \tag{6.5}$$

with $\mathbf{Y}_0$ defined as in (2.7), this is

$$\mathbf{Y}_0 := \{\boldsymbol{\tau} \in \mathbf{Y} : \quad \mathbf{div}\, \boldsymbol{\tau} = \mathbf{0}\} = \{\boldsymbol{\tau} \in \widetilde{\mathbf{Y}} : \quad \mathbf{div}\, \boldsymbol{\tau} = \mathbf{0} \quad \text{and} \quad \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathrm{t}}\}. \tag{6.6}$$

On the other hand, from [45, Section 3.4.3.1] we have that there exists $\beta_{\tilde{b}_2} > 0$ such that

$$\sup_{\mathbf{0} \neq (\underline{\boldsymbol{e}}, \boldsymbol{\tau}) \in \widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}} \frac{\tilde{b}_2((\underline{\boldsymbol{e}}, \boldsymbol{\tau}), \underline{\boldsymbol{v}})}{\|(\underline{\boldsymbol{e}}, \boldsymbol{\tau})\|_{\widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}}}} \geq \beta_{\tilde{b}_2} \|\underline{\boldsymbol{v}}\|_{\widetilde{\mathbf{Z}}} \quad \forall\, \underline{\boldsymbol{v}} \in \widetilde{\mathbf{Z}}. \tag{6.7}$$

In addition, similarly to (2.9), we note that for all $\boldsymbol{\tau} \in \mathbf{Y}_0$, it suffices to take $\boldsymbol{e} = \boldsymbol{\tau}$ to easily arrive at

$$\sup_{\mathbf{0} \neq \underline{\boldsymbol{e}} \in \widetilde{\mathbf{X}}} \frac{\tilde{b}_1(\underline{\boldsymbol{e}}, \boldsymbol{\tau})}{\|\underline{\boldsymbol{e}}\|_{\widetilde{\mathbf{X}}}} \geq \|\boldsymbol{\tau}\|_{\widetilde{\mathbf{Y}}} \quad \forall\, \boldsymbol{\tau} \in \mathbf{Y}_0. \tag{6.8}$$

Next, similarly to Section 3.2, for a given $r > 0$, let us define the bounded set

$$\widetilde{\mathbf{W}} := \left\{ \underline{\boldsymbol{w}} := (\boldsymbol{w}, s) \in \widetilde{\mathbf{X}} : \quad \|\underline{\boldsymbol{w}}\|_{\widetilde{\mathbf{X}}} \leq r \right\}, \tag{6.9}$$

then, for a fixed $\underline{\boldsymbol{w}} := (\boldsymbol{w}, s)$ in $\widetilde{\mathbf{W}}$, we define the bilinear form $\tilde{a}_{\underline{\boldsymbol{w}}} : \widetilde{\mathbf{X}} \times \widetilde{\mathbf{X}} \to \mathbb{R}$ as follows

$$\tilde{a}_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}}) := \int_\Omega C\boldsymbol{d} : \boldsymbol{e} + \int_\Omega \kappa(\underline{\boldsymbol{w}}) \nabla p \cdot \nabla q + c_0 \int_\Omega p\, q + \alpha \int_\Omega q\, \mathrm{tr}\, \boldsymbol{d} - \alpha \int_\Omega p\, \mathrm{tr}\, \boldsymbol{e}, \quad \forall\, \underline{\boldsymbol{d}}, \underline{\boldsymbol{e}} \in \widetilde{\mathbf{X}}. \tag{6.10}$$

Assuming the properties (3.1) hold for the nonlinear permeability, for all $\underline{\boldsymbol{d}}$ and $\underline{\boldsymbol{e}}$ in $\widetilde{\mathbf{X}}$, the following holds

$$\left|\tilde{a}_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{d}}, \underline{\boldsymbol{e}})\right| \leq C_a \|\underline{\boldsymbol{d}}\|_{\widetilde{\mathbf{X}}} \|\underline{\boldsymbol{e}}\|_{\widetilde{\mathbf{X}}} \quad \text{and} \quad \tilde{a}_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{e}}, \underline{\boldsymbol{e}}) \geq c_a \|\underline{\boldsymbol{e}}\|^2_{\widetilde{\mathbf{X}}}, \tag{6.11}$$

with C_a and c_a defined as in (3.6).

By adapting the fixed-point strategy introduced in Section 3.2 to the present case, proceeding similarly to Lemma 3.1 to analyse the linearised system, employing the stability properties of the forms $\tilde{b}_1$, $\tilde{b}_2$ and $\tilde{a}_{\underline{\boldsymbol{w}}}$, it can be proved the well-posedness of (6.4).

Theorem 6.1. *Let $\boldsymbol{f} \in \mathbf{L}^2(\Omega)$, $g \in L^2(\Omega)$, $\boldsymbol{u}_\Gamma \in \mathbf{H}^{1/2}(\Gamma)$ and $r_\Gamma \in \mathrm{H}^{-1/2}(\Gamma)$ such that*

$$\frac{\tilde{\gamma}_1}{r} \max\{\tilde{\gamma}_1 \kappa_2 r, 1\} \left(\|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega} \right) < 1, \tag{6.12}$$

where

$$\tilde{\gamma}_1 := \frac{(C_a + 1 + \tilde{\beta}_2 + \gamma_2)^2}{\tilde{\beta}_2^2\, \gamma_2}, \tag{6.13}$$

with γ_2 defined in (3.12). Then, the problem (6.4) has a unique solution $((\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \underline{\boldsymbol{u}}) \in \widetilde{\mathbf{X}} \times \widetilde{\mathbf{Y}} \times \widetilde{\mathbf{Z}}$ with $\underline{\boldsymbol{d}} \in \widetilde{\mathbf{W}}$. In addition, there holds

$$\|(\underline{\boldsymbol{d}}, \boldsymbol{\sigma}), \underline{\boldsymbol{u}}\| \lesssim \|g\|_{0,\Omega} + \|r_\Gamma\|_{-1/2,\Gamma} + \|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma} + \|\boldsymbol{f}\|_{0,\Omega}. \tag{6.14}$$

Proof. The proof follows using the same steps employed to prove Theorem 3.1. □

For the Galerkin scheme of problem (6.4), we consider the same space $\mathbf{X}_{2,h}$ (cf. (4.1)) for fluid pressure.

For each $K \in \mathcal{T}_h$ we consider the bubble space of order k, defined as

$$\mathbf{B}_k(K) := \begin{cases} \mathbf{curl}^{\mathrm{t}}(b_K \mathrm{P}_k(K)) & \text{in} \quad \mathbb{R}^2, \\ \nabla \times (b_K \mathbf{P}_k(K)) & \text{in} \quad \mathbb{R}^3, \end{cases}$$

where b_K is a suitably normalised cubic polynomial on K, which vanishes on the boundary of K (see [44]).

Next, we recall that the classical PEERS elements are described in [35]:

$$\begin{aligned} \widetilde{\mathbf{Y}}_h &:= \left\{ \boldsymbol{\tau}_h \in \mathbb{H}(\mathbf{div}, \Omega) : \quad \boldsymbol{\tau}_h|_K \in \mathbb{RT}_k(K) \oplus [\mathbf{B}_k(K)]^d \quad \forall\, K \in \mathcal{T}_h \right\}, \\ \widetilde{\mathbf{Z}}_{1,h} &:= \left\{ \boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \\ \widetilde{\mathbf{Z}}_{2,h} &:= \left\{ \boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) \cap \mathbf{C}(\overline{\Omega}) \quad \text{and} \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \end{aligned} \tag{6.15}$$

and are inf-sup stable for the bilinear form $\tilde{b}_2$. In addition, and according to [27], they are inf-sup stable together with the space

$$\widetilde{\mathbf{X}}_{1,h} := \left\{ \boldsymbol{e}_h \in \mathbb{L}^2(\Omega) : \quad \boldsymbol{e}_h|_K \in \mathbb{P}_k(K) \oplus [\mathbf{B}_k(K)]^d \quad \forall\, K \in \mathcal{T}_h \right\}, \tag{6.16}$$

with respect to b_1 (see also [28,30] where also the deviatoric part of the bubble functions is used in the enrichment).

Moreover, Arnold–Falk–Winther finite elements are in [36]:

$$\begin{aligned}\widetilde{\mathbf{Y}}_h &:= \left\{\boldsymbol{\tau}_h \in \mathbb{H}(\mathbf{div},\Omega) : \ \boldsymbol{\tau}_h|_K \in \mathbb{BDM}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h\right\},\\ \widetilde{\mathbf{Z}}_{1,h} &:= \left\{\boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \ \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\right\},\\ \widetilde{\mathbf{Z}}_{2,h} &:= \left\{\boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) : \ \boldsymbol{\eta}_h|_K \in \mathbb{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\right\},\end{aligned} \tag{6.17}$$

and, together with the space

$$\widetilde{\mathbf{X}}_{1,h} := \left\{\boldsymbol{e}_h \in \mathbb{L}^2(\Omega) : \ \boldsymbol{e}_h|_K \in \mathbb{BDM}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h\right\}, \tag{6.18}$$

they are inf-sup stable with respect to the bilinear forms $\tilde{b}_2$ and $\tilde{b}_1$.

In what follows, we will develop the analysis with the subspaces defined in (6.15)–(6.16). The analysis for the subspaces defined in (6.17)–(6.18) is conducted in an analogous manner.

Let us denote the product spaces $\widetilde{\mathbf{X}}_h := \widetilde{\mathbf{X}}_{1,h}\times\mathbf{X}_{2,h}$ and $\widetilde{\mathbf{Z}}_h := \widetilde{\mathbf{Z}}_{1,h}\times\widetilde{\mathbf{Z}}_{2,h}$, and note that the finite element subspaces $\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h\times\widetilde{\mathbf{Z}}_h$ are inf-sup stable for the bilinear form $\tilde{b}_2$ (cf. [45, Section 4.5])

$$\sup_{\mathbf{0}\neq(\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h)\in\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h} \frac{\tilde{b}_2((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\underline{\boldsymbol{v}}_h)}{\|(\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h)\|_{\widetilde{\mathbf{X}}\times\widetilde{\mathbf{Y}}}} \geq \tilde{\beta}^*_{b_2}\,\|\underline{\boldsymbol{v}}_h\|_{\widetilde{\mathbf{Z}}} \quad \forall\, \underline{\boldsymbol{v}}_h \in \widetilde{\mathbf{Z}}_h. \tag{6.19}$$

In addition, it is straightforward to see that the kernel of the bilinear form $\tilde{b}_2$ can be characterised by

$$\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_{h,0}, \quad \text{with} \quad \widetilde{\mathbf{Y}}_{h,0} = \{\boldsymbol{\tau}_h \in \widetilde{\mathbf{Y}}_h : \ \mathbf{div}\,\boldsymbol{\tau}_h = \mathbf{0} \quad \text{and} \quad \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathrm{t}}\}, \tag{6.20}$$

and, similarly to (4.6), using that $\widetilde{\mathbf{Y}}_{h,0}\subset\widetilde{\mathbf{X}}_{1,h}$, we can take $\boldsymbol{e}_h = \boldsymbol{\tau}_h$ to conclude that $\tilde{b}_1$ satisfies the inf-sup condition

$$\sup_{\mathbf{0}\neq\underline{\boldsymbol{e}}_h\in\widetilde{\mathbf{X}}_h} \frac{\tilde{b}_1(\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h)}{\|\underline{\boldsymbol{e}}_h\|_{\widetilde{\mathbf{X}}}} \geq \|\boldsymbol{\tau}_h\|_{\widetilde{\mathbf{Y}}} \quad \forall\, \boldsymbol{\tau}_h \in \widetilde{\mathbf{Y}}_{0,h}. \tag{6.21}$$

Finally, the scheme associated with the weak formulation (6.4) consists in finding $((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\underline{\boldsymbol{u}}_h)\in(\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h)\times\widetilde{\mathbf{Z}}_h$ such that

$$\begin{aligned}\tilde{a}(\underline{\boldsymbol{d}}_h,\underline{\boldsymbol{e}}_h) \;+\; \tilde{b}_1(\underline{\boldsymbol{e}}_h,\boldsymbol{\sigma}_h) &= \widetilde{G}(\underline{\boldsymbol{e}}_h),\\ \tilde{b}_1(\underline{\boldsymbol{d}}_h,\boldsymbol{\tau}_h) \;+\; \tilde{b}_2((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\underline{\boldsymbol{u}}_h) &= \widetilde{H}(\boldsymbol{\tau}_h),\\ \tilde{b}_2((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\underline{\boldsymbol{v}}_h) &= \widetilde{F}(\underline{\boldsymbol{v}}_h),\end{aligned} \tag{6.22}$$

for all $((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\underline{\boldsymbol{v}}_h)\in(\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h)\times\widetilde{\mathbf{Z}}_h$, with $\underline{\boldsymbol{d}}_h = (\boldsymbol{d}_h,p_h)$, $\underline{\boldsymbol{e}}_h = (\boldsymbol{e}_h,q_h)$, $\underline{\boldsymbol{u}}_h = (\boldsymbol{u}_h,\boldsymbol{\gamma}_h)$ and $\underline{\boldsymbol{v}}_h = (\boldsymbol{v}_h,\boldsymbol{\eta}_h)$.

Remark 6.1. Due to the definition of the bilinear forms b_1, b_2, $\tilde{b}_1$ and $\tilde{b}_2$ (cf. (2.2) and (6.3)), the subspaces $\mathbf{X}_{1,h}$ and $\widetilde{\mathbf{X}}_{1,h}$, defined in (4.3) and (6.16) (or (6.18)), can be chosen in many ways; it is sufficient that $\mathbf{div}\,\mathbf{Y}_h \subseteq \mathbf{X}_{1,h}$ and $\mathbf{div}\,\widetilde{\mathbf{Y}}_h \subseteq \widetilde{\mathbf{X}}_{1,h}$, with this, the inf-sup condition of b_1 and $\tilde{b}_1$ in the kernel of b_2 and $\tilde{b}_2$, respectively, can be ensured. For example, we can take the space $\mathbf{X}_{1,h} := \left\{\boldsymbol{e}_h \in \mathbb{L}^2_{\mathrm{sym}}(\Omega) : \ \boldsymbol{\tau}_h|_K \in \mathbb{P}_{k+2}(K)\ \forall\, K \in \mathcal{T}_h\right\}$, to approximate the strain tensor in the first Galerkin scheme, but it is more expensive.

To establish the unique solvability of (6.22), following a similar approach as in Section 4.2, we define the following set:

$$\widetilde{\mathbf{W}}_h := \left\{\underline{\boldsymbol{w}}_h := (\boldsymbol{w}_h,s_h)\in\widetilde{\mathbf{X}}_h \ : \ \|\underline{\boldsymbol{w}}_h\|_{\widetilde{\mathbf{X}}} \leq r\right\}. \tag{6.23}$$

Next, for a fixed $\underline{\boldsymbol{w}}_h := (\boldsymbol{w}_h,s_h)$ in $\widetilde{\mathbf{W}}_h$, we have that the bilinear form $\tilde{a}_{\underline{\boldsymbol{w}}}$ defined in (6.10), satisfies:

$$\tilde{a}_{\underline{\boldsymbol{w}}_h}(\underline{\boldsymbol{e}}_h,\underline{\boldsymbol{e}}_h) \geq c_a\|\underline{\boldsymbol{e}}_h\|^2_{\widetilde{\mathbf{X}}} \qquad \forall\, \underline{\boldsymbol{e}}_h \in \widetilde{\mathbf{X}}_h. \tag{6.24}$$

Using arguments analogous to those of Theorem 3.1 (see also Theorem 4.1), and the properties (6.19), (6.21), and (6.24) of the bilinear forms $\tilde{b}_2$, $\tilde{b}_1$, and $\tilde{a}_{\underline{\boldsymbol{w}}_h}$, adapting the fixed-point strategy introduced in Section 3.2 (see also Section 4.2) to the present case, we can assert the following result.

Theorem 6.2. *Let $\boldsymbol{f}\in\mathbf{L}^2(\Omega)$, $g\in L^2(\Omega)$, $\boldsymbol{u}_\Gamma\in\mathbf{H}^{1/2}(\Gamma)$ and $r_\Gamma\in\mathrm{H}^{-1/2}(\Gamma)$ such that*

$$\frac{\tilde{\gamma}_1^*}{r}\max\{\tilde{\gamma}_1^*\,\kappa_2\,r,1\}\left(\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma}+\|\boldsymbol{f}\|_{0,\Omega}\right)<1, \tag{6.25}$$

where $\tilde{\gamma}_1^*$ is the discrete counterpart of $\tilde{\gamma}_1$ (cf. (6.13)), defined by

$$\tilde{\gamma}_1^* := \frac{(C_a+1+\tilde{\beta}_2^*+\gamma_2)^2}{(\tilde{\beta}_2^*)^2\,\gamma_2}, \tag{6.26}$$

and with γ_2 defined in (3.12). Then, problem (6.22) has a unique solution $((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\underline{\boldsymbol{u}}_h)\in\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h\times\widetilde{\mathbf{Z}}_h$ with $\underline{\boldsymbol{d}}_h\in\widetilde{\mathbf{W}}_h$. In addition, there exists $\tilde{C}^*>0$ such that

$$\|(\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\underline{\boldsymbol{u}}_h\| \leq \tilde{C}^*\left(\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma}+\|\boldsymbol{f}\|_{0,\Omega}\right). \tag{6.27}$$

Employing the same arguments utilised in Section 5.1 and Theorem 5.1, we can provide the Céa estimate corresponding to the Galerkin scheme (6.22).

Theorem 6.3. *Assume that*

$$\kappa_2\tilde{\gamma}_1^*\tilde{\gamma}_1\left(\|g\|_{0,\Omega}+\|r_\Gamma\|_{-1/2,\Gamma}+\|\boldsymbol{u}_\Gamma\|_{1/2,\Gamma}+\|\boldsymbol{f}\|_{0,\Omega}\right)\leq\frac{1}{2}, \tag{6.28}$$

with $\tilde{\gamma}_1$ and $\tilde{\gamma}_1^*$ being the constants in (6.13) and (6.26). Assume that the hypotheses of Theorem 6.1 and Theorem 6.2 hold. Let $((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\underline{\boldsymbol{u}})\in(\widetilde{\mathbf{X}}\times\widetilde{\mathbf{Y}})\times\widetilde{\mathbf{Z}}$ and $((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\underline{\boldsymbol{u}}_h)\in(\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h)\times\widetilde{\mathbf{Z}}_h$ be the unique solutions of (6.4) and (6.22), respectively. Then, there exists $\tilde{C}_{Céa}>0$, such that

$$\|((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\boldsymbol{u})-((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\boldsymbol{u}_h)\|\leq\tilde{C}_{Céa}\inf_{((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\underline{\boldsymbol{v}}_h)\in(\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h)\times\widetilde{\mathbf{Z}}_h}\|((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\underline{\boldsymbol{u}})-((\underline{\boldsymbol{e}}_h,\boldsymbol{\tau}_h),\underline{\boldsymbol{v}}_h)\|, \tag{6.29}$$

with $\tilde{C}_{Céa}=2(\tilde{\gamma}_1^*(C_a+2)+1)$.

On the other hand, we observe the following approximation properties related to spaces (6.15)–(6.16).

($\widetilde{\mathbf{AP}}_h^{\boldsymbol{d}}$) For each $0\leq m\leq k+1$ and for each $\boldsymbol{e}\in\mathbb{H}^m(\Omega)$, there holds

$$\operatorname{dist}\left(\boldsymbol{e},\widetilde{\mathbf{X}}_{1,h}\right):=\inf_{\boldsymbol{e}_h\in\widetilde{\mathbf{X}}_{1,h}}\|\boldsymbol{e}-\boldsymbol{e}_h\|_{0,\Omega}\lesssim h^m\|\boldsymbol{e}\|_{m,\Omega}. \tag{6.30a}$$

($\widetilde{\mathbf{AP}}_h^{p}$) Coincides with ($\mathbf{AP}_h^{p}$).

($\widetilde{\mathbf{AP}}_h^{\boldsymbol{\sigma}}$) For each $0<m\leq k+1$ and for each $\boldsymbol{\tau}\in\mathbb{H}^m(\Omega)\cap\mathbb{H}(\mathbf{div};\Omega)$ with $\mathbf{div}\,\boldsymbol{\tau}\in\mathbf{H}^m(\Omega)$, there holds

$$\operatorname{dist}\left(\boldsymbol{\tau},\widetilde{\mathbf{Y}}_h\right):=\inf_{\boldsymbol{\tau}_h\in\widetilde{\mathbf{Y}}_h}\|\boldsymbol{\tau}-\boldsymbol{\tau}_h\|_{\widetilde{Y}}\lesssim h^m\left\{\|\boldsymbol{\tau}\|_{m,\Omega}+\|\mathbf{div}\,\boldsymbol{\tau}\|_{m,\Omega}\right\}. \tag{6.30b}$$

($\widetilde{\mathbf{AP}}_h^{\boldsymbol{u}}$) For each $0\leq m\leq k+1$ and for each $\boldsymbol{v}\in\mathbf{H}^m(\Omega)$, there holds

$$\operatorname{dist}\left(\boldsymbol{v},\widetilde{\mathbf{Z}}_{1,h}\right):=\inf_{\boldsymbol{v}_h\in\widetilde{\mathbf{Z}}_{1,h}}\|\boldsymbol{v}-\boldsymbol{v}_h\|_{0,\Omega}\lesssim h^m\|\boldsymbol{v}\|_{m,\Omega}. \tag{6.30c}$$

($\widetilde{\mathbf{AP}}_h^{\boldsymbol{\gamma}}$) For each $0\leq m\leq k+1$ and for each $\boldsymbol{\eta}\in\mathbb{H}^m(\Omega)\cap\mathbb{L}^2_{\mathrm{skew}}(\Omega)$, there holds

$$\operatorname{dist}\left(\boldsymbol{\eta},\widetilde{\mathbf{Z}}_{2,h}\right):=\inf_{\boldsymbol{\eta}_h\in\widetilde{\mathbf{Z}}_{2,h}}\|\boldsymbol{\eta}-\boldsymbol{\eta}_h\|_{0,\Omega}\lesssim h^m\|\boldsymbol{\eta}\|_{m,\Omega}. \tag{6.30d}$$

For (6.30a), (6.30b), (6.30c) and (6.30d) we refer to [27, Theorem 2.4].

Theorem 6.4. *Assume that the hypotheses of Theorem 6.3 hold and let $((\underline{\boldsymbol{d}},\boldsymbol{\sigma}),\underline{\boldsymbol{u}})\in\widetilde{\mathbf{X}}\times\widetilde{\mathbf{Y}}\times\widetilde{\mathbf{Z}}$ and $((\underline{\boldsymbol{d}}_h,\boldsymbol{\sigma}_h),\underline{\boldsymbol{u}}_h)\in(\widetilde{\mathbf{X}}_h\times\widetilde{\mathbf{Y}}_h)\times\widetilde{\mathbf{Z}}_h$ be the unique solutions of the continuous and discrete problems (6.4) and (6.22), respectively. Assume further that $\boldsymbol{d}\in\mathbb{H}^m(\Omega)$, $p\in\mathrm{H}^{m+1}(\Omega)$, $\boldsymbol{\sigma}\in\mathbb{H}^m(\Omega)$, $\mathbf{div}\,\boldsymbol{\sigma}\in\mathbf{H}^m(\Omega)$, $\boldsymbol{u}\in\mathbf{H}^m(\Omega)$ and $\boldsymbol{\gamma}\in\mathbb{H}^m(\Omega)$, for $0\leq m\leq k+1$. Then there exists $\widetilde{C}_{rate}>0$, independent of h, such that*

$$\|((\mathrm{e}_{\underline{\boldsymbol{d}}},\mathrm{e}_{\boldsymbol{\sigma}}),\mathrm{e}_{\underline{\boldsymbol{u}}})\|\leq\widetilde{C}_{rate}\,h^m\left\{\|\boldsymbol{d}\|_{m,\Omega}+\|p\|_{m+1,\Omega}+\|\boldsymbol{\sigma}\|_{m,\Omega}+\|\mathbf{div}\,\boldsymbol{\sigma}\|_{m,\Omega}+\|\boldsymbol{u}\|_{m,\Omega}+\|\boldsymbol{\gamma}\|_{m,\Omega}\right\}.$$

Proof. The result is a straightforward application of Theorem 6.3 and the approximation properties ($\widetilde{\mathbf{AP}}_h^{\boldsymbol{d}}$), ($\widetilde{\mathbf{AP}}_h^{p}$), ($\widetilde{\mathbf{AP}}_h^{\boldsymbol{\sigma}}$), ($\widetilde{\mathbf{AP}}_h^{\boldsymbol{u}}$) and ($\widetilde{\mathbf{AP}}_h^{\boldsymbol{\gamma}}$). □

7. Numerical results

This section contains selected computational examples that serve to confirm the theoretically obtained convergence rates of the two analysed mixed finite element formulations. We showcase tests in 2D and 3D including experimental convergence as well as an application-oriented simulation of filtration of interstitial fluid in soft tissue. The implementation has been carried out using the finite element library `Firedrake` [49]. Each solve of the discrete nonlinear coupled system was performed using Newton–Raphson's method with iterations terminated whenever the absolute or the relative ℓ^∞–norm of the discrete residual in the product space drops below the fixed tolerance 10^{-7}. Each linear system arising from linearisation was solved using the sparse LU factorisation algorithm MUMPS.

7.1. Verification of convergence with respect to smooth solutions

First we conduct a test of convergence where model parameters assume the following values $\mu=\lambda=\mu_f=1$ and $c_0=\alpha=\frac{1}{4}$. The nonlinear permeability is taken as the second form in (1.6) (the Kozeny–Carman law) with $k_0=k_1=0.1$. We use the following closed-form smooth solutions to the primal form of the coupled nonlinear problem

$$\boldsymbol{u}=\frac{1}{5}\begin{pmatrix}-x_1\cos(x_1)\sin(x_2)+x_1^2\\ x_1\sin(x_1)\cos(x_2)+x_2^2\end{pmatrix},\quad p=\sin(\pi x_1)\sin(\pi x_2),$$

Table 7.1
Verification of convergence for the method imposing stress symmetry strongly and using AW_k-based elements (using $k = 1$: this gives polynomial degree 3 for stress and strain and 1 for displacement). Errors and convergence rates are tabulated for strain, fluid pressure, poroelastic stress, and displacement. The symbol $\star$ in the first mesh refinement level indicates that no convergence rate is computed.

DoFs	h	$\mathrm{e}_0(\boldsymbol{d})$	rate	$\mathrm{e}_1(p)$	rate	$\mathrm{e}_{\mathbf{div}}(\boldsymbol{\sigma})$	rate	$\mathrm{e}_0(\boldsymbol{u})$	rate
303	0.5000	1.5e−03	⋆	3.8e−01	⋆	8.4e−02	⋆	6.3e−03	⋆
1063	0.2500	3.1e−04	2.24	1.2e−01	1.70	2.2e−02	1.92	1.6e−03	2.01
3975	0.1250	4.9e−05	2.68	3.2e−02	1.88	5.6e−03	1.98	3.9e−04	2.00
15367	0.0625	7.2e−06	2.76	8.2e−03	1.95	1.4e−03	2.00	9.8e−05	2.00
60423	0.0312	1.1e−06	2.72	2.1e−03	1.98	3.5e−04	2.00	2.5e−05	2.00
239623	0.0156	2.3e−07	2.03	5.3e−04	1.98	8.7e−05	2.00	6.1e−06	2.00

Table 7.2
Verification of convergence for the method imposing stress symmetry weakly and using PEERS_k (6.15) and AFW_k (6.17) elements with polynomial degrees $k = 0$ and $k = 1$. Errors and convergence rates are tabulated for strain, fluid pressure, poroelastic stress, displacement, and rotation Lagrange multiplier. The $\star$ in the first mesh refinement level indicates that no convergence rate is computed.

DoFs	h	$\mathrm{e}_0(\boldsymbol{d})$	rate	$\mathrm{e}_1(p)$	rate	$\mathrm{e}_{\mathbf{div}}(\boldsymbol{\sigma})$	rate	$\mathrm{e}_0(\boldsymbol{u})$	rate	$\mathrm{e}_0(\boldsymbol{\gamma})$	rate
PEERS_k-based FE scheme with $k = 0$											
130	0.7071	1.5e−01	⋆	1.1e+0	⋆	1.2e+0	⋆	4.6e−02	⋆	8.4e−02	⋆
482	0.3536	9.4e−02	0.66	7.4e−01	0.62	6.3e−01	0.90	2.3e−02	0.99	4.9e−02	0.79
1858	0.1768	5.2e−02	0.86	4.1e−01	0.85	3.2e−01	0.98	1.1e−02	1.02	2.7e−02	0.83
7298	0.0884	2.6e−02	0.97	2.1e−01	0.95	1.6e−01	1.00	5.6e−03	1.01	1.1e−02	1.26
28930	0.0442	1.3e−02	1.00	1.1e−01	0.98	8.0e−02	1.00	2.8e−03	1.00	4.3e−03	1.42
115202	0.0221	6.6e−03	1.00	5.4e−02	1.00	4.0e−02	1.00	1.4e−03	1.00	1.5e−03	1.48
PEERS_k-based FE scheme with $k = 1$											
386	0.7071	2.3e−02	⋆	3.8e−01	⋆	3.4e−01	⋆	5.0e−03	⋆	9.5e−03	⋆
1474	0.3536	8.1e−03	1.51	1.2e−01	1.70	9.0e−02	1.91	1.2e−03	2.02	3.1e−03	1.63
5762	0.1768	2.6e−03	1.66	3.2e−02	1.88	2.3e−02	1.97	3.0e−04	2.02	1.4e−03	1.64
22786	0.0884	7.3e−04	1.81	8.2e−03	1.95	5.8e−03	1.99	7.5e−05	2.02	4.9e−04	1.79
90626	0.0442	1.9e−04	1.91	2.1e−03	1.98	1.5e−03	1.99	1.9e−05	2.01	1.4e−04	1.89
361474	0.0221	5.0e−05	1.96	5.2e−04	1.99	3.7e−04	2.00	4.7e−06	2.00	3.8e−05	1.98
AFW_k-based FE scheme with $k = 0$											
161	0.7071	2.5e−02	⋆	1.1e+0	⋆	1.1e+0	⋆	4.4e−02	⋆	2.6e−02	⋆
569	0.3536	1.5e−02	0.72	7.4e−01	0.62	5.7e−01	0.93	2.2e−02	0.98	1.4e−02	0.93
2129	0.1768	5.9e−03	1.35	4.1e−01	0.85	2.9e−01	0.99	1.1e−02	1.00	6.3e−03	1.09
8225	0.0884	2.4e−03	1.30	2.1e−01	0.95	1.5e−01	1.00	5.6e−03	1.00	3.1e−03	1.05
32321	0.0442	1.1e−03	1.13	1.1e−01	0.98	7.3e−02	1.00	2.8e−03	1.00	1.5e−03	1.02
128129	0.0221	5.4e−04	1.04	5.4e−02	1.00	3.6e−02	1.00	1.4e−03	1.00	7.5e−04	1.00
AFW_k-based FE scheme with $k = 1$											
385	0.7071	5.3e−03	⋆	3.8e−01	⋆	3.3e−01	⋆	4.8e−03	⋆	3.0e−03	⋆
1425	0.3536	1.3e−03	2.05	1.2e−01	1.70	8.7e−02	1.92	1.2e−03	2.00	9.7e−04	1.65
5473	0.1768	2.1e−04	2.61	3.2e−02	1.88	2.2e−02	1.98	3.0e−04	2.00	1.9e−04	2.38
21441	0.0884	3.4e−05	2.62	8.2e−03	1.95	5.5e−03	2.00	7.5e−05	2.00	3.8e−05	2.30
84865	0.0442	6.5e−06	2.38	2.1e−03	1.98	1.4e−03	2.00	1.9e−05	2.00	8.7e−06	2.12
337665	0.0221	1.5e−06	2.15	5.2e−04	1.99	3.5e−04	2.00	4.7e−06	2.00	2.1e−06	2.04

which are used to generate exact mixed variables $\boldsymbol{d}, \boldsymbol{\sigma}, \boldsymbol{\gamma}$, and to produce non-homogeneous forcing term $\boldsymbol{f}$, boundary data $\boldsymbol{u}_\Gamma, r_\Gamma$, and source term g. Seven successively refined meshes (congruent right-angled triangular partitions) are generated for the domain $\Omega = (0,1)^2$ and we compute errors between approximate and exact solutions $e(\cdot)$ (measured in the H^1-norm for fluid pressure, in the tensor $\mathbb{H}(\mathbf{div})$-norm for poroelastic Cauchy stress, and in the tensorial and vectorial L^2-norms for strain, rotation, and displacement). The experimental rates of convergence at each mesh refinement are computed as

$$\mathtt{rate}(\%) = \frac{\log(\mathrm{e}(\%)/\mathrm{e}'(\%))}{\log(h/h')},$$

with $\% \in \{\boldsymbol{d}, p, \boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\gamma}\}$, and where e, e' stand for errors generated on two consecutive meshes of sizes h, h'. The mixed finite element methods are defined by the conforming AW_k, AFW_k, and PEERS_k-type of spaces specified in Section 6.

Tables 7.1–7.2 show the error history associated with the formulations with strong and weak symmetry imposition, and using for the latter case the two lowest-order polynomial degrees. In all runs our results confirm an error decay with a convergence rate of $O(h^{k+1})$ for all field variables in their natural norms, which is consistent with the theoretical error bounds from Theorems 5.2–6.4. We also depict examples of approximate solutions computed with the PEERS_k-based finite element family (setting $k = 1$). See Fig. 7.1, where the panels show also the outline of the domain before the deformation.

We also include a test (only for the lowest-order AFW scheme) where we take small storativity and small permeability $c_0 = 10^{-8}$, $k_0 = k_1 = 10^{-12}$. The error history in Table 7.3 indicates still an optimal convergence rate and suggesting that the method does not have pressure locking.

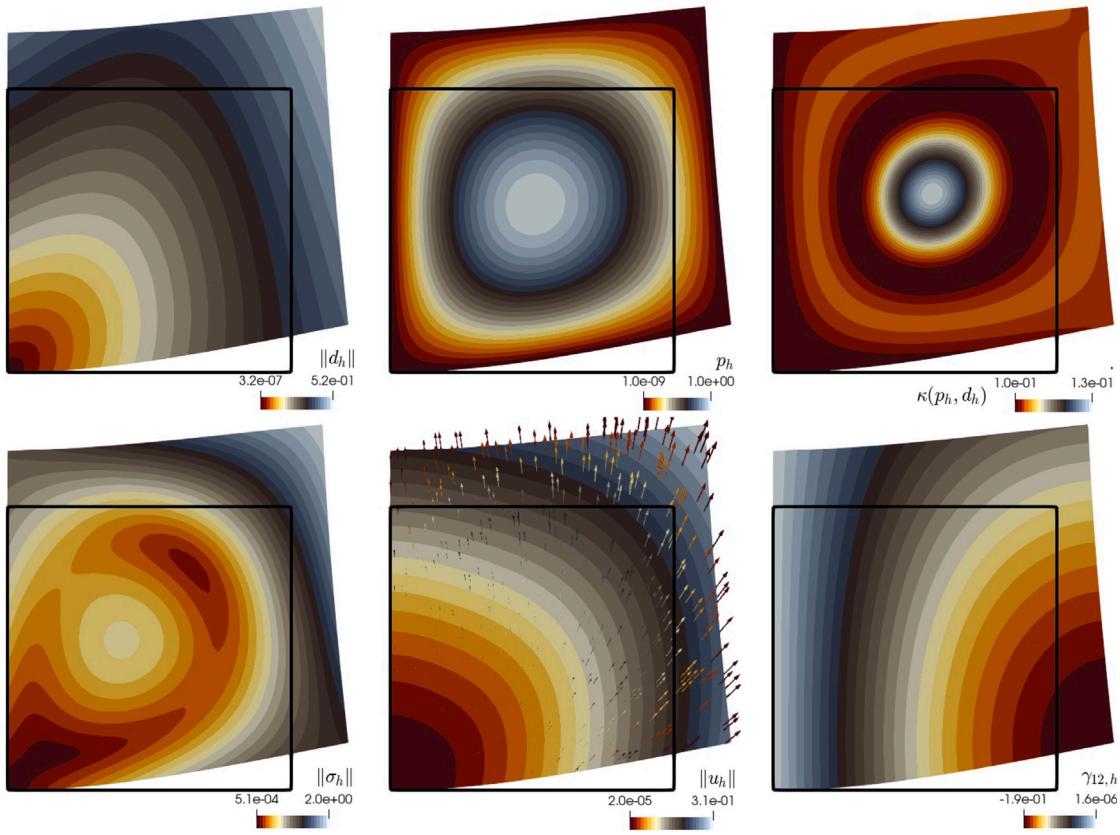

Fig. 7.1. Convergence tests. Approximate solutions computed with the second-order PEERS$_k$-based finite element scheme and rendered on the deformed configuration. Only the non-trivial component of the rotation tensor is shown in the bottom-right panel.

Table 7.3
Verification of convergence for the method imposing stress symmetry weakly and using the lowest-order AFW$_k$ (6.17) elements, and taking small storativity and permeability parameters. The ⋆ in the first mesh refinement level indicates that no convergence rate is computed.

DoFs	h	$\mathrm{e}_0(\boldsymbol{d})$	rate	$\mathrm{e}_1(p)$	rate	$\mathrm{e}_{\mathbf{div}}(\boldsymbol{\sigma})$	rate	$\mathrm{e}_0(\boldsymbol{u})$	rate	$\mathrm{e}_0(\boldsymbol{\gamma})$	rate
AFW$_k$-based FE scheme with $k = 0$											
161	0.7071	2.3e−02	⋆	1.2e+0	⋆	3.3e−01	⋆	4.1e−02	⋆	2.9e−02	⋆
569	0.3536	1.1e−02	1.05	8.1e−01	0.61	1.7e−01	0.99	2.1e−02	0.98	1.5e−02	0.98
2129	0.1768	5.3e−03	1.05	4.4e−01	0.89	8.4e−02	0.99	1.0e−02	0.99	7.4e−03	1.00
8225	0.0884	2.6e−03	1.02	2.2e−01	0.98	4.2e−02	1.00	5.2e−03	1.00	3.7e−03	1.00
32321	0.0442	1.3e−03	1.01	1.1e−01	1.00	2.1e−02	1.00	2.6e−03	1.00	1.9e−03	1.00
128129	0.0221	6.6e−04	1.00	5.6e−02	1.00	1.1e−02	1.00	1.3e−03	1.00	9.3e−04	1.00

7.2. Simulation of swelling of a porous structure

In the next test we replicate the swelling of a 3D block. The parameters and domain configuration are taken similarly to [50] and we simulate this behaviour with the second-order AFW$_k$-based finite element method using (6.17)–(6.18). The domain is $\Omega = (0,1) \times (0,1) \times (0, \frac{1}{2})$ and the deformation is induced by a fluid pressure gradient in the x_1−direction (we impose $p = 10^4$ at $x_1 = 0$, $p = 0$ at $x_1 = 1$, and zero-flux conditions on the remainder of Γ). The poroelastic body is allowed to slide on the sides $x_1 = 0$, $x_2 = 0$ and $x_3 = 0$ (these parts are called Γ_{slide}), whereas zero normal stress is considered elsewhere on Γ. The sliding condition $\boldsymbol{u} \cdot \boldsymbol{n}|_{\Gamma_{\text{slide}}} = 0$ is incorporated through the additional term

$$\langle (\boldsymbol{\tau}\boldsymbol{n}) \times \boldsymbol{n}, \boldsymbol{u} \times \boldsymbol{n} \rangle_{\Gamma_{\text{slide}}},$$

on the left-hand side of the second equation in the weak formulation (2.3) and taking $H(\boldsymbol{\tau}) = 0$. The model parameters for this test are the exponential permeability in (1.6) with $k_0 = 10^{-9}$, $k_1 = 10^{-6}$, $k_2 = -0.5$, the Young modulus $E = 8000$, Poisson ratio $\nu = 0.3$, storativity coefficient $c_0 = 0.001$, Biot–Willis parameter $\alpha = 0.5$, fluid viscosity $\mu_f = 10^{-3}$, and we take zero body loads and volumetric sources. Fig. 7.2 displays the approximate solutions rendered on the deformed configuration. We also show the

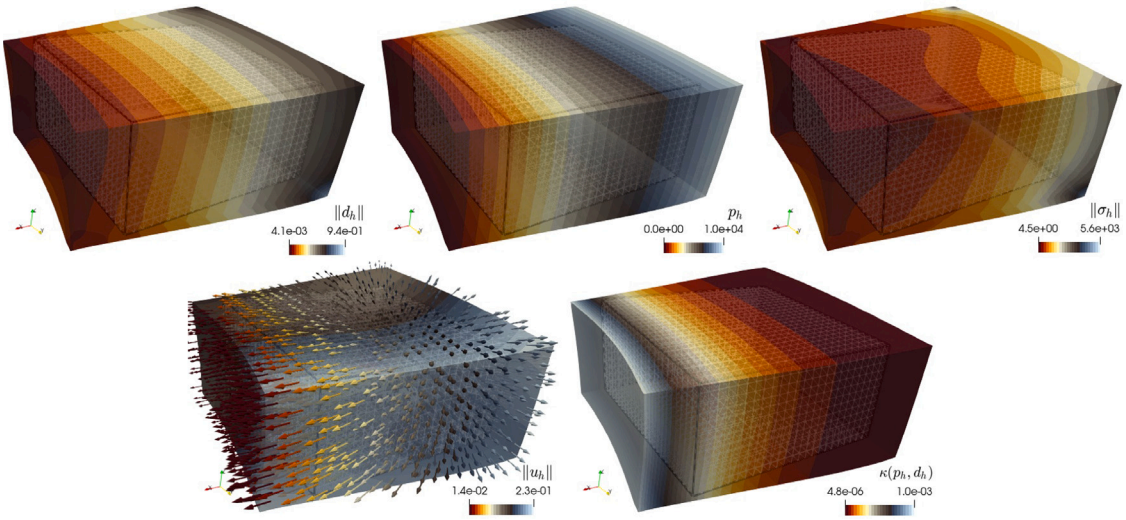

Fig. 7.2. Swelling of a poroelastic block. Approximate infinitesimal strain magnitude, fluid pressure, poroelastic stress magnitude, displacement magnitude and direction arrows, and permeability distribution.

contour of the undeformed domain for reference (the box with the light mesh shown in the background). The primary goal of this example is to showcase the performance of the proposed method in a 3D setting. From the figure we can observe that the swelling occurs in the x and y directions and the maximal displacements near the edge $x = 1$ (away from the slip displacement boundaries) are of approximately 20% of the domain diameter. Even if the permeability and pressure distribution are roughly linear in the x direction, the poroelastic regime adopted here produces a stress magnitude showing a non-uniform pattern. Note also that there are no nonphysically oscillating pressures.

7.3. Poroelastic filtration of slightly compressible trabecular meshwork

For our next example we consider a computational domain extracted and meshed from imaging of trabecular meshwork tissue in the canine eye in [51]. The characteristic length of the domain is $6.8 \cdot 10^{-3}$ [m]. In this test we use the strongly imposed symmetry with AW_k-based finite elements (4.2)–(4.3), and take piecewise linear and overall continuous elements for the fluid pressure. We set up an initial porosity field ϕ_0, randomly distributed between 0.3 and 0.45. A nonlinear permeability is prescribed depending on that initial porosity and on fluid pressure and skeleton dilation

$$\kappa(\boldsymbol{d}, p) = \frac{k_0}{\mu_f} + \frac{k_1}{\mu_f} \exp(-\frac{1}{2}(\phi_0 + (1 - \phi_0)[c_0 p + \alpha \operatorname{tr} \boldsymbol{d}])) \quad [\text{in m}^2],$$

where the dependence on the total amount of fluid is taken similarly as in [9]. The model parameters for this test are $k_0 = 10^{-10}$, $k_1 = 10^{-7}$, the Lamé constants $\lambda = 14388$ [Pa], $\mu = 1102$ [Pa], storativity coefficient $c_0 = 0.05$, Biot–Willis parameter $\alpha = 0.95$, fluid viscosity $\mu_f = 7.54^{-4}$ [Pa · s], and we take zero body loads and volumetric sources. The domain is assumed in contact, on a portion of the boundary on the top-left end, with the anterior chamber in the eye and therefore we set a traction of $\boldsymbol{\sigma n} = -3 \times 10^{-3} \boldsymbol{n}$ and a pore pressure $p = 2 \times 10^3$ [Pa]. On the outlet sub-boundary (a small region on the bottom-right end) we impose zero fluid pressure and traction-free conditions, and on the remainder of the boundary we set zero displacements and zero flux for the fluid pressure. From the results portrayed in Fig. 7.3 we observe that the pore pressure and strain concentration generation near the interfacial region imply a smaller permeability, which progressively increases as one approaches the outlet boundary. This behaviour coincides with the first round of tests with different permeability profiles explored in [51]. We also plot the off-diagonal entries of the Cauchy stress to illustrate the balance of angular momentum, and on the bottom-right panel we can see the deformation of the interfacial region and, as expected, a smaller expansion of the tissue towards the outlet.

7.4. Reproducing the mandel effect

To conclude this section, we utilise the proposed formulation, specifically focusing on the scenario of weakly symmetric stress, to simulate Mandel's effect (see, e.g., [52] or also [2,53–57]). Such a problem involves a specimen made of isotropic poroelastic material, which is positioned between two rigid frictionless impervious plates at the top and bottom. The slab is infinitely long with a cross section measuring $2L \times 2H$. The lateral sides of the specimen are free and permeable. In this simulation, a compressive force is exerted on the horizontal plates. As a result, the pore pressure at the centre of the specimen surpasses its starting value during the early stages of the process and subsequently diminishes until it reaches zero. This behaviour can be attributed to the drainage

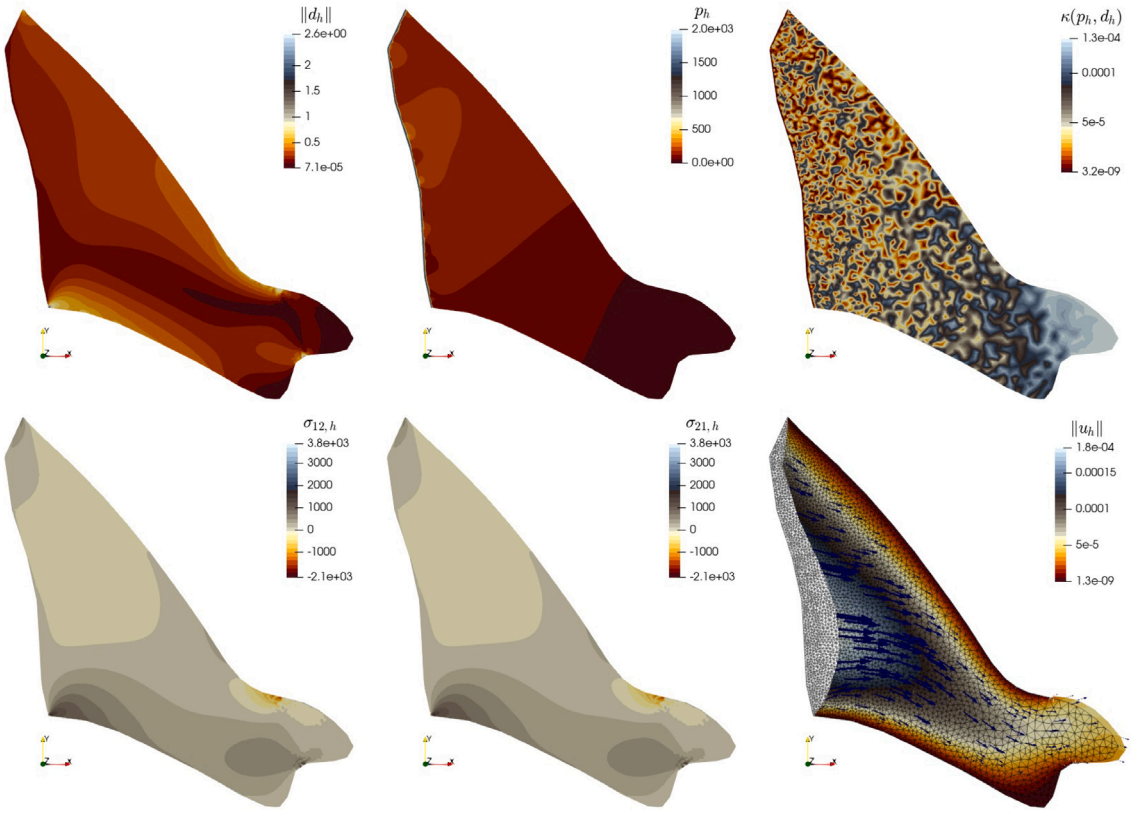

Fig. 7.3. Poroelastic filtration of trabecular meshwork tissue. Approximate infinitesimal strain magnitude, fluid pressure, nonlinear permeability (depending on an initial randomly distributed porosity), two off-diagonal entries of the poroelastic stress, and displacement magnitude with direction arrows.

of fluid from the specimen through the side edges. Consequently, a greater portion of the applied load is transferred towards the comparatively stiffer central region of the specimen.

To simplify the analysis, as usual, taking into account the symmetry of the geometry and problem set up, we only consider a quarter of the entire domain: $\Omega = (0, L)\times(0, H)$. This implies that the mechanical boundary conditions for the smaller domain Ω are as follows: on the boundary $x = L$ the pore pressure is fixed at zero and the normal poroelastic stress is also zero (both imposed as essential boundary conditions). At the boundaries $x = 0$ and $y = 0$ we impose a sliding condition $\boldsymbol{u}\cdot\boldsymbol{n} = 0$ for which, as in Section 7.2, the term $-\langle(\boldsymbol{\tau n})\cdot\boldsymbol{t}, \boldsymbol{u}\cdot\boldsymbol{t}\rangle$ appears on the right-hand side of the weak form (where $\boldsymbol{t}$ denotes the tangent vector to the boundary). A downward force of magnitude $2F$ is applied to the top plate (the boundary $y = H$) through the essential boundary condition for poroelastic stress $\boldsymbol{\sigma n} = (0, -F)^{\mathrm{t}}$. Zero-flux for the fluid phase is considered in all sub-boundaries except on $x = L$. This problem takes place in the quasi-steady regime and therefore we bring back in the time dependence, discretised using backward Euler's method with constant time step Δt and running until T_{end}, and necessitating an initial condition for pore pressure and strain (we initialise them both to zero). A coarse mesh is used together with the AFW_k finite element family with $k = 1$, and following e.g. [58] (where a thorough computational and experimental comparison is performed for poroelastic cartilage tissue in different regimes), we test the behaviour of the models with constant and nonlinear permeabilities $\kappa = \kappa_0$ and $\kappa = k_0\kappa_0 \exp(k_1[c_0 p + \alpha \operatorname{tr} \boldsymbol{d}])$.

We set the geometry and model parameters within the ranges used in [55]

$$L = H = 1\,[\mathrm{m}], \quad T_{\mathrm{end}} = 1\,[\mathrm{s}], \quad \Delta t = 0.01\,[\mathrm{s}], \quad c_0 = 4\times10^{-10}\,[\mathrm{Pa}^{-1}], \quad \kappa_0 = 5.1\times10^{-8}\,[\mathrm{m}^2],$$
$$k_0 = 5\,[-], \quad k_1 = 30\,[-], \quad \alpha = 0.9\,[-], \quad \mu_f = 10^{-3}\,[\mathrm{Pa}\cdot\mathrm{s}], \quad \rho = 1\,[\mathrm{Kg/m}^3],$$
$$E = 10^3\,[\mathrm{Pa}], \quad \nu = \frac{1}{3}\,[-], \quad F = 100\,[\mathrm{Pa}], \quad \boldsymbol{f} = \boldsymbol{0}, \quad g = 0.$$

The results of the simulations are collected in Fig. 7.4, where we plot the profiles of pore pressure, horizontal displacement, principal (axial) components of strain and of poroelastic stress over the horizontal mid-line of the domain (at $y = H/2$). The Mandel effect is clearly visible in the first plot, and the difference (between linear and nonlinear cases) in pore pressure build up is similar as the one observed in [58], that is, the nonlinear permeability produces a slightly lower pressure. Fig. 7.5 shows transients of the main variables over time at two spatial points (the left end of the horizontal mid-line and the top-right corner). Qualitatively the results agree with the expected behaviour in both linear and nonlinear regimes. We observe that the largest variation in the first point occurs for the pore pressure drop, whereas for the second point the largest variation is seen in the axial poroelastic stress. For

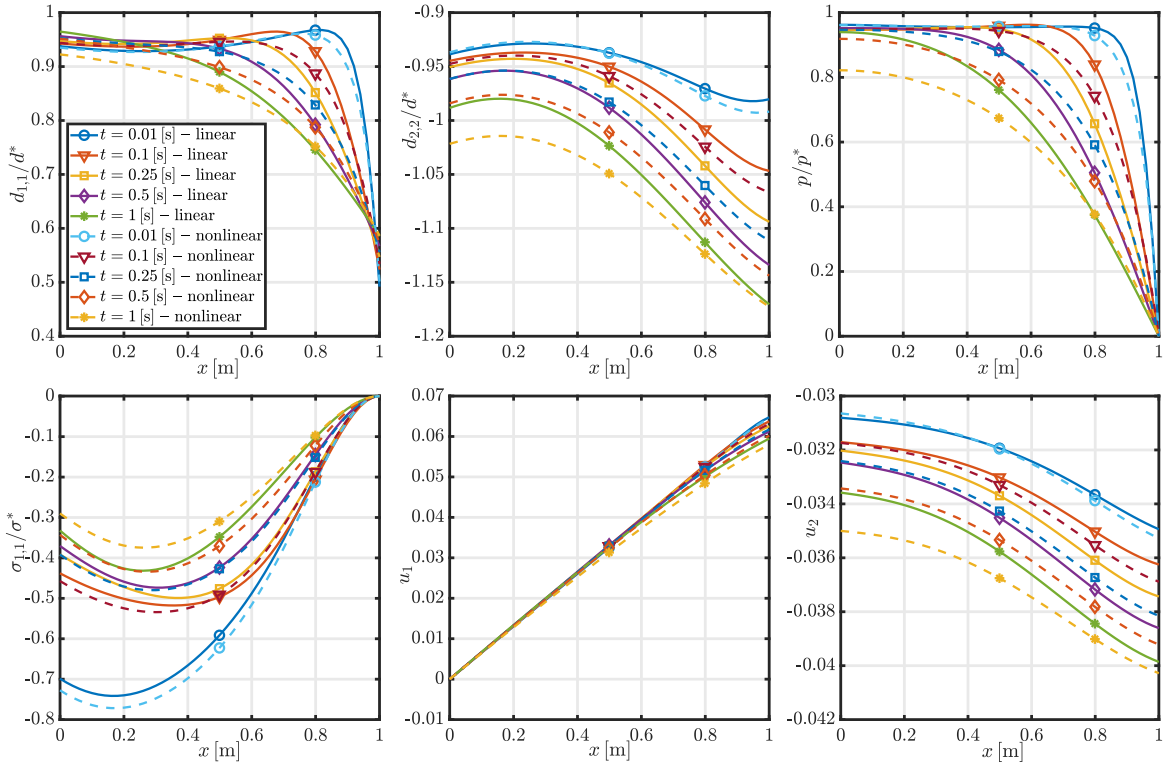

Fig. 7.4. Mandel's test. Plot over the domain horizontal mid-line of axial and radial strains (first and last components of $\boldsymbol{d}$) normalised through $d^* = 0.07$ [–], pore pressure profile normalised by $p^* = 60$ [Pa], axial poroelastic stress normalised by $\sigma^* = 4$ [Pa], and patterns of horizontal an vertical velocities (in [m]), for the constant and nonlinear permeability cases. The legend in the top-left panel (indicating the different times) applies to all panels in the figure.

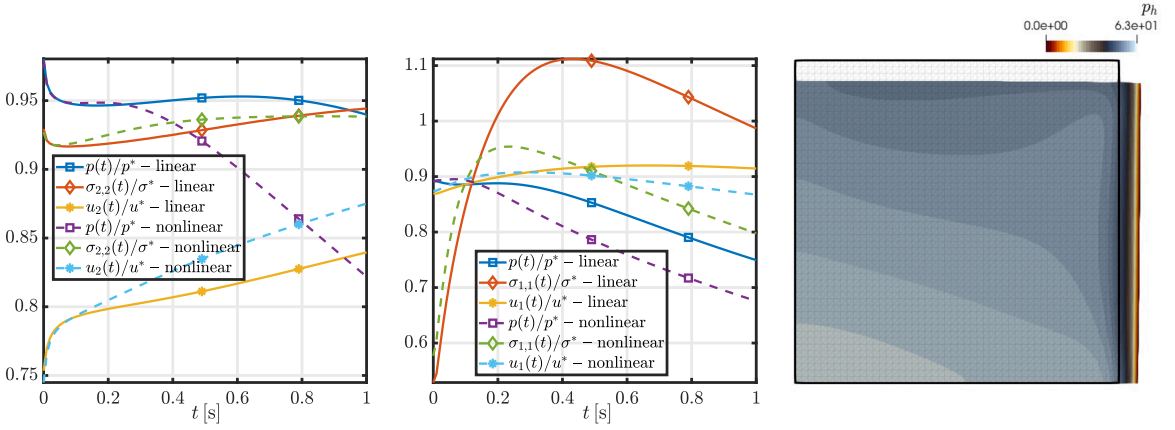

Fig. 7.5. Mandel's test. Variation of pore pressure normalised with $p^* = 60$ [Pa], poroelastic stress (axial component normalised with $\sigma^* = -100$ [Pa] and radial component with $\sigma^* = 8$ [Pa]) and displacement (vertical component normalised with $u^* = -0.04$ [m] and horizontal component with $u^* = 0.04$ [m]) against time for the constant and nonlinear permeability cases, and recorded at the points $(0, H/2)$ (left) and $(L/2, H)$ (middle). The right panel shows the patterns of pore pressure at the final time on the deformed configuration.

completeness we also depict the deformed configuration at the final time together with the pore pressure distribution (produced with the constant permeability case).

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.

Data availability

No data was used for the research described in the article.

References

[1] Borregales Reverón MA, Kumar K, Nordbotten JM, Radu FA. Iterative solvers for Biot model under small and large deformations. Comput Geosci 2021;25:687–99.
[2] Coussy O. Poromechanics. Chichester, UK: John Wiley & Sons Ltd; 2004.
[3] Bociu L, Guidoboni G, Sacco R, Webster JT. Analysis of nonlinear poro-elastic and poro-visco-elastic models. Arch Ration Mech Anal 2016;222:1445–519.
[4] Bociu L, Webster JT. Nonlinear quasi-static poroelasticity. J Differential Equations 2021;296:242–78.
[5] Cao Y, Chen S, Meir A. Analysis and numerical approximations of equations of nonlinear poroelasticity. Discrete Contin Dyn Syst-Ser B 2013;18(5):1253–73.
[6] Gaspar FJ, Lisbona FJ, Matus P, Tuyen VTK. Numerical methods for a one-dimensional non-linear Biot's model. J Comput Appl Math 2016;293:62–72.
[7] Showalter RE, Su N. Partially saturated flow in a poroelastic medium. Discrete Contin Dyn Syst Ser B 2001;1(4):403–20.
[8] Tavakoli A, Ferronato M. On existence-uniqueness of the solution in a nonlinear Biot's model. Appl Math 2013;7(1):333–41.
[9] van Duijn C, Mikelić A. Mathematical theory of nonlinear single-phase poroelasticity. J Nonlinear Sci 2023;33(3):44.
[10] Ambartsumyan I, Khattatov E, Yotov I. A coupled multipoint stress–multipoint flux mixed finite element method for the Biot system of poroelasticity. Comput Methods Appl Mech Engrg 2020;372:113407.
[11] 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):377–96.
[12] Elyes A, Radu F, Nordbotten J. Adaptive poromechanics computations based on a posteriori error estimates for fully mixed formulations of Biot's consolidation model. Comput Methods Appl Mech Engrg 2019;347:264–94.
[13] Lee JJ. Robust error analysis of coupled mixed methods for Biot's consolidation model. J Sci Comput 2016;69:610–32.
[14] Yi S-Y. Convergence analysis of a new mixed finite element method for Biot's consolidation model. Numer Methods Partial Differential Equations 2014;30(4):1189–210.
[15] Ambartsumyan I, Ervin VJ, Nguyen T, Yotov I. A nonlinear Stokes–Biot model for the interaction of a non-Newtonian fluid with poroelastic media. ESAIM Math Model Numer Anal 2019;53(6):1915–55.
[16] Caucao S, Li T, Yotov I. A multipoint stress-flux mixed finite element method for the Stokes-Biot model. Numer Math 2022;152(2):411–73.
[17] Li T, Yotov I. A mixed elasticity formulation for fluid-poroelastic structure interaction. ESAIM Math Model Numer Anal 2022;56:1–40.
[18] Hu H. On some variational principles in the theory of elasticity and the theory of plasticity. Scientia Sinica 1955;4:33–54.
[19] Washizu K. Variational methods in elasticity and plasticity. third ed.. Pergamon Press; 1982.
[20] Braess D. Finite elements. Theory, fast solver, and applications in solid mechanics. Second Edition: Cambridge University Press; 2001.
[21] Djoko J, Lamichhane B, Reddy B, Wohlmuth B. Conditions for equivalence between the Hu-Washizu and related formulations, and computational behavior in the incompressible limit. Comput Methods Appl Mech Engrg 2006;195:4161–78.
[22] Djoko J, Reddy B. An extended Hu–Washizu formulation for elasticity. Comput Methods Appl Mech Engrg 2006;195(44–47):6330–46.
[23] Lamichhane B, Reddy B, Wohlmuth B. Convergence in the incompressible limit of finite element approximations based on the Hu-Washizu formulation. Numer Math 2006;104:151–75.
[24] Lamperti A, Cremonesi M, Perego U, Russo A, Lovadina C. A Hu–Washizu variational approach to self-stabilized virtual elements: 2D linear elastostatics. Comput Mech 2023;71(5):935–55.
[25] Wagner W, Gruttmann F. An improved quadrilateral shell element based on the Hu–Washizu functional. Adv Model Simul Eng Sci 2020;7(1):1–27.
[26] Gatica GN, Gatica LF, Stephan EP. A dual-mixed finite element method for nonlinear incompressible elasticity with mixed boundary conditions. Comput Methods Appl Mech Engrg 2007;196(35–36):3348–69.
[27] Gatica GN, Márquez A, Rudolph W. A priori and a posteriori error analyses of augmented twofold saddle point formulations for nonlinear elasticity problems. Comput Methods Appl Mech Engrg 2013;264:23–48.
[28] Gómez-Vargas B, Mardal K-A, Ruiz-Baier R, Vinje V. Twofold saddle-point formulation of Biot poroelasticity with stress-dependent diffusion. SIAM J Numer Anal 2023;63(3):1449–81.
[29] Caucao S, Gatica G, Sandoval F. A fully-mixed finite element method for the coupling of the Navier-Stokes and Darcy-Forchheimer equations. Numer Methods Partial Differential Equations 2021;37, 3:2250–587.
[30] Gatica GN, Heuer N, Meddahi S. On the numerical analysis of nonlinear twofold saddle point problems. IMA J Numer Anal 2003;23:301–30.
[31] Camaño J, García C, Oyarzúa R. Analysis of a momentum conservative mixed-FEM for the stationary Navier–Stokes problem. Numer Methods Partial Differential Equations 2021;37, 5:2895–923.
[32] Caucao S, Oyarzúa R, Villa-Fuentes S. A new mixed-FEM for steady-state natural convection models allowing conservation of momentum and thermal energy. Calcolo 2020;57(4). article: 36.
[33] Howell JS, Walkington NJ. Inf–sup conditions for twofold saddle point problems. Numer Math 2011;118(4):663–93.
[34] Arnold DN, Winther R. Mixed finite elements for elasticity. Numer Math 2002;92(3):401–19.
[35] Arnold DN, Brezzi F, Douglas J. PEERS: A new mixed finite element method for plane elasticity. Japan J Appl Math 1984;1(347):347–67.
[36] Arnold DN, Falk RS, Winther R. Mixed finite element methods for linear elasticity with weakly imposed symmetry. Math Comput 2007;76:1699–723.
[37] Brenner S, Scott L. The mathematical theory of finite element methods. New York: Springer–Verlag; 1994.
[38] Showalter RE. Diffusion in poro-elastic media. J Math Anal Appl 2000;251(1):310–40.
[39] Ateshian GA, Weiss JA. Anisotropic hydraulic permeability under finite deformation. J Biomech Eng 2010;132(11):111004(7).
[40] Bociu L, Muha B, Webster JT. Weak solutions in nonlinear poroelasticity with incompressible constituents. Nonlinear Anal RWA 2022;67:103563.
[41] Boffi D, Brezzi F, Fortin M. Mixed finite element methods and applications, vol. 44, Springer; 2013.
[42] Fu S, Chung E, Mai T. Constraint energy minimizing generalized multiscale finite element method for nonlinear poroelasticity and elasticity. J Comput Phys 2020;417:109569.
[43] Hong Q, Kraus J, Lymbery M, Philo F. A new practical framework for the stability analysis of perturbed saddle-point problems and applications. Math Comp 2023;92(340):607–34.
[44] Ern A, Guermond J-L. Theory and practice of finite elements. Applied mathematical sciences, vol. 159, New York: Springer-Verlag; 2004.
[45] Gatica GN. A simple introduction to the mixed finite element method. Theory and applications. Berlin: Springer-Verlag; 2014.
[46] Arnold D, Winther R. Nonconforming mixed elements for elasticity. Math Models Methods Appl Sci 2003;13(03):295–307.
[47] Arnold D, Awanou G, Winther R. Nonconforming tetrahedral mixed finite elements for elasticity. Math Models Methods Appl Sci 2014;24(04):783–96.
[48] Aznaran F, Farrell P, Kirby R. Transformations for Piola-mapped elements. SMAI J Comput Math 2022;8:399–437.
[49] Rathgeber F, Ham DA, Mitchell L, Lange M, Luporini F, McRae AT, et al. Firedrake: Automating the finite element method by composing abstractions. ACM Trans Math Softw 2016;43(3):1–27.
[50] Oyarzúa R, Ruiz-Baier R. Locking-free finite element methods for poroelasticity. SIAM J Numer Anal 2016;54(5):2951–73.

[51] Ruiz-Baier R, Taffetani M, Westermeyer HD, Yotov I. The Biot–Stokes coupling using total pressure: Formulation, analysis and application to interfacial flow in the eye. Comput Methods Appl Mech Engrg 2022;389:e114384(1–30).
[52] Mandel J. Consolidation des sols (étude mathématique). Geotechnique 1953;3(7):287–99.
[53] Brun MK, Ahmed E, Berre I, Nordbotten JM, Radu FA. Monolithic and splitting solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport. Comput Math Appl 2020;80(8):1964–84.
[54] Castelletto N, White JA, Tchelepi H. Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics. Int J Numer Anal Methods Geomech 2015;39(14):1593–618.
[55] Fjær E, Holt RM, Horsrud P, Raaen AM. Petroleum related rock mechanics. Elsevier; 2021.
[56] Guo L, Vardakis JC, Chou D, Ventikos Y. A multiple-network poroelastic model for biological systems and application to subject-specific modelling of cerebral fluid transport. Internat J Engrg Sci 2020;147:103204.
[57] Teichtmeister S, Mauthe S, Miehe C. Aspects of finite element formulations for the coupled problem of poroelasticity based on a canonical minimization principle. Comput Mech 2019;64:685–716.
[58] Li L, Soulhat J, Buschmann M, Shirazi-Adl A. Nonlinear analysis of cartilage in unconfined ramp compression using a fibril reinforced poroelastic model. Clin Biomech 1999;14(9):673–82.