Mathematical framework for biological tissue perfusion modeling and simulation

Mathieu Barré

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX076

Thèse de doctorat

Cadre mathématique pour la modélisation et la simulation de tissus biologiques perfusés

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à École polytechnique

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 2 octobre 2023, par

MATHIEU BARRÉ

Composition du Jury :

Erik Burman Professor, University College London	Président
Alexandre Ern Professeur, École des Ponts	Rapporteur
Marie Rognes Chief research scientist, Simula	Rapporteure
Eliane Bécache Chargée de recherche, Inria	Examinatrice
Franz Chouly Professeur, Université de Bourgogne	Examinateur
Céline Grandmont Directrice de recherche, Inria	Directrice de thèse
Philippe Moireau Directeur de recherche, Inria	Directeur de thèse
Patrick Ciarlet Professeur, ENSTA Paris	Invité

À Pascal

Remerciements

Au cours des trois années qui viennent de s'écouler, j'ai travaillé avec deux scientifiques remarquables : Céline Grandmont et Philippe Moireau. Ils ont tous les deux été extrêmement disponibles pour moi, et je mesure la chance que j'ai d'avoir eu une directrice et un directeur de thèse aussi impliqués dans mon projet. J'ai énormément appris à leurs côtés et les remercie chaleureusement pour tout ce qu'ils m'ont transmis, que ce soit sur le plan scientifique ou humain. J'ai été impressionné par l'intuition mathématique et l'expertise technique de Céline. Je salue aussi sa rigueur scientifique et son intégrité intellectuelle : j'espère que je réussirai à porter ces valeurs aussi haut qu'elle dans la suite de ma carrière. Philippe m'a révélé les secrets de la théorie des semi-groupes et initié à tant d'autres sujets. Sa vaste culture scientifique m'a profondément nourri et va beaucoup me manquer. J'ai apprécié tous ses précieux conseils et le remercie pour son enthousiasme.

J'aimerais également adresser mes sincères remerciements à Alexandre Ern et Marie Rognes qui m'ont fait l'honneur de rapporter ma thèse. Leurs commentaires fournis m'ont permis d'améliorer ce manuscrit et d'ouvrir de nombreuses perspectives que j'ai hâte d'explorer. Marie, I would also like to thank you for welcoming me in your team at Simula during my first year of master, which encouraged me to pursue a career in mathematics related to biomedical applications.

Je remercie chaleureusement Eliane Bécache, Erik Burman et Franz Chouly d'avoir accepté de faire partie de mon jury. Merci beaucoup pour l'intérêt que vous avez porté à mon travail, pour votre venue à ma soutenance et pour vos riches questions.

Pendant mon doctorat, j'ai eu l'opportunité de collaborer avec Patrick Ciarlet, que je remercie de sa présence dans mon jury. Je suis honoré par la confiance qu'il m'a accordée et je salue la patience dont il a fait preuve. Merci aussi à Céline et Philippe pour la liberté qu'ils m'ont laissée afin de mener cette collaboration.

J'ai également eu l'occasion de travailler avec Claire A. Dessalles pour confronter les résultats théoriques de ma thèse à des données réelles. Cela a été une expérience très enrichissante pour moi et j'espère pouvoir continuer à collaborer avec Claire dans le futur.

Cette thèse a été financée par le Labex de Mathématique Hadamard. Je le remercie vivement pour son soutien, ainsi que tous les contribuables français qui rendent la recherche publique possible.

J'ai conduit mon doctorat au sein de l'équipe MΞDISIM d'Inria Saclay, dans laquelle j'ai trouvé un environnement de travail exceptionnel. Je garderai de très bons souvenirs de mes discussions de modélisation mécanique et d'analyse numérique avec Dominique Chapelle et Patrick Le Tallec. La porte de leur bureau m'a toujours été ouverte quand j'avais des questions à leur poser. Je remercie aussi Jean-Marc Allain, Frédérique Clément, Martin Genet et Sébastien Imperiale avec qui j'ai pu échanger sur la notion d'incompressibilité, la poromécanique, les schémas numériques et sur tant d'autres sujets autour d'un repas ou d'une pause café. Un grand merci à Jérôme Diaz et François Kimmig qui m'ont beaucoup dépanné quand j'avais des problèmes de code, en plus d'être devenus de bons amis dont l'humour a égayé ces années.

En rejoignant l'équipe, j'ai eu le bonheur de trouver ma place parmi les doctorants. Ceci a avant tout été possible grâce à l'accueil que m'ont réservé Cécile, Chloé, Guillaume, Nicole et bien sûr la mama Jessica. J'ai noué de magnifiques relations avec mes collègues de l'open space, un lieu

chaleureux où j'espère que les valeurs de solidarité et d'entraide perdureront encore longtemps. Tiphaine, ma chère sœur de thèse, je suis très heureux de la complicité que nous avons développée et te remercie d'avoir toujours été là pour me soutenir dans les moments difficiles. Merci à André pour les délicats massages, à Giulia pour les leçons de dialecte romain, à Alice pour sa gentillesse constante en dehors du terrain de pétanque, à Louis-Pierre pour son poignet magique, et à Gaël et Zineb qui prendront la suite. Enfin, merci à Bahar pour toute son aide sur le volet administratif et aux pâtissiers du Magnan pour leurs desserts décisifs.

D'autres rencontres ont émaillé mon parcours pendant cette thèse. Je voudrais notamment remercier l'équipe COMMEDIA d'Inria Paris, en particulier Miguel Fernández pour nos échanges sur les schémas de couplage et Fabien Vergnet pour son aide avec FEniCS. Par ailleurs, je dis merci à Alain Couvreur, Magalie Quet et Marine Spaak avec qui j'ai eu beaucoup de plaisir à travailler pendant la mission de médiation scientifique que j'ai menée pour Inria. Enfin, je remercie les professeurs de l'ENSTA qui m'ont fait aimer les EDPs et l'analyse numérique, en particulier Laurent Bourgeois, Christophe Hazard et Sonia Fliss. Ces professeurs sont devenus des collègues quand je suis passé de l'autre côté en donnant des TDs dans mon ancienne école. Merci à Sonia pour avoir coordonné ces TDs et m'avoir fait confiance, ainsi qu'à Marcella Bonazzoli qui m'a accompagné dans la création d'un sujet de TP.

Puisque je parle de la période ENSTA, comment continuer ces remerciements sans mentionner l'infatigable Habitaaaaaaaat ? Merci pour tous ces moments de vie, que ce soit à Palaiseau, Paris, Pornic, Marseille, Brest ou ailleurs. À l'ENSTA, j'ai aussi croisé le chemin de Pierre et Étienne, que j'ai eu le plaisir de retrouver comme collègues et amis par la suite. Par ailleurs, j'ai rencontré des personnes merveilleuses dans l'association Animath et y ai vécu de grandes aventures. Martin, Fabrice, Vincent, Raphaël, Guillaume, Matthieu, Cécile, Victor, Sophie, Lamia, Laetitia, Théo, Aline, Ilyes et tous les nouveaux, ces lignes sont pour vous et pour toutes les crêpes de Martin et Rémi, ce qui fait beaucoup. Ma thèse a également été marquée par le COMB et sa joyeuse équipe : un grand merci pour votre accueil original et votre générosité. J'ai aussi de la gratitude pour Arnaud, Arthur, Élias, Élise, Émeline, Étienne, Isabelle, Laura, Nabil, Sophie, Théo, Thomas et les nombreuses autres personnes avec qui j'ai exploré des mondes à venir. Enfin, les unions de Katia et Philipp, Zélia et Pierre ainsi que Constance et Lucas ont été de très beaux moments que j'ai aimé partager avec Yohann, JB, Galaad, Xavier, Cassandre, Lucas, Alain et Edwige pour ne citer qu'eux.

Je n'aurais pas pu terminer cette thèse sans le soutien indéfectible de ma famille. Maman, Papa et Maëlle, vous avez tous les trois été un élément fondamental dans ma réussite : merci pour tout. Julien, j'espère que tu continueras à veiller sur moi malgré la distance. Llewellyn, Tynawedd, Dwynwen, Hywel, Gwrvan, Aelwyd, Llawgad, Judwall, Arthvawr, le temps passé avec vous est toujours un bonheur trop court. Irène, Didier, Paul et Zoé, merci de m'avoir accueilli dans votre famille.

Mes dernières pensées vont à celles et ceux qui sont partis mais qui auraient été fiers de moi aujourd'hui. À celui qui nous aurait régalé de sa cuisine généreuse lors de mon pot de thèse, et à celui qui aurait sans aucun doute affiché mon diplôme de doctorat dans sa cuisine en bombant le torse.

Enfin, merci à toi qui te reconnaîtras, toi qui n'as jamais renoncé à la beauté du monde et qui m'a offert ton amour.

Table des matières

Introduction

This PhD thesis has been prepared in the MΞDISIM team at Inria and École polytechnique, funded by the LabEx Mathématique Hadamard, under the supervision of Céline Grandmont and Philippe Moireau.

Biological tissues modeled as porous media

In the human body, most biological tissues are composed of elastic components that interact with fluids. This structure of human organs has been revealed by anatomists over the centuries [Singer, 1925] as illustrated in Figure 1, and occurs both at microscopic and macroscopic scales. At the microscopic scale, biological tissues are constituted by cells surrounded by an extracellular matrix that is mainly made of water and proteins, see [Frantz et al., 2010] for further details. Each tissue has a specific extracellular matrix, whose composition and topology determine its macroscopic mechanical properties. For instance, the heart muscle is made of various biological fibers and is supplied in blood by the coronary circulation, namely the capillary vessels surrounding the heart represented in Figure 2. In the lungs, blood and gas exchanges occur at very small scales in the alveoli that are irrigated by the bronchial tree. As for the brain, its parenchyma is filled by the cerebrospinal fluid that plays a central role in the brain clearance process – see Figure 3 – and in protecting the brain tissue from injury. Such interaction of solids and fluids in biological tissues is called perfusion.

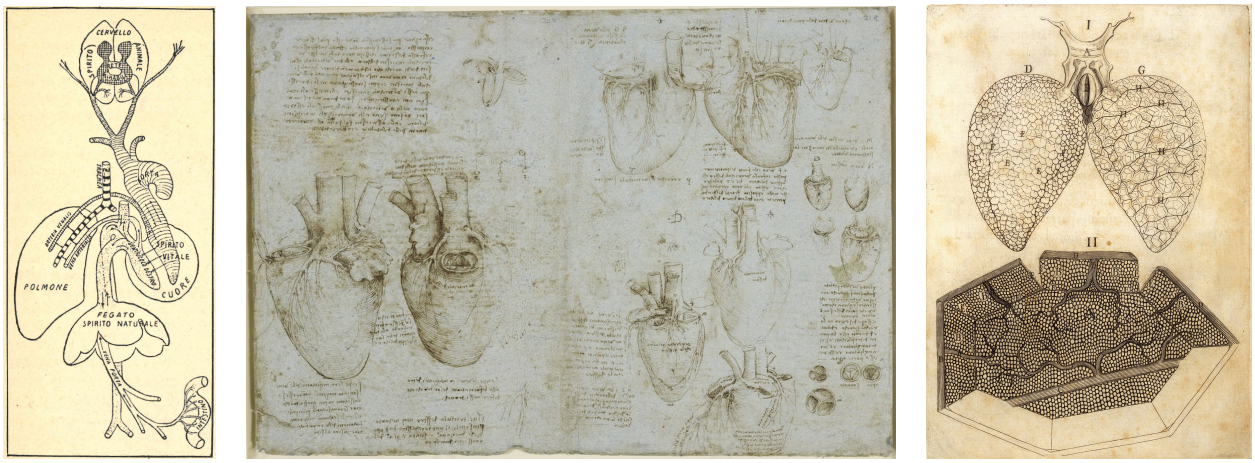


Figure 1 – (Left) Human physiology according to Galen (129 – 216) [Singer, 1925]. (Middle) Studies of the coronary vessels and heart valves by Leonardo da Vinci (1452 – 1519). Credits: British Royal Collection. (Right) Lung observation by Marcello Malpighi (1628 – 1694). Credits: Welcome Library, London.

Let us mention that all the fluids involved in the forementioned tissues – blood, pulmonary airflows, cerebrospinal fluid – can be considered as incompressible, see [Baffico et al., 2010] for a justification of this assumption for the airflows in the lungs. Therefore, many biological tissues in the human body are perfused by incompressible fluids. The perfusion of such tissues is crucial to understand their mechanical behavior and its dysfunction can lead to severe pathologies. Hence a heart perfusion malfunction, which can for example be caused by the occlusion of a coronary vessel as in Figure 2, is responsible for 20% of deaths in Europe [Nichols et al., 2014]. For the lungs, pathologies may affect the tissue parenchyma as in fibrosis [Nunes et al., 2015] or induce an inflammation at the alveolar level as for instance in COVID19 [Zhang et al., 2020]. Lastly, if the brain perfusion is disturbed, waste clearance is altered, which may be a cause of Alzheimer's disease [Iliff et al., 2013; Stone et al., 2015].

From the modeling point of view, a solid that is irrigated by a fluid is called a porous medium. The study of the mechanical behavior of porous media is named poromechanics [Coussy, 2004].

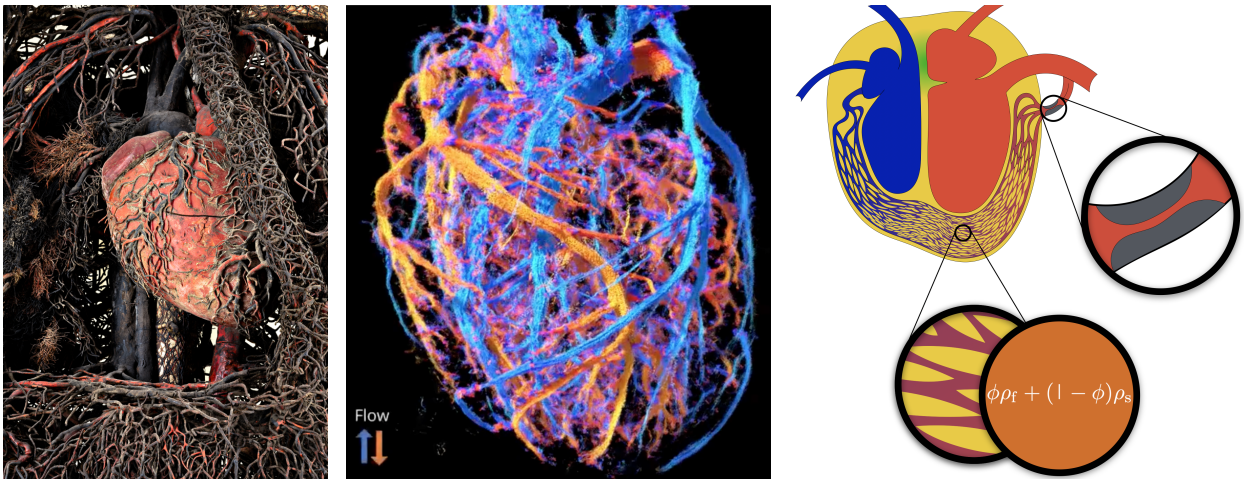


Figure 2 – (Left) Microcirculation networks in a human chest. Credits: Museo Sansevero, Naples. (Middle) Coronary flow in beating hearts [Demeulenaere et al., 2022]. (Right) Porous modeling of the coronary network and zoom on a coronary occlusion. Courtesy from [Burtschell, 2016].

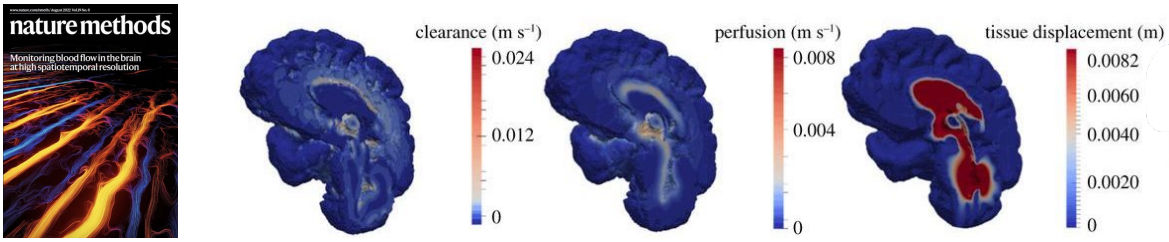


Figure 3 – (Left) *Nature methods* cover "Monitoring blood flow in the brain at high spatiotemporal resolution", August 2022. (Right) Simulation of CSF/ISF clearance, blood perfusion and parenchymal tissue displacement in the brain using multiple-network poroelastic theory [Guo et al., 2018].

Historically, the field of poromechanics was first developed for geosciences [Biot, 1941; Terzaghi, 1943] and only more recently did it focus on biomedical applications. The role of porous effects in the mechanical response of biological tissues is now clearly established, see the review [Khaled and Vafai, 2003] and references therein. Porous models have been used to simulate many perfused organs, starting with the heart [Yang and Taber, 1991; Huyghe et al., 1992; Nash and Hunter, 2000; Chapelle et al., 2010; Michler et al., 2013; Chabiniok et al., 2016; Di Gregorio et al., 2021; Barnafi Wittwer et al., 2022; Chabiniok et al., 2022]. Concerning the lungs, [Patte et al., 2022] considers a non-linear quasi-static porous model and formulates physiological boundary conditions, while [Wall et al., 2010] and [Berger et al., 2016] have proposed multiscaled coupled models of the ventilation process where the lung parenchyma is described by a porous media, and the use of poromechanics for patient-specific fibrosis diagnosis is considered in [Genet et al., 2020]. In the brain, poroelasticity can be used to model the waste clearance process [Basser, 1992; Tully and Ventikos, 2011; Vardakis et al., 2016; Guo et al., 2018; Lee et al., 2019; Kedarasetti et al., 2022] but also drug transport [Støverud et al., 2011] and stenosis [Chou et al., 2016]. When coupled with a fluid flow, porous models showed their ability to simulate lipid and drug transport in blood vessel walls [Koshiba et al., 2007; Calo et al., 2008; Badia et al., 2009; D'Angelo and Zunino, 2011; Čanić et al., 2021] and the interfacial flow in the eye [Boon et al., 2022; Ruiz-Baier et al., 2022]. More broadly, the poroelasticity framework was applied to tissue growth [Ambrosi and Preziosi, 2002; Armstrong et al., 2016; Sacco et al., 2017; Deville et al., 2018], to eye perfusion [Causin et al., 2014] and to tongue vascularization [Qohar et al., 2021].

In this context, the goal of this thesis is to

Develop a mathematical framework for biological tissue perfusion modeling and simulation.

To contribute to this broad objective, this thesis focuses on the linearized version of a non-linear poromechanics model formulated in [Chapelle and Moireau, 2014]. Compared to the forementioned works, the non-linear model from [Chapelle and Moireau, 2014] has two specificities. The first one is that it is valid for fast porous flows and large deformations. Hence it is particularly adapted to the perfusion of organs that are subject to large deformations such as the heart and the lungs. The second one is that it preserves energy estimates, which is an important physical feature and will be a key ingredient for the theoretical and numerical analysis performed in this thesis on its linearized version. This model is presented below, as well as its linearized counterpart.

Perfusion porous modeling

To understand the theoretical and numerical properties of the non-linear poromechanics model from [Chapelle and Moireau, 2014], a first step is to study its linearized version. The mathematical analysis and the formulation of numerical methods together with their numerical analysis for the corresponding linearized poromechanics model is the main topic of this thesis. In what follows, we first recall the governing equations derived in [Chapelle and Moireau, 2014]. These equations were obtained by revisiting Biot theory of poroelasticity [Biot, 1941, 1955; Coussy, 2004; Dormieux et al., 2006] for finite strains in order to retrieve a non-linear formulation compatible with the principles of thermodynamics, leading to a generic energy balance. Then, we show how to linearize this model. Note that such a linearization process was performed in [Burtschell et al., 2019] for a porous medium satisfying Terzaghi's effective stress principle. Here, the linearization is carried out without assuming that this principle is satisfied. This leads to a model where the Biot-Willis coefficient may be different from 1, which finally allows us to make the link between the resulting linearized model to the standard Biot equations. This linearization procedure was published as an appendix in [Barré et al., 2023].

A general finite strain poromechanics formulation adapted to biological soft tissue perfusion

We consider a deformable porous medium that occupies the space domain $\Omega(t)$ at time t. The deformed domain is obtained from a reference configuration domain $\hat{\Omega} \subset \mathbb{R}^d$, namely $\Omega(t) = \hat{\mathcal{A}}(\hat{\Omega})$, where

$$\hat{\mathcal{A}}(\cdot, t) \; : \; \left| \begin{array}{l} \hat{\Omega} \longrightarrow \Omega(t), \\ \hat{x} \longmapsto x = \hat{x} + \hat{u}_s(\hat{x}, t), \end{array} \right.$$

denotes the deformation mapping and $\hat{u}_s$ is the displacement field defined in the reference configuration, see Figure 4. We then introduce usual mechanical quantities in the reference configuration such as the deformation gradient tensor $\hat{F} = \nabla \hat{\mathcal{A}} = \hat{\mathbb{I}} + \nabla \hat{u}_s$, the Cauchy-Green deformation tensor $\hat{C} = \hat{F}^T \hat{F}$, the Green-Lagrange strain tensor $\hat{E} = \frac{1}{2}(\hat{C} - \hat{\mathbb{I}})$, and the apparent change of volume of the material $\hat{J} = \det \hat{F}$.

By convention, we use a hat superscript for lagrangian quantities (defined in the reference configuration), and no superscript for the corresponding eulerian quantities (defined in the deformed configuration). For instance, J is the function satisfying

$$J(x, t) = J\big(\hat{\mathcal{A}}(\hat{x}, t), t\big) = \hat{J}(\hat{x}, t).$$

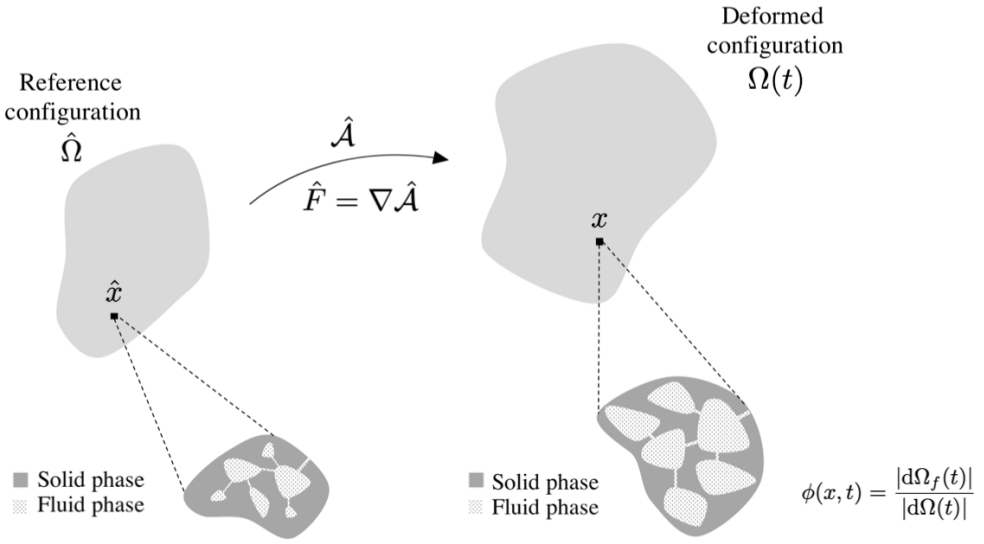


Figure 4 – Porous medium in deformed and reference configurations. Adapted from [Patte, 2020].

The porous medium is modeled as a mixture of a fluid phase and a solid phase called the skeleton. At each point of the deformed domain, we assume that the material contains a volume fraction ϕ of fluid and $1-\phi$ of solid, with $\phi(x,t)$ the porosity, and that the fluid and solid parts interact with each other. Moreover, we suppose that the fluid is homogeneous and incompressible, as it is the case in most of biomedical applications. Hence its density is independent of space and time.

The porous medium motion is described by the following set of unknows: the porosity ϕ mentioned previously (denoted by $\hat{\phi}$ in the reference configuration), the solid displacement $\hat{u}_s$, the solid velocity $\hat{v}_s$ in the reference configuration (denoted by v_s in the deformed configuration), the fluid velocity v_f in the deformed configuration (denoted by $\hat{v}_f$ in the reference configuration) and the interstitial pressure p (denoted by $\hat{p}$ in the reference configuration), namely the fluid pressure in the pores.

Moreover, let us denote by $\hat{\phi}_0(\hat{x}) = \hat{\phi}(\hat{x},0)$ the initial porosity in the reference configuration, by $\hat{\rho}_s$ the solid density in the reference configuration, by ρ_f the fluid density in the deformed configuration, by k_f the hydraulic conductivity tensor in the deformed configuration – and $\hat{k}_f$ in the reference configuration – which represents the friction between the fluid and solid phases, by

$$\hat{m} = \rho_f(\hat{J}\hat{\phi} - \hat{\phi}_0) \tag{1}$$

the added fluid mass per unit volume in the reference configuration, by f an exterior body force applied to the material and by θ a distributed fluid mass source term.

The macroscopic governing equations derived in [Chapelle and Moireau, 2014] for a general poromechanics formulation valid in large strains and adapted to soft tissue perfusion then read:

$$\begin{cases} \hat{\rho}_s(1-\hat{\phi})\dfrac{\partial \hat{v}_s}{\partial t} - \operatorname{div}\left(\hat{F}\,\hat{\Sigma}_s(\hat{u}_s,\hat{v}_s,\hat{p})\right) \\ \qquad - \hat{J}\hat{\phi}^2\,\hat{k}_f^{-1}(\hat{v}_f - \hat{v}_s) + \hat{p}\hat{J}\hat{F}^{-T}\,\nabla\hat{\phi} = \hat{\rho}_s(1-\hat{\phi})\,\hat{f}, & (\hat{x},t)\in\hat{\Omega}\times(0,T), \quad (2a) \\ \dfrac{1}{J}\dfrac{\partial}{\partial t}(J\rho_f\phi\,v_f)\big|_{\hat{x}} + \operatorname{div}\left(\rho_f\phi\,v_f\otimes(v_f-v_s)\right) \\ \qquad - \operatorname{div}\left(\phi\,\sigma_f^{\mathrm{tot}}(v_f,p)\right) + \phi^2\,k_f^{-1}(v_f-v_s) - \theta\,v_f = \rho_f\phi\,f, & (x,t)\in\Omega(t)\times(0,T), \quad (2b) \\ \dfrac{1}{J}\dfrac{\partial m}{\partial t}\Big|_{\hat{x}} + \operatorname{div}\left(\rho_f\phi(v_f-v_s)\right) = \theta, & (x,t)\in\Omega(t)\times(0,T). \quad (2c) \end{cases}$$

In this system, the first two equations correspond respectively to the solid and fluid phases momentum balance equations, and the third one to the mass conservation equation. These equations are complemented by the following constitutive laws. From [Chapelle and Moireau, 2014, Theorem 6], a general form of the fluid Cauchy stress tensor σ_f^{tot} is given by $\sigma_f^{\text{tot}} = \sigma_f(v_f) - p\,\mathbb{I}$, with

$$\sigma_f(v) = \lambda_f \operatorname{Tr}(\varepsilon(v))\,\mathbb{I} + 2\mu_f\,\varepsilon(v), \qquad \varepsilon(v) = \frac{1}{2}\big(\nabla v + \nabla v^T\big),$$

and the skeleton contribution to the second Piola-Kirchhoff stress tensor is given by

$$\hat{\Sigma}_s = \frac{\partial \hat{\Psi}_s}{\partial \hat{E}} + \frac{\partial \hat{\Psi}_{damp}}{\partial \dot{\hat{E}}} + \hat{\phi}\,\hat{p}\hat{J}\hat{C}^{-1}, \tag{3}$$

where $\hat{\Psi}_s$ denotes the skeleton free energy density potential and $\hat{\Psi}_{damp}$ is a viscous pseudo-potential, here chosen as a function of $\dot{\hat{E}}$ for simplicity. Finally, noting that $\hat{J}_s = (1-\hat{\phi})\hat{J}$, the system is closed by the relation

$$\hat{p} = -\frac{\partial \hat{\Psi}_s}{\partial \hat{J}_s}. \tag{4}$$

One fundamental property of such model is its thermodynamical compatibility. Mathematically speaking, this imply that the solution of (2) satisfies an energy balance as proved in [Chapelle and Moireau, 2014, Theorem 7]. Assuming in our case homogeneous Dirichlet conditions for the fluid and the solid, we have

$$\begin{aligned}\frac{\mathrm{d}\mathcal{K}}{\mathrm{d}t} + \frac{\mathrm{d}\widehat{\mathcal{W}}_s}{\mathrm{d}t} = &- \int_{\hat{\Omega}} \frac{\partial \hat{\Psi}_{damp}}{\partial \dot{\hat{E}}} : \dot{\hat{E}}\,\mathrm{d}\hat{x} - \int_{\Omega(t)} \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x \\ &- \int_{\Omega(t)} \phi^2 k_f^{-1}(v_f - v_s)\cdot(v_f - v_s)\,\mathrm{d}x + \mathcal{P}_{ext}^{total} + \mathcal{J}_{\mathcal{K}\theta} + \mathcal{J}_{\mathcal{G}\theta},\end{aligned} \tag{5}$$

with

$$\mathcal{K} = \frac{1}{2}\int_{\Omega(t)} \rho_s(1-\phi)\,|v_s|^2\,\mathrm{d}x + \frac{1}{2}\int_{\Omega(t)} \rho_f\phi\,|v_f|^2\,\mathrm{d}x, \qquad \widehat{\mathcal{W}}_s = \int_{\hat{\Omega}} \hat{\Psi}_s\,\mathrm{d}\hat{x}$$

the mixture's kinetic energy and the skeleton free energy,

$$\mathcal{J}_{\mathcal{K}\theta} = \frac{1}{2}\int_{\Omega(t)} \theta\,|v_f|^2\,\mathrm{d}x, \qquad \mathcal{J}_{\mathcal{G}\theta} = \int_{\Omega(t)} \frac{p}{\rho_f}\,\theta\,\mathrm{d}x,$$

the incoming rates of fluid kinetic energy and Gibbs free energy, and

$$\mathcal{P}_{ext}^{total} = \int_{\Omega(t)} \rho_f\phi f\cdot v_r\,\mathrm{d}x + \int_{\Omega(t)} \rho f\cdot v_s\,\mathrm{d}x = \int_{\Omega(t)} \rho_f\phi f\cdot v_f\,\mathrm{d}x + \int_{\Omega(t)} \rho_s(1-\phi) f\cdot v_s\,\mathrm{d}x,$$

the power of external forces, with $v_r = v_f - v_s$ the relative velocity between the fluid and the solid and $\rho = \rho_s(1-\phi) + \rho_f\phi$ the porous medium total density. One benefit of considering the linearized version of (2) is to keep such energy balance after linearization.

But before linearizing the coupled system, let us specify the skeleton constitutive law. Following the guidelines of [Chapelle and Moireau, 2014, Section 5.4], we consider a free energy density potential of the form

$$\hat{\Psi}_s = \widehat{W}^{skel}(\hat{E}) + \widehat{W}^{bulk}\left(\hat{J}_s\frac{1-\hat{\phi}_0}{\hat{\chi}_s(\hat{J})}\right), \tag{6}$$

where $\hat{\chi}_s(\hat{J})$ is a function representing the variations of solid volume directly due to macroscopic volume changes in the absence of pore pressure. In other words, (6) means that the energy bulk

term depends on the ratio between the change of volume for the solid part $\hat{J}_s$ and the change of volume $\hat{\chi}(\hat{J}) = \frac{\hat{\chi}_s(\hat{J})}{1-\hat{\phi}_0}$ occuring in each cell of the microstructure when assuming that the pore pressure is constant during the deformation. We suppose that $\hat{\chi}(\hat{J})$ is affine with respect to $\hat{J}$ and that $\hat{\chi}(\hat{J}) = 1$ when $\hat{J} = 1$, namely

$$\hat{\chi}(\hat{J}) - 1 = \beta(\hat{J} - 1) \qquad \Leftrightarrow \qquad \hat{\chi}_s(\hat{J}) = (1 - \hat{\phi}_0)\big(1 + \beta(\hat{J} - 1)\big),$$

with $0 \leq \beta < 1$ a coefficient vanishing for incompressible materials.

Combining (4) and (6), we see that

$$\hat{p} = -\frac{\partial \widehat{W}^{bulk}}{\partial \hat{J}_s}.$$

From (3), it follows that

$$\hat{\Sigma}_s = \frac{\partial \widehat{W}^{skel}}{\partial \hat{E}} - \hat{p}\frac{\partial \hat{J}_s}{\partial \hat{E}} + \frac{\partial \widehat{W}^{bulk}}{\partial \hat{J}} \cdot \frac{\partial \hat{J}}{\partial \hat{E}} + \frac{\partial \hat{\Psi}_{damp}}{\partial \dot{\hat{E}}} + \hat{\phi}\,\hat{p}\hat{J}\hat{C}^{-1}.$$

Observing that $\frac{\partial}{\partial \hat{E}}(\hat{J}\hat{\phi}) = \frac{\partial}{\partial \hat{E}}(\frac{\hat{m}}{\rho_f} + \hat{\phi}_0) = 0$ because $\hat{m}$ corresponds to the fluid mass entering into the pores in the reference configuration and thus does not depend on the deformation of the material, we get

$$\frac{\partial \hat{J}_s}{\partial \hat{E}} = \frac{\partial \hat{J}}{\partial \hat{E}} = \hat{J}\hat{C}^{-1}.$$

Since $\frac{\partial \widehat{W}^{bulk}}{\partial \hat{J}} = \frac{\hat{J}_s \hat{\chi}_s'(\hat{J})}{\hat{\chi}_s(\hat{J})}$, we obtain

$$\hat{\Sigma}_s = \frac{\partial \widehat{W}^{skel}}{\partial \hat{E}} + \frac{\partial \hat{\Psi}_{damp}}{\partial \dot{\hat{E}}} - \left(1 - \frac{\hat{J}_s \hat{\chi}_s'(\hat{J})}{\hat{\chi}_s(\hat{J})} - \hat{\phi}\right)\hat{p}\hat{J}\hat{C}^{-1}. \tag{7}$$

Note that this expression is valid for any potentials $\widehat{W}^{skel}$ and $\hat{\Psi}_{damp}$. To further develop the computations, we can for example use a Ciarlet-Geymonat-like potential [Ciarlet, 1988] for the bulk potential $\widehat{W}^{bulk}$, which yields

$$\widehat{W}^{bulk}\left(\hat{J}_s \frac{1 - \hat{\phi}_0}{\hat{\chi}_s(\hat{J})}\right) = \hat{\gamma}\,\kappa\left(\frac{\hat{J}_s}{\hat{\chi}_s(\hat{J})} - 1 - \log\left(\frac{\hat{J}_s}{\hat{\chi}_s(\hat{J})}\right)\right),$$

where κ denotes the solid grains bulk modulus and $\hat{\gamma}$ is a scaling factor. In order to recognize the storage coefficient in the pressure equation, we choose from now on $\hat{\gamma} = \frac{1-\hat{\phi}}{1-\beta}$. Using (4), we get

$$\hat{p} = -\frac{\partial \widehat{W}^{bulk}}{\partial \hat{J}_s} = -\frac{(1-\hat{\phi})\kappa}{1-\beta}\left(\frac{1}{\hat{\chi}_s(\hat{J})} - \frac{1}{\hat{J}_s}\right) = \frac{(1-\hat{\phi})\kappa}{1-\beta} \cdot \frac{\hat{\chi}_s(\hat{J}) - \hat{J}_s}{\hat{J}_s \hat{\chi}_s(\hat{J})}.$$

Remarking that $\hat{\chi}_s(\hat{J}) - \hat{J}_s = \hat{J}\hat{\phi} - \hat{\phi}_0 + \beta(1-\hat{\phi}_0)(\hat{J}-1) + 1 - \hat{J} = \rho_f^{-1}\hat{m} - \big(1 - \beta(1-\hat{\phi}_0)\big)(\hat{J}-1)$, we obtain

$$\hat{p} = \frac{\kappa}{(1-\beta)(1-\hat{\phi}_0)} \cdot \frac{\rho_f^{-1}\hat{m} - \big(1 - \beta(1-\hat{\phi}_0)\big)(\hat{J}-1)}{\hat{J}\big(1 + \beta(\hat{J}-1)\big)}. \tag{8}$$

This closure relation will be the cornerstone of the linearization process. As a matter of fact, it involves the interstitial pressure $\hat{p}$, the fluid added mass $\hat{m}$ that is related to the porosity $\hat{\phi}$, and the change of volume $\hat{J}$ that is close to 1 when linearizing the coupled system.

The linearized model

We linearize (2) for infinitesimal transformations around the configuration $(\hat{u}_s, \hat{v}_s, v_f, \hat{\phi}) = (0, 0, 0, \hat{\phi}_0)$. In particular we have

$$\hat{\mathcal{A}} = \hat{\mathbb{I}} + \mathcal{O}\big(|\nabla \hat{u}_s|\big), \quad \hat{E} = \hat{\varepsilon} + \mathcal{O}\big(|\nabla \hat{u}_s|^2\big),$$

where $|\nabla \hat{u}_s|^2 = \nabla \hat{u}_s : \nabla \hat{u}_s$, with the notation

$$A : B = \sum_{i,j=1}^{d} A_{ij} B_{ij}, \quad \forall A \in \mathcal{M}_d(\mathbb{R}), \forall B \in \mathcal{M}_d(\mathbb{R}).$$

Thus, the reference and deformed configurations reduce to a single domain, which will be denoted by Ω, allowing us to drop from now on the hat superscripts used previously to distinguish variables defined on $\hat{\Omega}$ or $\Omega(t)$. Besides, it holds

$$J = 1 + \operatorname{Tr} \varepsilon + \mathcal{O}\big(|\varepsilon|^2\big) = 1 + \operatorname{div} u_s + \mathcal{O}\big(|\nabla u_s|^2\big)$$

and, in virtue of (8) and (1),

$$\phi = \phi_0 + \mathcal{O}\big(|(\operatorname{div} u_s, p)|\big).$$

Furthermore, any choice of potentials $W^{skel}(E)$ and $\Psi_{damp}(\dot{E})$ satisfies

$$\frac{\partial W^{skel}}{\partial E} = \lambda \operatorname{Tr}(\varepsilon)\, \mathbb{I} + 2\mu\, \varepsilon + \mathcal{O}\big(|\nabla u_s|^2\big) \text{ and } \frac{\partial \Psi_{damp}}{\partial \dot{E}} = \nu \operatorname{Tr}(\dot{\varepsilon})\, \mathbb{I} + 2\eta\, \dot{\varepsilon} + \mathcal{O}\big(|\nabla v_s|^2\big),$$

for some Lamé constants λ, μ, ν and η, where η represents the solid grains viscosity. To simplify, we suppose that $\nu = 0$, but note that choosing $\nu = \lambda^* > 0$ would mean taking into account secondary consolidation effects as in [Murad and Cushman, 1996].

Since

$$\frac{J_s \chi_s'(J)}{\chi_s(J)} = \frac{J_s \beta (1 - \phi_0)}{\chi_s(J)} \quad \text{and} \quad \frac{J_s}{\chi_s(J)} = 1 + \mathcal{O}\big(|(\operatorname{div} u_s, p)|\big),$$

the solid stress tensor expression (7) implies that

$$\sigma_s = \lambda \operatorname{Tr}(\varepsilon)\ \mathbb{I} + 2\mu\, \varepsilon + 2\eta\, \dot{\varepsilon} - \big(1 - \beta(1 - \phi_0) - \phi_0\big) p\, \mathbb{I} + \mathcal{O}\big(|(\nabla u_s, \nabla v_s, p)|^2\big).$$

The Biot-Willis coefficient, denoted by α, is then defined as the coefficient multiplying the pressure term in the porous medium linearized total stress tensor $\Sigma = \sigma_s + \phi_0\, \sigma_f^{\text{tot}}$. Therefore

$$\alpha = 1 - \beta(1 - \phi_0), \tag{9}$$

and (2a) becomes

$$\begin{aligned} \rho_s (1 - \phi_0)\, \partial_t v_s - \operatorname{div} \Big(\lambda \operatorname{Tr}(\varepsilon(u_s))\ \mathbb{I} + 2\mu\, \varepsilon(u_s) + 2\eta\, \varepsilon(v_s) \Big) \\ - \ \phi_0^2\, k_f^{-1} (v_f - v_s) + (\alpha - \phi_0)\, \nabla p = \rho_s (1 - \phi_0) f, \quad \forall (x,t) \in \Omega \times (0,T). \end{aligned} \tag{10}$$

The fluid equation readily results from the linearization of (2b), leading to

$$\rho_f \phi_0\, \partial_t v_f - \operatorname{div} \big(\phi_0\, \sigma_f(v_f) \big) + \phi_0^2\, k_f^{-1} (v_f - v_s) - \theta\, v_f + \phi_0\, \nabla p = \rho_f \phi_0\, f, \text{ in } \Omega \times (0,T). \tag{11}$$

To recover a mass balance equation, we infer from (8) that

$$\frac{(1 - \beta)(1 - \phi_0)}{\kappa}\, p = \frac{\rho_f^{-1} m - \alpha \operatorname{div} u_s + \mathcal{O}\big(|\nabla u_s|^2\big)}{1 + \mathcal{O}\big(|\operatorname{div} u_s|\big)}.$$

Since $(1-\beta)(1-\phi_0) = \alpha - \phi_0$ and $p\,\mathrm{div} u_s = \mathcal{O}\big(|(\nabla u_s, p)|^2\big)$, it follows that

$$\rho_f^{-1} m = \frac{\alpha - \phi_0}{\kappa}\, p + \alpha\, \mathrm{div}\, u_s + \mathcal{O}\big(|(\nabla u_s, p)|^2\big),$$

Differentiating this relation with respect to time and using (2c), we finally get

$$\frac{\alpha - \phi_0}{\kappa}\, \partial_t p + \mathrm{div}\,\big((\alpha - \phi_0) v_s + \phi_0 v_f\big) = \rho_f^{-1}\theta, \quad \forall (x,t) \in \Omega \times (0,T). \tag{12}$$

The coefficient $\frac{\alpha-\phi_0}{\kappa}$ is known as the storage coefficient and is often denoted by c_0. The Biot-Willis coefficient, given by (9), is usually computed by the formula

$$\alpha = 1 - \frac{\kappa_0}{\kappa},$$

with κ_0 the drained bulk modulus. Hence $\kappa_0 = \beta(1-\phi_0)\kappa$ – see (9), and we have in particular

$$\kappa \to +\infty \Leftrightarrow \alpha \to 1 \Leftrightarrow \beta \to 0 \Leftrightarrow \chi_s(J) \to 1 - \phi_0,$$

so that (6) reduces to

$$\Psi_s = W^{skel}(E) + W^{bulk}(J_s)$$

for nearly-incompressible materials. Such a decomposition for Ψ_s corresponds to a material satisfying Terzaghi's effective stress principle, for which the microstructure Poisson effects are not taken into account. This was the assumption made in [Burtschell et al., 2019].

Gathering (10), (11) and (12), the linearized model reads:

$$\begin{cases} \rho_s(1-\phi_0)\,\partial_t v_s - \mathrm{div}\,\Big(\lambda\,\mathrm{Tr}(\varepsilon(u_s))\,\mathbb{I} + 2\mu\,\varepsilon(u_s) + 2\eta\,\varepsilon(v_s)\Big) \\ \qquad\qquad - \phi_0^2\, k_f^{-1}(v_f - v_s) + (\alpha - \phi_0)\,\nabla p = \rho_s(1-\phi_0) f, \quad \forall (x,t) \in \Omega \times (0,T), \quad \text{(13a)} \\ \rho_f \phi_0\,\partial_t v_f - \mathrm{div}\,\Big(\lambda_f \phi_0\,\mathrm{Tr}(\varepsilon(v_f))\,\mathbb{I} + 2\mu_f \phi_0\,\varepsilon(v_f)\Big) \\ \qquad\qquad + \phi_0^2\, k_f^{-1}(v_f - v_s) - \theta\, v_f + \phi_0\,\nabla p = \rho_f \phi_0\, f, \qquad \forall (x,t) \in \Omega \times (0,T), \quad \text{(13b)} \\ \dfrac{\alpha - \phi_0}{\kappa}\,\partial_t p + \mathrm{div}\,\big((\alpha - \phi_0) v_s + \phi_0 v_f\big) = \rho_f^{-1}\theta, \qquad \forall (x,t) \in \Omega \times (0,T). \quad \text{(13c)} \end{cases}$$

Note that in (13), the porosity ϕ_0 is no longer an unknown of the model since the non-linear system (2) is linearized around a given porosity. Hence, in the rest of the thesis, the porosity will be considered as a given data depending only on space, and will be denoted by ϕ to simplify the notation.

The theoretical and numerical analysis of Problem (13) is the main goal of this thesis. In particular, this thesis focuses on the incompressible case for which $\kappa = +\infty$ and the pressure equation (13c) reduces to the incompressibility constraint

$$\mathrm{div}\,\big((\alpha - \phi_0) v_s + \phi_0 v_f\big) = \rho_f^{-1}\theta. \tag{14}$$

Before recalling the literature related to this system, let us connect it to more standard porous media models.

Although it is obtained by linearizing a recent poromechanics model, system (13) is in fact strongly related to commonly used porous models such as Biot equation. More precisely, in Chapter 1, we show thanks to an appropriate change of variable that if $\eta = \mu_f = \lambda_f = 0$, then (13) reduces to

$$\begin{cases} \rho\,\partial_{tt}^2 u_s + \rho_f \phi\,\partial_t v_r - (\lambda + \mu)\nabla(\mathrm{div}\, u_s) - \mu\,\Delta u_s + \alpha\,\nabla p = \rho f, \\ \rho_f \partial_{tt}^2 u_s + \rho_f \partial_t v_r + \phi\, k_f^{-1} v_r + \nabla p = \rho_f f, \\ \partial_t(c_0 p + \alpha\,\mathrm{div}\, u_s) + \mathrm{div}\,\big(\phi\, v_r\big) = \rho_f^{-1}\theta, \end{cases} \tag{15}$$

where we recall that $v_r = v_f - v_s$ denotes the relative velocity between the fluid and the solid, $\rho = \rho_s(1-\phi) + \rho_f\phi$ the porous medium total density and $c_0 = (\alpha - \phi)/\kappa$ the storage coefficient. Problem (15) corresponds exactly to Biot poroacoustic equations [Biot, 1956b].

Therefore, system (13) can also be interpreted as a poroacoustic model with additional viscosity effects in both fluid and solid phases. Moreover, note that if the fluid inertial effects are neglected, (15) further simplifies into

$$\begin{cases} \rho\,\partial_{tt}^2 u_s - (\lambda+\mu)\nabla(\operatorname{div} u_s) - \mu\,\Delta u_s + \alpha\nabla p = g, \\ \partial_t(c_0 p + \alpha \operatorname{div} u_s) - \operatorname{div}(k_f \nabla p) = h, \end{cases} \tag{16}$$

which is nothing else than Biot's consolidation model, one of the most widely used poromechanics models. We refer to Chapter 3 for further details.

Before going into the contributions of this thesis, we give a short state of the art on modeling, theoretical and numerical aspects of porous media problems.

State of the art

From the mechanical modeling point of view, the first works dealing with porous media subject to finite strains arised from mixture theory [Bowen, 1980], a theory in which the fluid and solid constituents of the porous medium are treated equally. Mixture theory formulations allowing large structural deformations and strong inertial effects were proposed in [Wilmanski, 2005; Rajagopal and Tao, 2005] but without a full thermodynamical justification since they do not satisfy Clausius-Duhem inequality. In the framework of Biot theory in which the solid skeleton plays a special role, also referred to as Theory of Porous Media (TPM), such formulations were given in [Bourgeois, 1997; Lopatnikov and Cheng, 2004; Gajo and Denzer, 2011]. Yet, these models assume that the fluid viscous effects within the fluid can be neglected with respect to the frictional effects between the two phases, and in [Gajo and Denzer, 2011] the fluid is supposed to be compressible. Therefore, to our knowledge, the non-linear poromechanics model derived in [Chapelle and Moireau, 2014] is the first one to take into account inertial and viscous effects both for the incompressible fluid and the solid consituting the porous medium. Moreover, it is compatible with the laws of thermodynamics. Let us mention that a new model based on [Chapelle and Moireau, 2014] was designed in [Vuong et al., 2015], and that recent extensions of the TPM were considered for the study of subcutaneous injection [Gil, 2020; Gil et al., 2022] and fluid-porous structure interaction [Zakerzadeh and Zunino, 2019].

Concerning the mathematical studies of non-linear porous systems, existence and uniqueness results are available. These works include non-linear constitutive laws [Showalter and Stefanelli, 2004; Barucq et al., 2005], permeability depending on pressure [Showalter and Su, 2001] or porosity and solid dilatation [Tavakoli and Ferronato, 2013; Bociu et al., 2016; Bociu and Webster, 2021; Bociu et al., 2022], coupling with heat flow [Brun et al., 2019] and fluid-porous structure interaction [Benešová et al., 2023]. The existence and uniqueness of Biot-type problems, namely systems of the form (15) or (16), has also been largely studied. For the unsteady Biot's consolidation model, namely (16) with $\rho > 0$, the existence of strong solutions goes back to [Dafermos, 1968] where the proof relies on Laplace transform. The existence of weak solutions was then shown in [Barucq et al., 2004] thanks to a Galerkin method together with a regularization technique. Biot's quasi-static equation, namely (16) with $\rho = 0$, was analyzed with homogenization theory in [Auriault, 1980], leading to the existence of strong solutions. Existence of weak solutions using a Galerkin approach was obtained in [Ženíšek, 1984] and more regularity on the weak solution was retrieved in [Owczarek, 2010]. One of the most comprehensive work is [Showalter, 2000], in which a semigroup approach allows the author to establish the existence of strong but also weak solutions, in particular in the incompressible case $c_0 = 0$. Concerning Biot poroacoustic equations, see (15), the existence of

solutions was proved in [Santos, 1986] and [Ezziani, 2005] using respectively Galerkin and semigroup approaches. Yet, in these two articles, the incompressible case $c_0 = 0$ is not considered, whereas it is essential for the biomedical applications targeted in this thesis.

In parallel with these theoretical works, the conception of robust discretization techniques for poromechanics problems became an active field of research. The first developments of this field can be found in [Russell and Wheeler, 1983] and [Lewis and Schrefler, 1987]. Among others, more recent studies include finite differences using MAC grids [Harlow and Welch, 1965; Gaspar et al., 2003], discontinuous Galerkin elements [Phillips and Wheeler, 2008; Liu et al., 2009; Chen et al., 2013; Khan and Zanotti, 2020], nonconforming discretizations based on Raviart-Thomas elements [Yi, 2013; Hu et al., 2017; Khan and Zanotti, 2020], finite volumes [Nordbotten, 2014, 2016] and stabilization techniques [Rodrigo et al., 2016, 2018]. Particular attention has been paid to better understand the causes of the so-called poroelastic locking, which occurs when the incompressible or low-permeability regime is reached [Phillips and Wheeler, 2009; Ferronato et al., 2010; Haga et al., 2012; Yi, 2017; Bertrand et al., 2022]. To overcome this locking, [Oyarzúa and Ruiz-Baier, 2016] introduces an additional unknown related to the total stress of the porous medium, while [Lee, 2018] and [Lee et al., 2019] consider an additional unknown named total pressure, which allows to retrieve robust estimates with respect to the locking physical parameters. Moreover, splitting schemes were proposed in [Zienkiewicz et al., 1988; Settari and Mourits, 1998]. These splitting procedures, known respectively as the undrained split and fixed-stress split algorithms, were shown to be convergent in [Mikelić and Wheeler, 2013] and [Girault et al., 2019]. Recent theoretical and numerical optimizations were given in [Both et al., 2017; Storvik et al., 2019; Both et al., 2019a], which has led to renewed interest in such methods. Other splitting schemes based on projection methods were designed in [Zienkiewicz et al., 1993; Huang et al., 2001; Li et al., 2003] for the unsteady Biot's consolidation model, namely (16) with $\rho > 0$, and in [Markert et al., 2009] for Biot incompressible poroacoustics equations, namely (15) with $c_0 = 0$. However, to our knowledge, no convergence or complete stability analysis was shown for these projection methods. The above references focus on Biot-type systems. Yet, under certain physical assumptions, (16) further simplifies to Darcy equation, for which many numerical methods have been proposed. In particular, Darcy equation can be used to simulate fractured porous media, see [Angot et al., 2009; Bukač et al., 2016; Köppel et al., 2018; Van Duijn et al., 2019; Bonaldi et al., 2021] just to name a few.

We also note that the non-linear poromechanics model (2) shows some similiarities with fluid-structure interaction problems, as shown in [Burtschell, 2016, Remark 13]. Indeed, it can be interpreted as a fluid-structure interaction problem in which the fluid and solid parts share the same porous domain and hence interact at each point of the domain. Moreover, as in fluid-structure interaction modeling, Problem (2) can be formulated within an Arbitrary Lagrangian Eulerian framework. Therefore, numerical strategies developed for fluid-structure interaction problems may provide some hints for the system considered here. The numerical analysis of the linearized version of fluid-structure interaction problems in the incompressible case was performed in [Le Tallec and Mani, 2000]. For non-linear fluid-structure interaction problems dealing with viscous incompressible flows, the first numerical approaches consisted in fixed point algorithms [Le Tallec and Mouro, 2001; Mok and Wall, 2001; Deparis et al., 2003], that were improved in [Gerbeau and Vidrascu, 2003] using a quasi-Newton algorithm. In these problems, when the densities of the fluid and of the structure are close, numerical unstabilities arise in loosely coupled strategies from the so-called added-mass effect, which was studied in [Causin et al., 2005] using a toy model together with a spectral analysis. These unstabilities were treated in semi-implicit coupling schemes based on projection schemes [Fernández et al., 2007; Astorino and Grandmont, 2010], other operator splitting approaches [Guidoboni et al., 2009; Bukač et al., 2014], algebraic factorization [Quaini and Quarteroni, 2007; Badia et al., 2008] or Robin method [Astorino et al., 2010]. Then, a fully explicit coupling was proposed in [Burman and Fernández, 2009] using Nitsche's method. The convergence

of this explicit strategy, in particular the dependency of the error with respect to the mesh size, was further improved in [Burman and Fernández, 2014] and in [Burman et al., 2022a] thanks to a Robin-Robin coupling approach. Note moreover that Nitsche's method for fluid-structure interaction problems, originally described in [Hansbo et al., 2004], has also been used in the development of immersed boundary methods [Burman et al., 2015; Massing et al., 2015].

Concerning works specifically relating to the non-linear poromechanics model (2) or to its linearized counterpart (13) studied here, some mathematical and numerical results are already available. For the non-linear model, [Burtschell et al., 2017] proposes a partioned method in which the fluid viscous step is treated explicitely while the fluid and the solid remain coupled implicitly in their respective projection steps. This study considers the case of total stress boundary conditions, which are treated thanks to a Robin method inspired from fluid-structure interaction [Astorino et al., 2010]. In [Chabiniok et al., 2022], a model reduction of (2) is derived assuming that the porous medium presents a spherical symmetry. Moreover, the model is slightly modified in order to take into account the blood perfusion provided by the coronary network, thus paving the way to clinical applications. For the linearized poromechanics model (13) with Dirichlet boundary conditions, [Burtschell et al., 2019] shows the existence and uniqueness of strong solutions for a compressible and viscous skeleton, namely $\kappa < +\infty$ and $\eta > 0$. In [Barnafi et al., 2021], the compressible non-viscous case $\kappa < +\infty$ and $\eta = 0$ is investigated. The existence of weak solutions is obtained thanks to a Galerkin method provided that the permeability k_f is large enough and that the additional fluid mass input θ is small enough. Yet, the existence of solutions for an incompressible material $\kappa = +\infty$ is not addressed. From a numerical point of view, [Burtschell et al., 2019] proposes a Newmark discretization for the solid together with a Backward Euler method for the fluid, while [Barnafi et al., 2021] considers a Backward Euler method both for the solid and fluid parts. Spatial and temporal convergence analysis are established assuming that a discrete inf-sup condition associated with the incompressibility constraint (14) is satisfied. However, note that the inf-sup condition obtained in these works depends on the porosity ϕ, so that it could not be generalized to the non-linear case in which the porosity is an unknown of the system. Finally, in [Both et al., 2022], the authors extend the undrained and fixed-stress algorithms for Biot's equations to the porous model (13). This results in an alternating minimization algorithm iterating between the solid and the fluid steps until convergence. The number of iterations before convergence is very sensitive to the various physical parameters and in particular to permeability, which is a limitation of this scheme in the low-permeable regime.

In this context, the major contributions of this thesis are detailed in what follows.

Main contributions

This thesis is divided into five chapters, the contents of which are summarized below.

Chapter 1 – Analysis of a linearized poromechanics model for incompressible and nearly incompressible materials In Chapter 1, we prove the existence and uniqueness of weak and strong solutions to Problem (13) in all cases $\eta \geq 0$ and $\kappa \leq +\infty$. In the compressible regime $\kappa < +\infty$, we extend the previous existence results by dropping the conditions made on the permeability and the additional fluid mass input, while the study of the incompressible case $\kappa = +\infty$ requires a specific analysis of the incompressibility constraint (14). The proofs hinges on a combination of semigroup theory and energy estimates together with $\mathtt{T}$-coercivity. $\mathtt{T}$-coercivity is a notion generalizing coercivity that enables us to handle all the difficulties of the problem in a compact way, and which is equivalent to the inf-sup conditions. First, it allows us to deal with the hyperbolic – parabolic coupling of the solid and fluid equations that occurs for a non-viscous solid. Indeed when $\eta = 0$ the solid equation (13a) becomes hyperbolic, while the fluid equation (13b) is always parabolic. Second, we use this notion to deal with the incompressible case $\kappa = +\infty$ in which the

solid and fluid phases are coupled through the divergence constraint (14) by retrieving an inf-sup condition indepedent of porosity. Moreover, we show how to pass to the limit on the weak solutions between the compressible and incompressible regimes.

Chapter 2 – The T-coercivity method for mixed problems (joint work with Patrick Ciarlet, ENSTA Paris) The notion of T-coercivity was originally introduced for unconstrained static problems [Bonnet-Ben Dhia et al., 2010a; Ciarlet Jr, 2012; Chesnel and Ciarlet, 2013; Bonnet-Ben Dhia et al., 2014]. Chapter 2 is devoted to the extension of this concept for saddle-point problems, that appears in particular in system (13) when $\kappa = +\infty$. We start by applying T-coercivity to Stokes problem, and then deduce a general framework from this example. In particular, we prove the equivalence between the T-coercivity theory and the usual Ladyzhenskaya–Babuška–Brezzi inf-sup condition. Furthermore, it is shown that T-coercivity simplifies some proofs of the standard theory for perturbed saddle-point problems. The T-coercivity method appears to be a flexible tool that can be applied to various physical problems, including electromagnetism, nearly-incompressible elasticity and neutron diffusion. In particular, the method allows to go from the continuous to the discrete level in a simple way by invoking a Fortin operator.

Chapter 3 – Numerical analysis of an incompressible soft material poromechanics model using T-coercivity This third chapter deals with the numerical approximation of Problem (13) in the incompressible regime using mixed finite elements. Two monolithic schemes are considered with either a Newmark discretization for solid and fluid quantities, or a Backward Euler discretization for the fluid. Thanks to the T-coercivity mappings built in Chapter 1, we define a discrete projection operator adapted to the bilinear form involved in the problem. Using this projection operator together with discrete energy balances, a spatial and temporal error analysis for both schemes is performed as long as the standard divergence inf-sup condition is satisfied. Doing so, we obtain a discretization that is robust with respect to incompressibility, porosity and permeability. Moreover, we study both theoretically and numerically the influence of the additional fluid mass input on the stability and convergence of the schemes.

Chapter 4 – A projection scheme for an incompressible soft material poromechanics model After the monolithic approach followed in Chapter 3, this chapter aims at proposing a splitting scheme decoupling the computation of solid, fluid and pressure quantities at each time step regardless of boundary conditions. By taking into account the specific saddle-point structure of the problem, a projection scheme is designed for Dirichlet, Neumann and total stress boundary conditions. For this last case, we use a Robin-Robin coupling method inspired from fluid-structure interaction problems [Burman et al., 2022a] in order to ensure stability with respect to possible added-mass effects. The stability analysis is carried out paying attention on the explicit or implicit treatment of permeability. In the case of Dirichlet boundary conditions, we prove the convergence of the scheme for both its non-incremental and incremental versions. Numerical results are provided.

Chapter 5 – Modeling and simulation of artificial microvessel perfusion (joint work with Claire A. Dessalles, University of Geneva) The last chapter of this thesis is turned towards biomedical applications. More precisely, the goal of Chapter 5 is to use the poromechanics model (13) to simulate microvessel-on-chip platforms. The microvessel-on-chip platform under study is a microfluidics experiment in which a porous hydrogel is perforated by a cylindrical water channel symbolizing a microvessel. Using the monolithic scheme analyzed in Chapter 3 with suitable boundary conditions, we show that our model is able to reproduce such experiments. The simulation results are first validated qualitatively by a realistic test case and then more quantitatively by a parametric study.

The key concepts involved in this thesis and the links between its chapters are represented schematically in Figure 5.

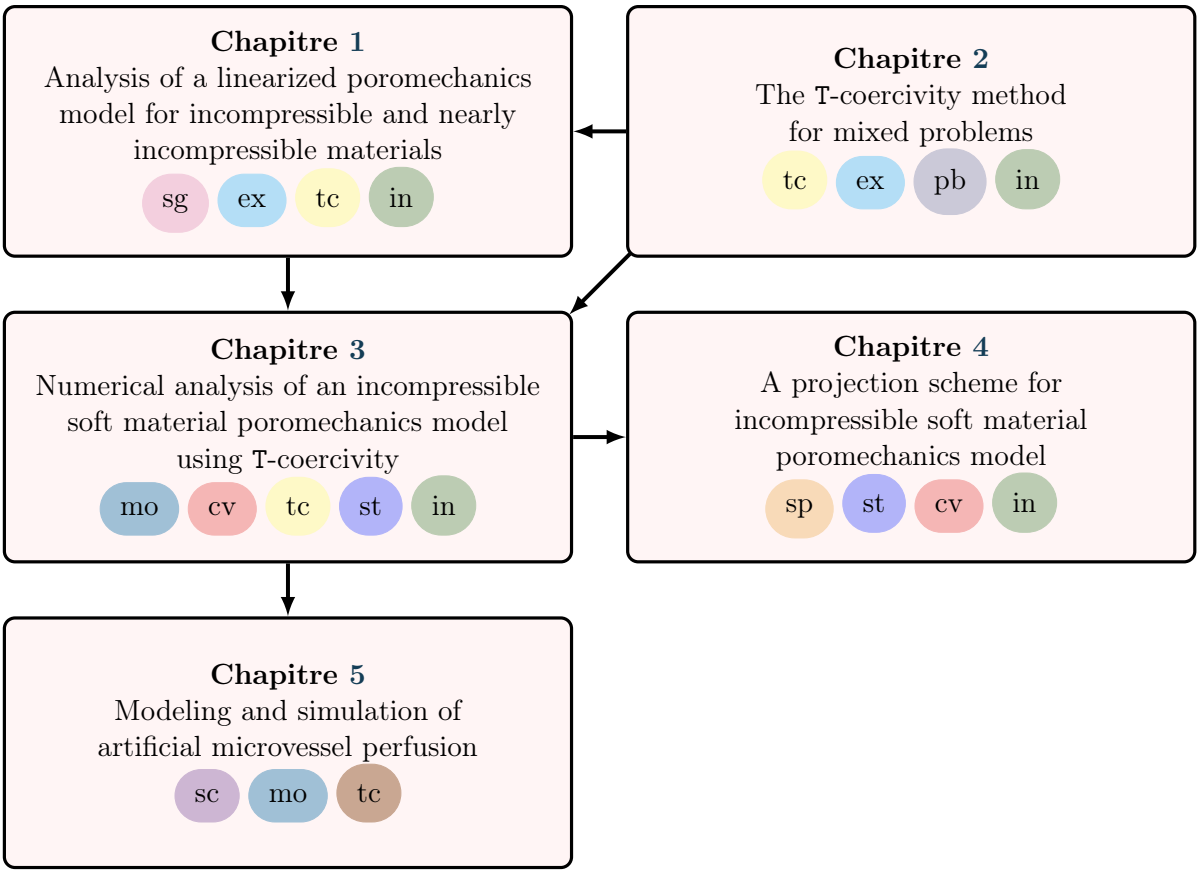


Figure 5 – Thesis organization and key concepts. sg semigroup theory, ex existence and uniqueness of PDEs solutions, tc T-coercivity, in incompressibility and inf-sup conditions, pb perturbed saddle-point problems, mo monolithic scheme, cv spatial and temporal convergence analysis, st stability analysis, sp splitting scheme, sc scientific computing, tc transmission conditions.

Note that the first three chapters of this thesis have been published or submitted.

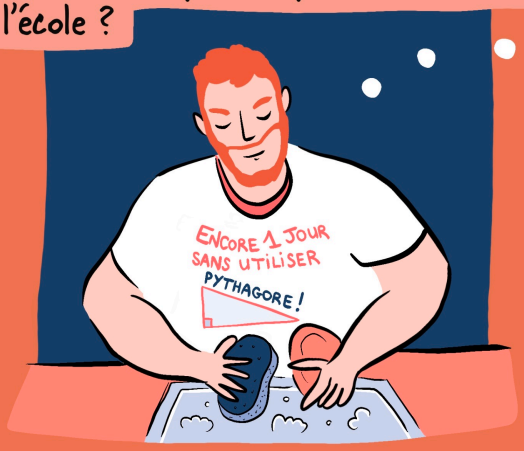
- Barré, M., Grandmont, C. and Moireau, P. (2023) Analysis of a linearized poromechanics model for incompressible and nearly incompressible materials. *Evolution Equations and Control Theory*, 12(3):846-906, DOI:10.3934/eect.2022053.
- Barré, M., Grandmont, C. and Moireau, P. (2023) Numerical analysis of an incompressible soft material poromechanics model using T-coercivity. *Comptes Rendus. Mécanique*, 351(S1), 1-36, DOI:10.5802/crmeca.194.
- Barré, M. and Ciarlet Jr, P. (Submitted) The T-coercivity approach for mixed problems.

Moreover, each of the numerical schemes proposed in Chapters 1, 3, 4 and 5 have been implemented using the FEniCS finite element software [Logg et al., 2012; Alnæs et al., 2015]. Altogether, these implementations represent about 10 000 lines of code.

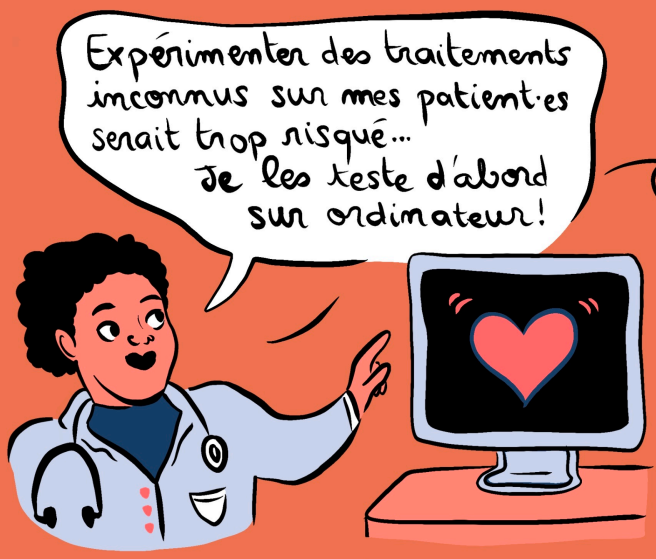
To conclude this introduction, we reproduce a comic strip created with the illustrator and scriptwriter Marine Spaak. This comics strip was part of a scientific outreach project for Inria and Université Paris-Saclay. It aims at explaining with simple words the subject and motivations of this thesis.

LES MATHS FONT BATTRE MON CŒUR (1/2)

Vous pensez que les maths ne servent plus à rien une fois qu'on a quitté les bancs de l'école ?



Et si je vous disais que les mathématiques peuvent aider les médecins à sauver des vies ?



Pour faire des simulations informatiques d'organes, il faut modéliser leur comportement par des équations

Plusieurs équations peuvent décrire le même organe selon le phénomène étudié

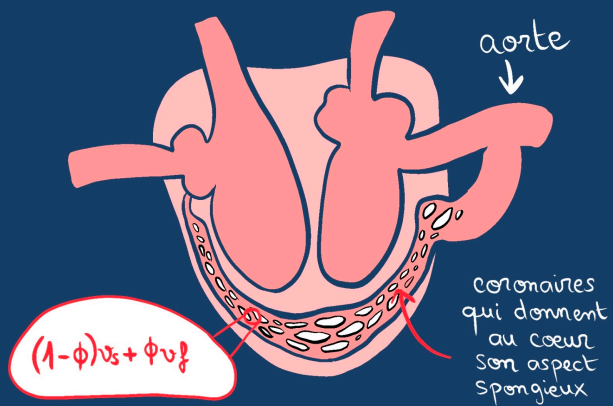
Thèse conduite et adaptée en BD par : Mathieu BARRÉ | co-scénarisation et illustrations : Marine SPAAK

Avec l'accompagnement de : université PARIS-SACLAY | LA DIAGONALE

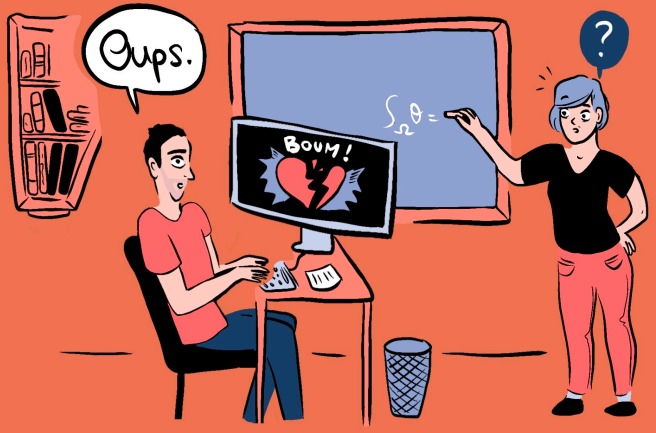
LES MATHS FONT BATTRE MON CŒUR (2/2)

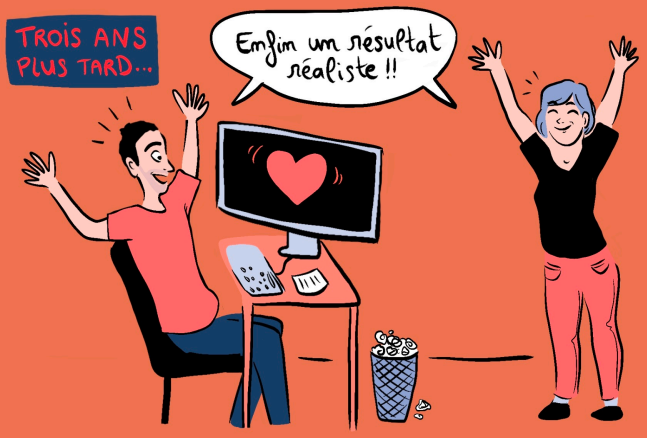
Dans ma thèse, j'étudie les équations des éponges biologiques en mouvement

Ces équations s'appliquent en particulier au cœur humain

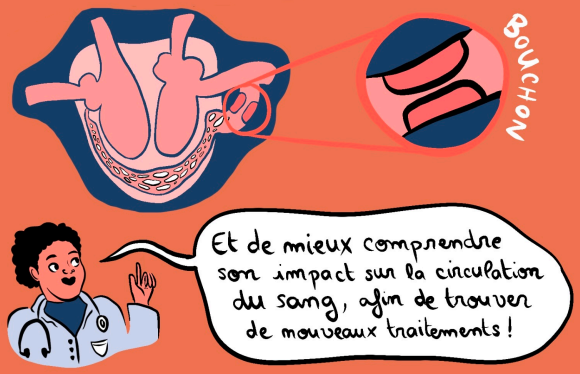


Sans mathématiciennes et mathématiciens pour comprendre leurs secrets, impossible d'avoir une simulation sur ordinateur qui fonctionne !





Mes travaux pourraient un jour permettre de simuler numériquement un bouchon dans les coronaires



Thèse conduite et adaptée en BD par : Mathieu BARRÉ | co-scénarisation et illustrations : Marine SPAAK

Avec l'accompagnement de : université PARIS-SACLAY | LA DIAGONALE

CHAPTER 1

Analysis of a linearized poromechanics model for incompressible and nearly incompressible materials

This chapter reproduces results published in *Evolution Equations and Control Theory*, 12(3):846-906 (2023) and obtained in collaboration with Céline Grandmont and Philippe Moireau. Moreover, in June 2021, I presented this work at the *International Conference on Coupled Problems in Science and Engineering (ECCOMAS)* in Chia Laguna, Italy (online).

Contents

Analysis of a linearized poromechanics model for incompressible and nearly incompressible materials

Mathieu Barré[1,2], Céline Grandmont[3,4,5] and Philippe Moireau[1,2]

[1] Inria, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
[2] LMS, École Polytechnique, CNRS, Institut Polytechnique de Paris Route de Saclay, 91120 Palaiseau, France
[3] Inria, 2 Rue Simone Iff, 75012 Paris, France
[4] LJLL, Sorbonne Université, CNRS, 4 Place Jussieu, 75005 Paris, France
[5] Département de Mathématique, Université Libre de Bruxelles CP 214, Boulevard du Triomphe, 1050 Bruxelles, Belgium

Published in *Evolution Equations and Control Theory*
DOI:10.3934/eect.2022053

Abstract

In this work, we thoroughly analyze the linearized version of a poromechanics model developed to simulate soft tissues perfusion. This is a fully unsteady model in which the fluid and solid equations are strongly coupled through the interstitial pressure. As such, it generalizes Darcy, Brinkman and Biot equations of poroelasticity. The mathematical and numerical analysis of this model was initially performed for a compressible porous material. Here, we focus on the nearly incompressible case with a semigroup approach, which also allows us to prove the existence of weak solutions. We show the existence and uniqueness of strong and weak solutions in the incompressible limit case, for which a divergence constraint on the mixture velocity appears. Due to the special form of the coupling, the underlying problem is not coercive. Nevertheless, by using the notion of $\mathtt{T}$-coercivity, we obtain stability estimates and well-posedness results. Our study also provides guidelines to propose stable and robust approximations of the problem with mixed finite elements. In particular, we recover an inf-sup condition that is independent of the porosity. Finally, we numerically investigate the elliptic regularity of the associated steady-state problem and illustrate the sensitivity of the solution with respect to the different model parameters.

Keywords — Mixture theory, incompressible limit, weak and strong solutions, semigroup theory, $\mathtt{T}$-coercivity.
Mathematics Subject Classification (2020) — 35M31, 35A15, 47D06, 74F10.

Introduction

Poromechanical models aim at describing the mechanical behavior of saturated porous media with the interaction of a fluid flow within a deformable porous structure through the definition of a multi-phase continuum framework [Coussy, 2004; De Boer, 2005]. The initial introduction of such models concerns geophysics [Biot, 1941; Terzaghi et al., 1996], but these models have been recently used for biomechanical applications, in particular to represent perfused living tissues. If the heart perfusion remains a leading example of application [Yang and Taber, 1991; Huyghe et al., 1992; Nash and Hunter, 2000; Chapelle et al., 2010; Michler et al., 2013; Chabiniok et al., 2016], poroelastic models

have also been considered to simulate lipid and drug transport in blood vessel walls [Koshiba et al., 2007; Calo et al., 2008; Badia et al., 2009; D'Angelo and Zunino, 2011; Čanić et al., 2021], water transport and drug delivery in the brain [Basser, 1992; Støverud et al., 2011; Tully and Ventikos, 2011; Vardakis et al., 2016; Chou et al., 2016; Guo et al., 2018; Lee et al., 2019], ocular diseases such as glaucoma [Causin et al., 2014; Ruiz-Baier et al., 2022], fibrosis diagnosis in the lungs [Berger et al., 2016; Genet et al., 2020], or also tissue growth [Ambrosi and Preziosi, 2002; Sacco et al., 2017; Deville et al., 2018].

In these biomedical applications, physical phenomena such as fluid inertia and solid quasi-incompressibility, generally neglected in civil engineering, may play an important role. Therefore, the original poroelasticity model derived by Terzaghi [Terzaghi, 1943] and Biot [Biot, 1941] must be revised to include inertial effects. Note that in the many applications of poroelasticity, unsteady behavior for the fluid and the solid is typically included when studying wave propagation in porous media, see [Schanz, 2009] and references therein. It is also an important topic for the simulation of fluid-porous structure interaction (FPSI) occurring in living tissues [Showalter, 2005; Badia et al., 2009; Bukač et al., 2015b,a, 2016; Cesmelioglu, 2017; Angot, 2018; Ambartsumyan et al., 2018; Ager et al., 2019; Čanić et al., 2021; Bociu et al., 2021]. In addition to inertial effects, perfused organs such as the heart or the lungs are subject to finite strains, so their modeling must also account for these non-linear effects and, in particular, consider porosity – which represents the fraction of fluid in the porous material – as a primary variable. Such modeling extensions were proposed within the framework of Biot theory where the solid skeleton plays a special role [Bourgeois, 1997; Lopatnikov and Cheng, 2004; Gil et al., 2022], or in the context of mixture theory treating equivalently all components of the mixture [Bowen, 1980; Wilmanski, 2005; Rajagopal and Tao, 2005]. All these models suppose – explicitly or implicitly – that the frictional effects within the fluid can be neglected due to its viscosity, and rarely take into consideration the influence of solid viscosity. Recently, authors in [Chapelle and Moireau, 2014] have revisited the framework of Biot theory at finite strain to derive general formulations adapted to soft tissues perfusion, including inertial and viscous effects both for the fluid and the solid.

Their formulation is compatible with thermodynamical principles. In particular, the solution of the linearized version of the fully coupled model proposed in [Chapelle and Moireau, 2014] satisfies energy estimates, opening the way to prove well-posedness. In [Burtschell et al., 2019; Barnafi et al., 2021] the case where the structure is compressible is considered for a linearized system close to the one considered here. Still, the general resulting formulation can exhibit – when solid viscosity is neglected – a hyperbolic-parabolic coupling between the structure and the fluid, with – when the skeleton is incompressible – an additional incompressibility constraint involving a mixture velocity, and therefore leads to challenging questions of analysis.

From a mathematical point of view, there is a large literature related to the existence and uniqueness of solutions for linear Biot's consolidation models, namely systems of the form

$$\begin{cases} \rho\,\partial_{tt}^2 u_s - (\lambda+\mu)\nabla(\operatorname{div} u_s) - \mu\,\Delta u_s + \alpha\nabla p = f, & \text{(1.1a)} \\ \partial_t(c_0 p + \alpha \operatorname{div} u_s) - \operatorname{div}(k_f \nabla p) = g, & \text{(1.1b)} \end{cases}$$

where the two unknowns are the displacement of the structure u_s and the interstitial pressure p, which corresponds to the fluid pressure in the pores. For the unsteady system ($\rho > 0$), the existence of strong solutions was first derived in [Dafermos, 1968] using Laplace transform and then completed by [Fichera, 1974], and the existence of weak solutions was obtained in [Barucq et al., 2004] with a Galerkin method and a regularization technique. The quasi-static case ($\rho = 0$) was first studied in [Auriault, 1980] where it was recovered using homogenization techniques, leading to the existence of strong solutions. Existence of weak solutions was shown in [Ženíšek, 1984] using a Galerkin approach, which was recently refined to get a more regular solution [Owczarek, 2010]. In [Showalter, 2000], existence of strong but also weak solutions is established by means of a semigroup approach. This article also handles secondary consolidation phenomena occuring in clays [Murad

and Cushman, 1996], modeled by the presence of an extra term $-\nabla\big(\lambda^*\partial_t(\operatorname{div} u_s)\big)$ in (1.1a). Non-linear extensions of (1.1) were also analyzed [Showalter and Stefanelli, 2004; Showalter, 2013; Cao et al., 2013; Bociu et al., 2016; Both et al., 2021; Bociu and Webster, 2021; Bociu et al., 2022]. Yet, in the previous models, fluid inertial effects are neglected and, apart from [Showalter, 2000; Bociu et al., 2022], little attention is paid to the incompressible case $c_0 = 0$. Moreover, fluid inertial effects are included in porous wave propagation models [Biot, 1956a], whose well-posedness was studied in [Santos, 1986] and [Ezziani, 2005] using respectively Galerkin and semigroup approaches. However, the fluid viscosity is still not considered and the existence of solutions is carried out only for a compressible fluid, while the fluids present in biomedical applications (blood, lymph, cerebrospinal fluid) are mostly incompressible. Finally, [Burtschell et al., 2019] and [Barnafi et al., 2021] take into account inertial and viscous fluid effects as their formulation are derived from the linearization of [Chapelle and Moireau, 2014] and show respectively the existence of a strong solution when solid viscosity is included, and the existence of a weak solution in absence of solid viscosity, both for a compressible solid. The existence result for incompressible or nearly-incompressible materials was not covered by their results.

In the present work, we study the well-posedness for a linearized system, obtained by linearizing the fully coupled system introduced in [Chapelle and Moireau, 2014], by unifying semigroup and variational approaches. The considered model takes into account both fluid and structure inertia, the fluid viscosity, possible damping in the structure, a friction force between both phases, and the interstitial pressure. The elastic or viscoelastic skeleton can be compressible or incompressible, so that we consider four different cases. Our results include the compressible fully viscous case originally studied in [Burtschell et al., 2019] and generalize, by relaxing the condition on the fluid mass source term, the results on the compressible elastic case obtained in [Barnafi et al., 2021]. Note moreover that the linearized system here considered differs slightly from the one studied in [Burtschell et al., 2019; Barnafi et al., 2021], since it incorporates the Biot-Willis coefficient that models pressure-deformation coupling, hence relating the proposed model to the forementioned Biot-type systems. In addition to the compressible case, we fully analyze the incompressible limit case, which corresponds to the physiological regime when considering living tissues. Our approach exploits the notion of $\mathtt{T}$-coercivity [Ciarlet Jr, 2012; Chesnel and Ciarlet, 2013] to prove, when no damping is added to the structure, the surjectivity of the underlying operator that involves the resolution of a non-coercive problem. Furthermore, we also take advantage of the recent parallel between inf-sup conditions and $\mathtt{T}$-coercivity [Barré and Ciarlet Jr, 2022] to prove the fundamental inf-sup condition associated with the mixture velocity constraint that we have to deal with in the incompressible case. It appears that the inf-sup condition is ultimately independent of the porosity. This result, already conjectured in [Burtschell et al., 2019] and partially justified in [Barnafi et al., 2021], is crucial to be able to use generic finite-element discretization. It would also be essential when considering the discretization of the non-linear model from [Chapelle and Moireau, 2014], in which the porosity is an unknown of the system.

The paper is organized as follows. Section 1.1 presents the poromechanics model under study, its connection with standard Biot models, and general preliminaries such as the formal derivation of energy estimates on the system. Further details concerning the full non-linear model introduced in [Chapelle and Moireau, 2014] and its linearization can be found in the thesis introduction. In Section 1.2, we unify the semigroup and variational approaches used in [Burtschell et al., 2019] and [Barnafi et al., 2021] by proving the existence and uniqueness of strong and weak solution for a compressible porous material. We highlight the role of solid viscosity on the model by pointing out the differences that appear in the variational formulation when this coefficient vanishes. Section 1.3 is devoted to the incompressible regime and more specifically to the saddle-point structure of the problem arising in this case, with a particular attention dedicated to the existence and regularity of pressure. Next, in Section 1.4, we establish a link between the results of Sections 1.2 and 1.3 by passing to the incompressible limit obtained when the bulk modulus of the structure skeleton

goes to infinity. Finally, these theoretical results are complemented with numerical experiments exploring the regularity of solutions and the domain of the underlying semigroup operator.

1.1 Problem setting

The model, close to the one we consider here, was introduced in [Burtschell et al., 2019] and further explored in [Barnafi et al., 2021; Both et al., 2022]. This model comes from the linearization of the poromechanical model developed in [Chapelle and Moireau, 2014]. For the sake of completeness, we refer the reader to the thesis introduction for a brief presentation of the non-linear model proposed in [Chapelle and Moireau, 2014] and details about the linearization process in which we introduce the Biot-Willis coefficient that was not taken into account in [Burtschell et al., 2019; Barnafi et al., 2021; Both et al., 2022]. We will explain in Section 1.1.1 that the resulting linearized model is a variant of the well-known Biot systems [Biot, 1941, 1955; Biot and Temple, 1972]. Its peculiarity compared to Biot-type models is that it incorporates inertial and viscous effects for both the fluid and the solid, and satisfies an energy balance which is formally derived in Section 1.1.2 and further rigorously justified.

We consider a porous medium in a bounded domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ with Lipschitz boundary. In each point of the domain Ω, we consider a mixture of fluid and structure and we denote by ϕ the porosity. For all $x \in \Omega$, $\phi(x) \in (0, 1)$ represents the fraction of fluid whereas $1 - \phi(x)$ represents the fraction of elastic medium. The fluid phase is assumed to be an homogeneous, viscous, Newtonian and incompressible fluid. We denote by v_f its velocity, ρ_f its density and μ_f its viscosity. Since the fluid is incompressible (resp. homogeneous), ρ_f is independent of time (resp. of space). We also assume that the structure is elastic and, to simplify, that its macroscopic behavior law is linear and isotropic and, thus, characterized by two Lamé constants λ and μ. They stand for the elastic parameters characterizing the macroscopic behavior of the elastic part of the mixture (*i.e.* the homogenized behavior of a perforated elastic medium with no fluid). We denote by u_s the structure displacement and by $v_s = \partial_t u_s$ the structure velocity. The density of the structure is denoted by ρ_s and its viscosity by η. The fluid and the structure are coupled through a friction force that depends linearly on the relative velocity $v_f - v_s$ and reads $\phi^2 k_f^{-1}(v_f - v_s)$, where k_f is the hydraulic conductivity tensor. In addition, they are coupled through the interstitial pressure p, that is further linked to the incompressibility of the whole fluid-structure mixture. Finally, $\alpha(x) \in (\phi(x), 1)$ is the Biot-Willis coefficient, which takes into account the pressure-deformation coupling. This coefficient depends on space for a compressible material but tends to 1 in the incompressible limit as the skeleton elastic bulk modulus, denoted κ, tends to $+\infty$, see the thesis introduction.

The fully coupled model then reads

$$\begin{cases} \rho_s(1-\phi)\,\partial_{tt}^2 u_s - \operatorname{div}\left(\sigma_s(u_s)\right) - \operatorname{div}\left(\sigma_s^{\mathrm{vis}}(\partial_t u_s)\right) \\ \qquad - \phi^2\, k_f^{-1}(v_f - \partial_t u_s) + (\alpha - \phi)\,\nabla p = \rho_s(1-\phi)\, f, & \text{in } \Omega \times (0,T), \quad (1.2a) \\ \rho_f \phi\, \partial_t v_f - \operatorname{div}\left(\phi\, \sigma_f(v_f)\right) \\ \qquad + \phi^2\, k_f^{-1}(v_f - \partial_t u_s) - \theta\, v_f + \phi\, \nabla p = \rho_f \phi\, f, & \text{in } \Omega \times (0,T), \quad (1.2b) \\ \dfrac{\alpha - \phi}{\kappa}\, \partial_t p + \operatorname{div}\left((\alpha - \phi)\, \partial_t u_s + \phi\, v_f\right) = \dfrac{\theta}{\rho_f}, & \text{in } \Omega \times (0,T), \quad (1.2c) \end{cases}$$

where the structure stress tensor is given by Hooke's law

$$\sigma_s(u) = \lambda \operatorname{Tr}(\varepsilon(u))\, \mathcal{I} + 2\mu\, \varepsilon(u),$$

with $\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^T)$, the structure additional viscosity is given by

$$\sigma_s^{\mathrm{vis}}(v) = 2\eta\, \varepsilon(v),$$

and the fluid stress tensor reads

$$\sigma_f(v) = \lambda_f \operatorname{Tr}(\varepsilon(v))\, \mathcal{I} + 2\mu_f\, \varepsilon(v).$$

In the above system, the data are the applied exterior force f and the additional fluid mass input θ. The coupled system (1.2) describes the mixture of an elastic, possibly viscous medium and an incompressible Newtonian flow. The first equation (1.2a) represents the momentum conservation law of the elastic phase including inertial effects, macroscopic elastic behavior, possible viscous damping, friction force between the fluid and the structure, and the gradient of the interstitial pressure. The second equation (1.2b) stands for the momentum conservation law of the fluid phase including inertial effects, macroscopic viscous effects, friction force and the gradient of the interstitial pressure. The third equation (1.2c) traduces the total mass conservation dynamic, it involves the parameter κ, that represents the bulk modulus of the elastic medium constituting the porous matrix, and the Biot-Willis parameter α. When $\kappa < +\infty$ it corresponds to a compressible skeleton, whereas when $\kappa = +\infty$ (that implies $\alpha = 1$) we have an incompressible elastic skeleton. Since the fluid is assumed to be incompressible we deal in the limit case $\kappa = +\infty$ with an incompressible porous medium. This latter case is crucial when considering living tissues since they are nearly incompressible. Note that when $\kappa = +\infty$ there is no dynamic of the pressure since the term $\partial_t p$ in (1.2c) vanishes, but the pressure is the Lagrange multiplier associated with the mixture constraint $\rho_f \operatorname{div}\big((1-\phi)\,\partial_t u_s + \phi\, v_f\big) = \theta$ involving the mixture velocity $v_m = (1-\phi)\,\partial_t u_s + \phi\, v_f$. Further details on the derivation of this linearized coupled system are gathered in the thesis introduction.

Remark 1.1. Note that the skeleton incompressibility is equivalent to $\kappa \to +\infty$, but does not necessarily imply that $\lambda \to +\infty$. As a matter of fact, λ is the *drained* Lamé coefficient, namely the mechanical parameter of a skeleton perforated by holes corresponding to the fluid phase in the porous medium. In particular, in the targeted biomedical applications, the porous medium is mostly composed of the fluid phase, so that λ may remain small even if the solid phase is incompressible.

For a presentation of typical boundary conditions for such systems, we refer to [Burtschell et al., 2017, 2019; Sacco et al., 2019; Čanić et al., 2021] and their analysis will imply further development. In the present work, we limit our analysis to the case of homogeneous Dirichlet boundary conditions for the structure and for the fluid:

$$\begin{cases} u_s = 0, & \text{on } \partial\Omega \times (0,T), \quad (1.3a) \\ v_f = 0, & \text{on } \partial\Omega \times (0,T). \quad (1.3b) \end{cases}$$

This coupled problem has to be completed with initial data:

$$\begin{cases} u_s(0) = u_{s0}, & \text{in } \Omega, \quad (1.4a) \\ \partial_t u_s(0) = v_{s0}, & \text{in } \Omega, \quad (1.4b) \\ v_f(0) = v_{f_0}, & \text{in } \Omega, \quad (1.4c) \end{cases}$$

and in the case $\kappa < +\infty$

$$p(0) = p_0, \quad \text{in } \Omega. \quad (1.5)$$

Before detailing the well-posedness analysis of the considered coupled system, let us first emphasize its links to other systems modeling porous media.

1.1.1 Related poromechanics models

As shown in [Rajagopal, 2007], Darcy, Brinkman and Biot equations can be derived within the framework of mixture theory under specific assumptions. The system (1.2), which arises from Biot theory, can be seen as a combination of these models. Indeed, (1.2) is close to the fully dynamic

Biot system introduced in [Biot, 1956a] for the study of acoustic waves in saturated porous media, but also includes a viscous fluid term as in Brinkman equation.

More precisely, denoting by u_f the displacement of fluid particles within the porous medium and by $w = \phi(u_f - u_s)$ the relative displacement of the fluid phase with respect to the solid one, the model from [Biot, 1956a] reads

$$\begin{cases} \rho\,\partial_{tt}^2 u_s + \rho_f\,\partial_{tt}^2 w - \operatorname{div}\big(\sigma_s(u_s)\big) + \alpha\nabla p = g, & \text{(1.6a)} \\ \rho_f\,\partial_{tt}^2 u_s + a\rho_f\,\partial_{tt}^2\left(\dfrac{w}{\phi}\right) + k_f^{-1} q + \nabla p = h, & \text{(1.6b)} \\ c_0 p + \alpha \operatorname{div} u_s + \operatorname{div} w = k, & \text{(1.6c)} \end{cases}$$

where $\rho = \rho_s(1-\phi)+\rho_f\phi$ corresponds to the density of the mixture, $a \geq 1$ is a coefficient describing tortuosity effects, and

$$c_0 = \frac{\phi}{\kappa_f} + \frac{\alpha - \phi}{\kappa},$$

is the storage coefficient, with κ_f the fluid bulk modulus.

In our case, the fluid is assumed to be incompressible and thus $\kappa_f = +\infty$, so that $c_0 = \frac{\alpha-\phi}{\kappa}$. To link (1.2) and (1.6), let us assume that we have no additional fluid mass input, namely $\theta = 0$, and that we can *neglect viscous effects*, which amounts to take $\eta = \mu_f = \lambda_f = 0$. Introducing the new unknown

$$q = \phi(v_f - \partial_t u_s) = \partial_t w,$$

which corresponds to the filtration velocity, (1.2) becomes

$$\begin{cases} \rho_s(1-\phi)\,\partial_{tt}^2 u_s - \operatorname{div}\big(\sigma_s(u_s)\big) - \phi\, k_f^{-1} q + (\alpha - \phi)\,\nabla p = \rho_s(1-\phi)\, f, & \text{(1.7a)} \\ \rho_f\phi\,\partial_{tt}^2 u_s + \rho_f \partial_t q + \phi\, k_f^{-1} q + \phi\,\nabla p = \rho_f \phi\, f, & \text{(1.7b)} \\ c_0\,\partial_t p + \operatorname{div}\big(\alpha\,\partial_t u_s + q\big) = \rho_f^{-1}\theta. & \text{(1.7c)} \end{cases}$$

Replacing (1.7a) by (1.7a) + (1.7b) and dividing (1.7b) by ϕ, we get

$$\begin{cases} \rho\,\partial_{tt}^2 u_s + \rho_f \partial_t q - \operatorname{div}\big(\sigma_s(u_s)\big) + \alpha\,\nabla p = \rho f, & \text{(1.8a)} \\ \rho_f\,\partial_{tt}^2 u_s + \rho_f\,\partial_t\left(\dfrac{q}{\phi}\right) + k_f^{-1} q + \nabla p = \rho_f\, f, & \text{(1.8b)} \\ \partial_t(c_0 p + \alpha \operatorname{div} u_s) + \operatorname{div} q = \rho_f^{-1}\theta, & \text{(1.8c)} \end{cases}$$

which, provided that $a = 1$, corresponds exactly to (1.6) since $q = \partial_t w$ and (1.8c) $= \partial_t$(1.6c). Note that if $c_0 > 0$, equation (1.6c) can be used to eliminate the pressure unknown as done in [Zienkiewicz and Shiomi, 1984; Santos, 1986; Ezziani, 2005], but it is no longer the case if we consider (1.8c). The assumption $a = 1$ indicates that (1.2) does not take into account tortuosity effects since they are not compatible with the first principle of continuum mechanics introduced in [Chapelle and Moireau, 2014], see [Gil et al., 2022, Section 5.3.4] for a discussion on the thermodynamical compatibility of these effects and [Lopatnikov and Cheng, 2004] for a fully unsteady poromechanical model in which they are included.

If the fluid and solid inertial effects are also neglected, (1.8a) reduces to (1.1a) and (1.8b) implies that $q = -k_f \nabla p + \rho_f k_f f$. Substituting this result in (1.8c), we recover the quasi-static Biot's consolidation model, namely (1.1) with $\rho = 0$. Therefore, the model studied in this paper is connected to Darcy, Brinkman and Biot equations, but the presence of inertial and viscous terms both for the fluid and the solid requires a separate study. In particular, because of these extra terms, the functional setting adapted to the problem differs from the one developped for Biot models. This functional setting is guided by the energy balance presented below.

Remark 1.2. Darcy, Brinkman and Biot models have been justified *a posteriori* using homogeneization techniques, see for instance [Auriault, 1997; Hornung et al., 1997; Mikelić, 2000; Rohan et al., 2019] and references therein. The justification of (1.2) by homogeneization is an open problem.

Remark 1.3. For a physical discussion about when viscous effects can be neglected or not, we refer the reader to [Rajagopal, 2007] or [Markert, 2007, Section 3].

1.1.2 Energy estimate

Existence of solutions of such a coupled system, in the compressible case $\kappa < +\infty$, has been partially obtained in [Burtschell et al., 2019; Barnafi et al., 2021]. More precisely, the case $\kappa < +\infty$, $\eta > 0$ and $\theta = 0$ has already been studied in [Burtschell et al., 2019], where existence of strong solutions thanks to the semigroup formalism has been derived. The case $\kappa < +\infty$, $\eta = 0$ is treated in [Barnafi et al., 2021], where existence of variational solutions is obtained under a smallness assumption on θ. Here, we consider all the different cases $\kappa \leq +\infty$, $\eta \geq 0$, and any given θ sufficiently smooth. We prove existence of unique strong and mild solutions – in a sense to be made precise later – using semigroup theory, from which we deduce existence of a unique variational solution. We eventually show that one can pass to the limit in the weak formulation as κ goes to infinity.

Before going through the proofs, let us first derive formally some energy bounds satisfied by any smooth enough solutions of the coupled problem. We first derive them in the case $\kappa < +\infty$ and then in the limit case $\kappa = +\infty$. Let us multiply (1.2a) by the structure velocity $\partial_t u_s$, integrate over Ω and integrate by parts in space. No boundary terms appear thanks to the homogeneous Dirichlet boundary conditions (1.3a) and we obtain

$$\frac{\rho_s}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega (1-\phi)|\partial_t u_s|^2\,\mathrm{d}x + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x + 2\eta\int_\Omega \varepsilon(\partial_t u_s) : \varepsilon(\partial_t u_s)\,\mathrm{d}x$$
$$- \int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot \partial_t u_s\,\mathrm{d}x - \int_\Omega p\,\mathrm{div}\left((\alpha-\phi)\,\partial_t u_s\right)\mathrm{d}x = \int_\Omega \rho_s(1-\phi) f\cdot \partial_t u_s\,\mathrm{d}x.$$

Let us also multiply (1.2b) by the fluid velocity v_f, integrate over Ω and integrate by parts in space. No boundary terms appear thanks to the homogeneous Dirichlet boundary conditions (1.3b) and we get

$$\frac{\rho_f}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi|v_f|^2\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x$$
$$+ \int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot v_f\,\mathrm{d}x - \int_\Omega \theta|v_f|^2\,\mathrm{d}x - \int_\Omega p\,\mathrm{div}\,(\phi\, v_f)\,\mathrm{d}x = \int_\Omega \rho_f \phi f\cdot v_f\,\mathrm{d}x.$$

The last equation (1.2c) is multiplied by p and integrated over Ω, which leads to

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \frac{\alpha-\phi}{\kappa}|p|^2\,\mathrm{d}x + \int_\Omega \mathrm{div}\left((\alpha-\phi)\,\partial_t u_s + \phi\, v_f\right) p\,\mathrm{d}x = \int_\Omega \frac{\theta}{\rho_f} p\,\mathrm{d}x. \tag{1.9}$$

Adding these three contributions, we see that the terms involving the divergence of the mixture velocity $v_{m,\alpha} = (\alpha-\phi)\partial_t u_s + \phi\, v_f$ cancel, and we have the following energy equality

$$\frac{\rho_s}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega (1-\phi)|\partial_t u_s|^2\,\mathrm{d}x + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x$$
$$+ 2\eta\int_\Omega \varepsilon(\partial_t u_s) : \varepsilon(\partial_t u_s)\,\mathrm{d}x + \frac{\rho_f}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi|v_f|^2\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x$$
$$+ \int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot(v_f - \partial_t u_s)\,\mathrm{d}x + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \frac{\alpha-\phi}{\kappa}|p|^2\,\mathrm{d}x$$
$$= \int_\Omega \rho_s(1-\phi) f\cdot \partial_t u_s\,\mathrm{d}x + \int_\Omega \rho_f \phi f\cdot v_f\,\mathrm{d}x + \int_\Omega \frac{\theta}{\rho_f} p\,\mathrm{d}x + \int_\Omega \theta|v_f|^2\,\mathrm{d}x. \tag{1.10}$$

Remark 1.4. The energy identity (1.10) corresponds exactly to the linearized counterpart of the energy balance (5) derived for the non-linear poromechanics model from [Chapelle and Moireau, 2014].

Consequently, in order to obtain an energy estimate, we impose the following assumptions on the data:

($h1$) The constants ρ_s, ρ_f, μ_f, λ, μ are assumed to be strictly positive, whereas $\eta \geq 0$;

($h2$) The porosity $\phi \in \mathrm{H}^{d/2+r}(\Omega)$ with $r > 0$, and is such that there exists $(\phi_{\min}, \phi_{\max})$ satisfying

$$0 < \phi_{min} \leq \phi(x) \leq \phi_{\max} < 1, \quad \forall x \in \Omega;$$

($h3$) The friction tensor k_f is invertible and there exists $k_{\min}^{-1} > 0$ such that

$$k_f^{-1} v \cdot v \geq k_{\min}^{-1} |v|^2, \quad \forall v \in \mathbb{R}^d;$$

($h4$) $f \in L^2((0,T) \times \Omega)$;

($h5$) $\theta \in \mathrm{C}^0([0,T] \times \Omega)$;

($h6$) The (non-homogeneous) Biot-Willis coefficient $\alpha \in \mathrm{H}^{d/2+r}(\Omega)$ with $r > 0$, and is such that there exists $\big((\alpha-\phi)_{\min}, (\alpha-\phi)_{\max}\big)$ satisfying

$$0 < (\alpha-\phi)_{\min} \leq \alpha(x) - \phi(x) \leq (\alpha-\phi)_{\max} < 1, \quad \forall x \in \Omega;$$

Remark 1.5. The hypotheses ($h2$) and ($h6$) imply that the porosity ϕ and the Biot-Willis coefficient α belong to a multiplior space of $\mathrm{H}^1(\Omega)$. These assumptions are needed to define the term $\operatorname{div}\big((\alpha-\phi)\,\partial_t u_s + \phi\, v_f\big)$ in (1.9). Indeed, if α, $\phi \in \mathrm{H}^{d/2+r}(\Omega)$ with $r > 0$, then for any $(w_s, w_f) \in [\mathrm{H}_0^1(\Omega)]^d$ we have $\operatorname{div}\big((\alpha-\phi)\, w_s + \phi\, w_f\big) \in \mathrm{L}^2(\Omega)$.

Under these assumptions, using Young inequality to bound the right-hand side of (1.10) by

$$\begin{aligned}
&\frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|f|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|\partial_t u_s|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,|f|^2\,\mathrm{d}x \\
&\qquad + \frac{1}{2}\int_\Omega \rho_f\phi\,|v_f|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\kappa}{\rho_f^2(\alpha-\phi)_{\min}}\,|\theta|^2\,\mathrm{d}x \\
&\qquad\qquad + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}\,|p|^2\,\mathrm{d}x + \frac{2\|\theta\|_{\mathrm{C}^0([0,T]\times\Omega)}}{\rho_f\phi_{\min}} \cdot \frac{1}{2}\int_\Omega \rho_f\phi|v_f|^2\,\mathrm{d}x,
\end{aligned}$$

integrating in time from 0 to t and applying Grönwall Lemma, we obtain the following energy bound

$$\begin{aligned}
&\frac{\rho_s}{2}\int_\Omega (1-\phi)|\partial_t u_s(t)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \sigma_s(u_s(t)) : \varepsilon(u_s(t))\,\mathrm{d}x \\
&\quad + 2\eta\int_0^t\int_\Omega \varepsilon(\partial_t u_s) : \varepsilon(\partial_t u_s)\,\mathrm{d}x\,\mathrm{d}s + \frac{\rho_f}{2}\int_\Omega \phi|v_f(t)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}|p(t)|^2\,\mathrm{d}x \\
&\quad + \int_0^t\int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x\,\mathrm{d}s + \int_0^t\int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot(v_f - \partial_t u_s)\,\mathrm{d}x\,\mathrm{d}s \\
&\leq \exp\left(\max\left(1, \frac{2\|\theta\|_{\mathrm{C}^0([0,T]\times\Omega)}}{\rho_f\phi_{\min}}\right)t\right)\left(\left(\frac{\rho_s}{2}(1-\phi_{\min}) + \frac{\rho_f}{2}\phi_{\max}\right)\int_0^t\int_\Omega |f|^2\,\mathrm{d}x\,\mathrm{d}s\right. \\
&\qquad + \frac{\kappa}{2\rho_f^2(\alpha-\phi)_{\min}}\int_0^t\int_\Omega |\theta|^2\,\mathrm{d}x\,\mathrm{d}s + \frac{\rho_s}{2}\int_\Omega (1-\phi)|v_{s0}|^2\,\mathrm{d}x \\
&\qquad\left. + \frac{1}{2}\int_\Omega \sigma_s(u_{s0}) : \varepsilon(u_{s0})\,\mathrm{d}x + \frac{\rho_f}{2}\int_\Omega \phi|v_{f_0}|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}|p_0|^2\,\mathrm{d}x\right). \qquad (1.11)
\end{aligned}$$

Note that the friction contribution induces dissipation in the system since

$$\int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot(v_f - \partial_t u_s)\,\mathrm{d}x \geq 0,$$

in virtue of $(h3)$.

Moreover, Korn inequality [Duvaut and Lions, 1972; Ciarlet, 1988] implies that the fluid and structure dissipative terms are coercive in $[\mathrm{H}_0^1(\Omega)]^d$. Namely, there exists $C_d > 0$ such that

$$\forall v \in [\mathrm{H}_0^1(\Omega)]^d, \quad \int_\Omega \varepsilon(v) : \varepsilon(v)\,\mathrm{d}x \geq C_d \|v\|^2_{[\mathrm{H}_0^1(\Omega)]^d}, \tag{1.12}$$

which implies that the bilinear elastic form is coercive in $[\mathrm{H}_0^1(\Omega)]^d$ and verifies

$$\forall v \in [\mathrm{H}_0^1(\Omega)]^d, \quad \int_\Omega \sigma_s(v) : \varepsilon(v)\,\mathrm{d}x \geq 2\mu C_d \|v\|^2_{[\mathrm{H}_0^1(\Omega)]^d}. \tag{1.13}$$

For the fluid part, thanks to assumption $(h2)$, one also has

$$\forall v \in [\mathrm{H}_0^1(\Omega)]^d, \quad \int_\Omega \phi\,\sigma_f(v) : \varepsilon(v)\,\mathrm{d}x \geq 2\mu_f \phi_{\min} C_d \|v\|^2_{[\mathrm{H}_0^1(\Omega)]^d}. \tag{1.14}$$

Consequently, assuming $(u_{s0}, v_{s0}, v_{f_0}, p_0) \in [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times \mathrm{L}^2(\Omega)$, it follows that $u_s \in \mathrm{L}^\infty(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, $\partial_t u_s \in \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d)$, $v_f \in \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, $p \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$, and that, moreover, $\partial_t u_s \in \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$ if $\eta > 0$.

Note that the energy bound (1.11) depends on the bulk modulus κ. Nonetheless, if θ is regular enough, we can recover an energy estimate independent of κ by coming back to the case where the right-hand side of (1.2c) is equal to zero, as we are now going to perform it in the incompressible case.

Let us now focus on the case $\kappa = +\infty$ for which $\alpha = 1$. The equation (1.2c) reduces to

$$\mathrm{div}\left((1-\phi)\,v_s + \phi\,v_f\right) = \frac{\theta}{\rho_f}, \quad \text{in } \Omega. \tag{1.15}$$

Without loss of generality we can assume that the right-hand side of (1.15) is equal to zero. Indeed, provided that θ is regular enough and that $\int_\Omega \theta\,\mathrm{d}x = 0$, there exists v_θ such that $\mathrm{div}\, v_\theta = \frac{\theta}{\rho_f}$. Considering the system satisfied by $v_s - v_\theta$ and $v_f - v_\theta$, namely defining the new displacement

$$u_{s0} + \int_0^t (v_s - v_\theta)\,\mathrm{d}s = u_s - \int_0^t v_\theta\,\mathrm{d}s,$$

we end up with a system for which the constraint reads $\mathrm{div}\left((1-\phi)\,v_s + \phi\,v_f\right) = 0$. To obtain the energy estimates, we proceed as for the case $\kappa < +\infty$ by multiplying (1.2a) by the structure velocity $\partial_t u_s$, and (1.2b) by the fluid velocity v_f. After integration over the domain and integration by parts, adding these two contributions and taking into account the mixture incompressibility constraint $\mathrm{div}\left((1-\phi)\,v_s + \phi\,v_f\right) = 0$ yields

$$\begin{aligned}
&\frac{\rho_s}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega (1-\phi)|\partial_t u_s|^2\,\mathrm{d}x + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x + 2\eta\int_\Omega \varepsilon(\partial_t u_s) : \varepsilon(\partial_t u_s)\,\mathrm{d}x \\
&\quad + \frac{\rho_f}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi|v_f|^2\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x + \int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot(v_f - \partial_t u_s)\,\mathrm{d}x \\
&\qquad = \int_\Omega \rho_s(1-\phi) f\cdot\partial_t u_s\,\mathrm{d}x + \int_\Omega \rho_f \phi f\cdot v_f\,\mathrm{d}x + \int_\Omega \theta|v_f|^2\,\mathrm{d}x.
\end{aligned} \tag{1.16}$$

Grönwall Lemma then implies

$$\begin{aligned}
&\frac{\rho_s}{2}\int_\Omega (1-\phi)|\partial_t u_s(t)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \sigma_s(u_s(t)) : \varepsilon(u_s(t))\,\mathrm{d}x \\
&\qquad + 2\eta\int_0^t\int_\Omega \varepsilon(\partial_t u_s) : \varepsilon(\partial_t u_s)\,\mathrm{d}x\,\mathrm{d}s + \frac{\rho_f}{2}\int_\Omega \phi|v_f(t)|^2\,\mathrm{d}x \\
&\quad + \int_0^t\int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x\,\mathrm{d}s + \int_0^t\int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot(v_f - \partial_t u_s)\,\mathrm{d}x\,\mathrm{d}s \\
&\leq \exp\left(\max\left(1, \frac{2\|\theta\|_{\mathrm{C}^0([0,T]\times\Omega)}}{\rho_f\phi_{\min}}\right)t\right)\left(\left(\frac{\rho_s}{2}(1-\phi_{\min}) + \frac{\rho_f}{2}\phi_{\max}\right)\int_0^t\int_\Omega |f|^2\,\mathrm{d}x\,\mathrm{d}s\right. \\
&\qquad \left. + \frac{\rho_s}{2}\int_\Omega (1-\phi)|v_{s0}|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \sigma_s(u_{s0}) : \varepsilon(u_{s0})\,\mathrm{d}x + \frac{\rho_f}{2}\int_\Omega \phi|v_{f_0}|^2\,\mathrm{d}x\right). \quad (1.17)
\end{aligned}$$

Thanks to Korn inequality (1.12), coercivities (1.13), (1.14), assumptions $(h1)-(h5)$ and assuming that $(u_{s0}, v_{s0}, v_{f_0}) \in [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d$, we have $u_s \in \mathrm{L}^\infty(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, $\partial_t u_s \in \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d)$, $v_f \in \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, and if $\eta > 0$, $\partial_t u_s \in \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$. Here the energy bounds does not give bounds on the pressure, which is the main difference between the cases $\kappa < +\infty$ and $\kappa = +\infty$.

We then propose the following milestones for our analysis. We start by considering the compressible case for which $\kappa < +\infty$. In this case the pressure p has its own dynamic. Then the incompressible case, namely $\kappa = +\infty$, is treated and we have to deal with a divergence-free constraint on the mixture velocity. Each case is split into two cases: the viscous one (namely $\eta > 0$) for which we have a parabolic-parabolic coupling between the solid and fluid equations, and the inviscid one (namely $\eta = 0$) for which we have a hyperbolic-parabolic coupling. For each four cases we prove existence of strong, mild and variational solutions and give the link between the three types of solutions. In particular, existence of strong and mild solutions relies on the study of the first order system of the form $\dot{z} + Az = g$ associated with (1.2) and the underlying unbounded operator A using semigroup theory. By strong solution, we mean that the solution is regular in time and that the equations are satisfied almost everywhere in the sense that all the components of $\dot{z}$ and Az are defined almost everywhere, whereas mild solutions are solutions satisfying the Duhamel formula. Note that in the case $\eta = 0$ in order to prove that the operator is maximal accretive we need to take care of the non coercivity of the associated bilinear form. This issue is solved thanks to the notion of $\mathtt{T}$-coercivity introduced in [Ciarlet Jr, 2012; Chesnel and Ciarlet, 2013]. Next the variational solutions are obtained by an approximation strategy as the limit of a sequence of strong solutions. Note that, as we will see, the definition of the variational formulations is different when considering $\eta > 0$ or $\eta = 0$. The main difference comes from the fact that, in the latter case, the structure velocity is not in $[\mathrm{H}_0^1(\Omega)]^d$ in space but only in $[\mathrm{L}^2(\Omega)]^d$. We end up with the study of the incompressible limit, which allows to pass to the limit in the weak formulation for $\kappa < +\infty$, to recover the weak formulation associated with $\kappa = +\infty$. The theoretical results are further completed by numerical illustrations to investigate the regularity of the solutions.

1.2 Existence of solutions for a compressible skeleton $\kappa < +\infty$.

In this section, we study the poromechanical problem for a compressible skeleton, that corresponds to $\kappa < +\infty$. First, we write (1.2) as a first-order evolution equation and we define the associated unbounded operator. Then, by investigating the properties of this operator, we use a semigroup approach to show existence and uniqueness of strong and mild solutions to the system. The existence of variational solutions is then obtained by an approximation strategy. The cases $\eta > 0$ and $\eta = 0$

are treated separately in order to emphasize the influence of solid viscosity on the model. But let us start with some general notation and definitions valid for both cases.

1.2.1 Semigroup framework

The system (1.2) can be rewritten as a first order system as follows

$$\begin{cases} \partial_t u_s - v_s = 0, & \text{in } \Omega \times (0,T), \quad (1.18a)\\ \rho_s(1-\phi)\,\partial_t v_s - \operatorname{div}\big(\sigma_s(u_s)\big) - \operatorname{div}\big(\sigma_s^{\mathrm{vis}}(v_s)\big) \\ \quad - \phi^2\, k_f^{-1}(v_f - v_s) + (\alpha-\phi)\,\nabla p = \rho_s(1-\phi)\, f, & \text{in } \Omega \times (0,T), \quad (1.18b)\\ \rho_f \phi\,\partial_t v_f - \operatorname{div}\big(\phi\,\sigma_f(v_f)\big) \\ \quad + \phi^2\, k_f^{-1}(v_f - v_s) - \theta\, v_f + \phi\,\nabla p = \rho_f \phi\, f, & \text{in } \Omega \times (0,T), \quad (1.18c)\\ \dfrac{\alpha-\phi}{\kappa}\,\partial_t p + \operatorname{div}\big((\alpha-\phi)\,v_s + \phi\, v_f\big) = \dfrac{\theta}{\rho_f}, & \text{in } \Omega \times (0,T). \quad (1.18d) \end{cases}$$

Let $z = (u_s, v_s, v_f, p)$ and $z_0 = (u_{s0}, v_{s0}, v_{f_0}, p_0)$ denote respectively the unknown variable and the initial condition of (1.18). We formulate (1.18) as an abstract evolution problem

$$\begin{cases} \dot z(t) + A_\eta^\kappa z(t) + G(t) z(t) = g(t), \quad t \in [0,T], \\ z(0) = z_0, \end{cases} \tag{1.19}$$

where A_η^κ is an unbounded operator specified with respect to the solid viscosity η and the bulk modulus κ, and $G(t)$ is a bounded perturbation defined below.

Let us first define the energy space

$$Z = [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times \mathrm{L}^2(\Omega),$$

associated with (1.18). Since the functions $\rho_s(1-\phi)$, $\rho_f\phi$ and $\frac{\alpha-\phi}{\kappa}$ are bounded and bounded from below by strictly positive constants, the space Z can be endowed with the scalar product defined by

$$(z,y)_Z = \int_\Omega \sigma_s(u_s) : \varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\, v_s \cdot w_s\,\mathrm{d}x + \int_\Omega \rho_f \phi\, v_f \cdot w_f\,\mathrm{d}x + \int_\Omega \frac{\alpha-\phi}{\kappa}\, p\, q\,\mathrm{d}x,$$

for any $z = (u_s, v_s, v_f, p)$ and $y = (d_s, w_s, w_f, q)$ belonging to Z. The associated norm reads

$$\|z\|_Z^2 = \|u_s\|_{\mathrm{s}}^2 + \int_\Omega \rho_s(1-\phi)\,|v_s|^2\,\mathrm{d}x + \int_\Omega \rho_f\phi\,|v_f|^2\,\mathrm{d}x + \int_\Omega \frac{\alpha-\phi}{\kappa}\, p^2\,\mathrm{d}x, \tag{1.20}$$

with

$$\|u_s\|_{\mathrm{s}}^2 = \int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x. \tag{1.21}$$

This norm is equivalent to the canonical norm on Z according to Korn inequality (1.13).

Setting

$$Y = [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d \times \mathrm{L}^2(\Omega),$$

as an intermediate space, we introduce the bilinear form a_η^κ defined for all $z = (u_s, v_s, v_f, p) \in Y$ and $y = (d_s, w_s, w_f, q) \in Y$ by

$$\begin{aligned} a_\eta^\kappa(z,y) = &- \int_\Omega \sigma_s(v_s) : \varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s) : \varepsilon(w_s)\,\mathrm{d}x + 2\eta \int_\Omega \varepsilon(v_s) : \varepsilon(w_s)\,\mathrm{d}x \\ &+ \int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(w_f)\,\mathrm{d}x + \int_\Omega \phi^2\, k_f^{-1}(v_f - v_s)\cdot(w_f - w_s)\,\mathrm{d}x \\ &+ \int_\Omega \operatorname{div}\big((\alpha-\phi)\,v_s + \phi\, v_f\big)\, q\,\mathrm{d}x - \int_\Omega p \operatorname{div}\big((\alpha-\phi)\,w_s + \phi\, w_f\big)\,\mathrm{d}x. \end{aligned} \tag{1.22}$$

The bilinear form a_η^κ is continuous over $Y \times Y$.

Associated with this bilinear form, we introduce the unbounded operator $\big(A_\eta^\kappa, D(A_\eta^\kappa)\big)$ defined by

$$(A_\eta^\kappa z, y)_Z = a_\eta^\kappa(z,y), \quad \forall z \in D(A_\eta^\kappa), \forall y \in Y, \tag{1.23}$$

in the domain

$$D(A_\eta^\kappa) = \big\{ z \in Y : \exists g \in Z,\, a_\eta^\kappa(z,y) = (g,y)_Z, \quad \forall y \in Y \big\}. \tag{1.24}$$

Finally, for all $t \in [0,T]$, we define the time-dependent operator

$$G(t) \,:\, z = (u_s, v_s, v_f, p) \in Z \longmapsto \left(0, 0, -\frac{\theta(t)}{\rho_f \phi}\, v_f, 0\right). \tag{1.25}$$

Taking $g = \left(0, f, f, \frac{\kappa}{\alpha-\phi} \cdot \frac{\theta}{\rho_f}\right)$, the state-space formulation (1.19) is equivalent to (1.18) in a sense that will be specified in Corollary 1.9.

Remark 1.6. Note that in the domain of operator the equation stating that the time derivative of the structure displacement is equal to the structure velocity (that comes from the first order rewriting of a second order in time problem) will hold true in $[\mathrm{H}_0^1(\Omega)]^d$ in the space variable. This is the reason of the presence of the term $-\int_\Omega \sigma_s(v_s) : \varepsilon(d_s)\,\mathrm{d}x$ in (1.22). Yet, even if the solid velocity is considered in $[\mathrm{H}_0^1(\Omega)]^d$ in the latter integral, we will see that when $\eta = 0$ the resulting weak solution does not satisfy (1.18a) in $[\mathrm{H}_0^1(\Omega)]^d$ but only in $[\mathrm{L}^2(\Omega)]^d$. The same issue appears when studying the wave equation.

For $z = (u_s, v_s, v_f, p) \in D(A_\eta^\kappa)$, we can write

$$A_\eta^\kappa z = \begin{pmatrix} -v_s \\ (\rho_s(1-\phi))^{-1}\big(-\mathrm{div}\,(\sigma_s(u_s)) - \mathrm{div}\,(\sigma_s^{\mathrm{vis}}(v_s)) \\ \qquad\qquad +\phi^2\, k_f^{-1}(v_s - v_f) + (\alpha-\phi)\,\nabla p\big) \\ (\rho_f \phi)^{-1}\big(-\mathrm{div}\,(\phi\,\sigma_f(v_f)) + \phi^2\, k_f^{-1}(v_f - v_s) + \phi\,\nabla p\big) \\ \frac{\kappa}{\alpha-\phi}\mathrm{div}\,\big((\alpha-\phi)\,v_s + \phi\, v_f\big) \end{pmatrix}, \tag{1.26}$$

so that the operator A_η^κ can be expressed in matrix form as

$$A_\eta^\kappa = N_0^{-1} \times \begin{pmatrix} 0 & -\mathbb{I} & 0 & 0 \\ -\mathrm{div}\,(\sigma_s(\cdot)) & -\mathrm{div}\,(\sigma_s^{\mathrm{vis}}(\cdot)) + \phi^2\, k_f^{-1} & -\phi^2\, k_f^{-1} & (\alpha-\phi)\,\nabla \\ 0 & -\phi^2\, k_f^{-1} & -\mathrm{div}\,(\phi\,\sigma_f(\cdot)) + \phi^2\, k_f^{-1} & \phi\,\nabla \\ 0 & \mathrm{div}\,((\alpha-\phi)\,\cdot) & \mathrm{div}\,(\phi\,\cdot) & 0 \end{pmatrix},$$

where $\mathbb{I}$ denotes the identity operator of the space $[\mathrm{H}_0^1(\Omega)]^d$ endowed with the norm (1.21), and

$$N_0 = \begin{pmatrix} \mathbb{I} & 0 & 0 & 0 \\ 0 & \rho_s(1-\phi) & 0 & 0 \\ 0 & 0 & \rho_f \phi & 0 \\ 0 & 0 & 0 & \frac{\alpha-\phi}{\kappa} \end{pmatrix}.$$

Moreover, from (1.24) and (1.26), it follows that

$$D(A_\eta^\kappa) = \left\{ \begin{array}{l} u_s, v_s, v_f \in [\mathrm{H}_0^1(\Omega)]^d, p \in \mathrm{L}^2(\Omega) \text{ such that} \\ \qquad -\mathrm{div}\,(\sigma_s(u_s)) - \mathrm{div}\,(\sigma_s^{\mathrm{vis}}(v_s)) + (\alpha-\phi)\,\nabla p \in [\mathrm{L}^2(\Omega)]^d, \\ \qquad -\mathrm{div}\,(\phi\,\sigma_f(v_f)) + \phi\,\nabla p \in [\mathrm{L}^2(\Omega)]^d, \\ \qquad \mathrm{div}\,((\alpha-\phi)\,v_s + \phi\, v_f) \in \mathrm{L}^2(\Omega) \end{array} \right\}. \tag{1.27}$$

Note that belonging to $D(A_\eta^\kappa)$ does not mean that all the above terms individually belong to $\mathrm{L}^2(\Omega)$, but only that their sum does. For instance, (v_f, p) does not necessarily belong to $[\mathrm{H}^2(\Omega)]^d \times \mathrm{H}^1(\Omega)$, but we know that $-\mathrm{div}\,(\phi\,\sigma_f(v_f)) + \phi\,\nabla p \in [\mathrm{L}^2(\Omega)]^d$. Specifying $D(A_\eta^\kappa)$ in terms of classical Sobolev spaces requires to study the regularity of the solution to the static problem $A_\eta^\kappa z = g$ with $g \in Z$. This issue, delicate from a theoretical point of view, will be explored in more details in numerical experiments, see Section 1.5.

In what follows, we exploit the previous framework to prove that Problem (1.18) has a unique strong and mild solution for $\kappa < +\infty$. We also recover the existence of variational solutions as the limit of a sequence of strong solutions. If $\eta > 0$, the solid equation (1.18b) is parabolic, while it becomes hyperbolic when $\eta = 0$. For this reason, we distinguish the cases $\eta > 0$ and $\eta = 0$.

1.2.2 The case $\eta > 0$

Let us start with the parabolic-parabolic coupling configuration. This case was treated in [Burtschell et al., 2019] for $\theta = 0$ and $\alpha = 1$. Here, we propose a proof of existence and uniqueness which is valid for a time-dependent θ. Note that considering $\alpha \neq 1$ does not induce additional difficulties.

Theorem 1.7. *Assume that $(h1), (h2), (h3)$ and $(h6)$ hold true and that $\eta > 0$.*

(i) If $\theta \in \mathrm{C}^1([0,T];\mathrm{L}^\infty(\Omega))$, $z_0 \in D(A_\eta^\kappa)$ and $f \in \mathrm{H}^1(0,T;[\mathrm{L}^2(\Omega)]^d)$, then there exists a unique strong solution $z \in \mathrm{C}^1([0,T];Z) \cap \mathrm{C}^0([0,T];D(A_\eta^\kappa))$ satisfying (1.19).

(ii) If $\theta \in \mathrm{C}^0([0,T]\times\Omega)$, $z_0 \in Z$ and $f \in \mathrm{L}^2(0,T;[\mathrm{L}^2(\Omega)]^d)$, then Problem (1.19) has a unique mild solution $z \in \mathrm{C}^0([0,T];Z)$ such that $z(0) = z_0$ and

$$\int_0^T z(t)\psi(t)\,\mathrm{d}t \in D(A_\eta^\kappa), \tag{1.28}$$

$$-\int_0^T z(t)\dot\psi(t)\,\mathrm{d}t + A_\eta^\kappa\left(\int_0^T z(t)\psi(t)\,\mathrm{d}t\right) + \int_0^T G(t)z(t)\psi(t)\,\mathrm{d}t = \int_0^T g(t)\psi(t)\,\mathrm{d}t, \tag{1.29}$$

for all $\psi \in \mathrm{C}_c^1([0,T];\mathbb{R})$. Moreover, z verifies the Duhamel formula

$$z(t) = \Phi_\eta^\kappa(t)z_0 + \int_0^t \Phi_\eta^\kappa(t-s)\big(-G(s)z(s) + g(s)\big)\,\mathrm{d}s, \tag{1.30}$$

where Φ_η^κ denotes the continuous semigroup generated by A_η^κ in the sense that

$$A_\eta^\kappa x = -\frac{\mathrm{d}}{\mathrm{d}t}\big(\Phi_\eta^\kappa(t)x\big)_{|t=0^+}, \quad x \in Z. \tag{1.31}$$

Proof. Let us prove (ii). We shall first show that the operator A_η^κ defined by (1.23) is maximal-accretive, namely:

- $(A_\eta^\kappa z, z)_Z \geq 0, \quad \forall z \in D(A_\eta^\kappa)$;
- $A_\eta^\kappa + \lambda_0 I$ is surjective from $D(A_\eta^\kappa)$ to Z, for all $\lambda_0 > 0$.

For any $z = (u_s, v_s, v_f, p) \in D(A_\eta^\kappa)$, we have by definition of the bilinear form a_η^κ and the operator A_η^κ

$$(A_\eta^\kappa z, z)_Z = a(z,z) = 2\eta\int_\Omega |\varepsilon(v_s)|^2\,\mathrm{d}x + \int_\Omega \phi^2\,k_f^{-1}(v_f - v_s)\cdot(v_f - v_s)\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x.$$

Since $k_f^{-1}(v_f - v_s) \cdot (v_f - v_s) \geq 0$, we find that $(A_\eta^\kappa z, z)_Z \geq 0$.

Let $\lambda_0 > 0$ be a positive real number and let g be an element of Z. To prove that $A_\eta^\kappa + \lambda_0 I$ is surjective from $D(A_\eta^\kappa)$ to Z, we consider the variational problem

$$\begin{cases} \text{Find } z \in Y \quad \text{such that} \\ \forall y \in Y, \quad a_\eta^\kappa(z,y) + \lambda_0 (z,y)_Z = (g,y)_Z. \end{cases} \tag{1.32}$$

Using Poincaré inequality, we see that the linear form $y \mapsto (g,y)_Z$ is continuous over Y and that the bilinear form $a_\eta^\kappa(\cdot,\cdot) + \lambda_0(\cdot,\cdot)_Z$ is continuous over $Y \times Y$. Moreover,

$$\begin{aligned} a_\eta^\kappa(z,z) + \lambda_0(z,z)_Z &= 2\eta \int_\Omega |\varepsilon(v_s)|^2 \,\mathrm{d}x + \int_\Omega \phi^2 k_f^{-1}(v_f - v_s)\cdot(v_f - v_s)\,\mathrm{d}x + \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x \\ &\quad + \lambda_0 \Big(\|u_s\|_{\mathrm{s}}^2 + \int_\Omega \rho_s (1-\phi)\,|v_s|^2 \,\mathrm{d}x + \int_\Omega \rho_f \phi\, |v_f|^2 \,\mathrm{d}x + \int_\Omega \frac{\alpha - \phi}{\kappa}\, p^2 \,\mathrm{d}x \Big) \\ &\geq \lambda_0 \|u_s\|_{\mathrm{s}}^2 + 2\eta \|\varepsilon(v_s)\|^2 + 2\mu_f \phi_{\min} \|\varepsilon(v_f)\|^2 + \lambda_0 \frac{(\alpha-\phi)_{\min}}{\kappa} \|p\|^2, \end{aligned} \tag{1.33}$$

where $\|\cdot\|$ denotes the L^2 norm indifferently in $[\mathrm{L}^2(\Omega)]^d$ or $\mathrm{L}^2(\Omega)$. Consequently, the bilinear form $a_\eta^\kappa(\cdot,\cdot) + \lambda_0(\cdot,\cdot)_Z$ is coercive on Y thanks to Korn inequality (1.12).

From Lax-Milgram theorem, we deduce that there exists a unique $z \in Y$ solution of (1.32). Since by construction $a_\eta^\kappa(z,y) = (g - \lambda_0 z, y)_Z$ for all $y \in Y$ and $g - \lambda_0 z \in Z$, we finally get that $z \in D(A_\eta^\kappa)$ in view of (1.24).

Hence, A_η^κ is maximal-accretive and Lumer-Phillips theorem (see for instance [Pazy, 2012, Chapter 1, Theorem 4.3]) implies that A_η^κ is the infinitesimal generator – in the sense of (1.31) – of a C^0-semigroup of contraction $(\Phi_\eta^\kappa(t))_{t\geq 0}$. In particular, we have

$$\|\Phi_\eta^\kappa(t)\|_{\mathcal{L}(Z)} \leq 1, \quad t \in [0,T]. \tag{1.34}$$

Then, we observe that $G(t)$ is a bounded perturbation of A_η^κ. Indeed, for any $z \in Z$,

$$\|G(t)z\|_Z^2 = \int_\Omega \rho_f \phi \left(\frac{\theta(t)}{\rho_f \phi} \right)^2 |v_f|^2 \,\mathrm{d}x \leq \omega^2 \|z\|_Z^2,$$

with $(\rho_f \phi_{\min})^{-1}\omega = \|\theta\|_{L^\infty((0,T)\times\Omega)}$. Thus $G \in \mathrm{C}^0([0,T];\mathcal{L}(Z))$ and

$$\|G(t)\|_{\mathcal{L}(Z)} \leq \omega, \quad t \in [0,T]. \tag{1.35}$$

Therefore, the assertion (ii) follows from [Bensoussan et al., 2007, Part II, Chapter 1, Proposition 3.4] and [Burq and Gérard, 2002, Corollary 2.19].

If $\theta \in \mathrm{C}^1([0,T];\mathrm{L}^\infty(\Omega))$ then $G \in \mathrm{C}^1([0,T];\mathcal{L}(Z))$, which proves (i) by an application of [Bensoussan et al., 2007, Part II, Chapter 1, Proposition 3.5]. □

Remark 1.8. The bilinear form $a_\eta^\kappa(\cdot,\cdot) + \lambda_0(\cdot,\cdot)_Z$ is coercive on Y precisely because $\eta > 0$. It will not be the case when $\eta = 0$. In particular, this implies that, here, $(\Phi_\eta^\kappa(t))_{t\geq 0}$ is an analytic semigroup [Bensoussan et al., 2007, Part II, Chapter 1, Theorem 2.12].

The solution $z \in \mathrm{C}^1([0,T];Z) \cap \mathrm{C}^0([0,T];D(A_\eta^\kappa))$, called strong solution in the foregoing, is sometimes referred to as *strict* solution to account for the C^1 regularity in time – see for instance [Bensoussan et al., 2007, Part II, Chapter 1, Definition 3.1]. The next result clarifies in which sense this solution satisifies the original equation under study.

Corollary 1.9. *If $\theta \in \mathrm{C}^1([0,T];\mathrm{L}^\infty(\Omega))$, $z_0 \in D(A_\eta^\kappa)$ and $f \in \mathrm{H}^1(0,T;[\mathrm{L}^2(\Omega)]^d)$, then the strong solution defined above satisfies (1.18) almost everywhere in $(0,T)\times\Omega$.*

Proof. The strong solution satisfies

$$\begin{cases} \dot{z}(t) + A_\eta^\kappa z(t) + G(t)z(t) = g(t), \quad t \in [0,T], \\ z(0) = z_0. \end{cases}$$

Since $z \in \mathrm{C}^1([0,T];Z)$, we know that $\partial_t u_s \in \mathrm{C}^0([0,T];[\mathrm{H}_0^1(\Omega)]^d])$, $\partial_t v_s \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$, $\partial_t v_f \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$ and $\partial_t p \in \mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$. In view of (1.27), the regularity $z \in \mathrm{C}^0([0,T];D(A_\eta^\kappa))$ implies that

$$\begin{aligned} -\mathrm{div}\,(\sigma_s(u_s)) - \mathrm{div}\,(\sigma_s^{\mathrm{vis}}(v_s)) + (\alpha-\phi)\,\nabla p \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d), \\ -\mathrm{div}\,(\phi\,\sigma_f(v_f)) + \phi\,\nabla p \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d), \end{aligned}$$

and $\mathrm{div}\,\big((\alpha-\phi)v_s + \phi\, v_f\big) \in \mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$. Thus for every $t \in [0,T]$, (1.18b), (1.18c) and (1.18d) are verified in $[\mathrm{L}^2(\Omega)]^d$, and in particular almost everywhere. □

In other words, Corollary 1.9 does not mean that each individual term appearing in (1.18) is defined almost everywhere. However, each line of (1.18) is satisfied almost everywhere since $\dot{z}$ and $A_\eta^\kappa z$ are both defined almost everywhere.

Theorem 1.7 provides the existence and uniqueness of two types of solutions: the *strong* solution and the *mild* solution. The strong solution is regular since it belongs to $\mathrm{C}^1([0,T];Z) \cap \mathrm{C}^0([0,T];D(A_\eta^\kappa))$ but it requires high regularity assumptions on the source terms and on the initial conditions, in particular $z_0 \in D(A_\eta^\kappa)$. The mild solution requires weaker assumptions, but the Duhamel formula (1.30) is quite abstract. The next theorem establishes the existence and uniqueness of a third notion of solution: the *variational* solution, that satisfies a weak formulation in the following sense.

Theorem 1.10. *Assume that $(h1)-(h6)$ hold true and that $\eta > 0$. If $z_0 = (u_{s0}, v_{s0}, v_{f_0}, p_0) \in Z$, there exists a variational solution $u_s \in \mathrm{C}^0([0,T];[\mathrm{H}_0^1(\Omega)]^d)$, and $\partial_t u_s \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, and $v_f \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$ and $p \in \mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$ such that*

$$\big(u_s(0), \partial_t u_s(0), v_f(0), p(0)\big) = (u_{s0}, v_{s0}, v_{f_0}, p_0), \tag{1.36}$$

and such that the following equations hold true in $\mathcal{D}'(0,T)$:

$$\begin{cases} \forall (w_s, w_f, q) \in [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d \times \mathrm{L}^2(\Omega), \\ \dfrac{\mathrm{d}^2}{\mathrm{d}t^2}\displaystyle\int_\Omega \rho_s(1-\phi)\,u_s(t)\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s(u_s(t)) : \varepsilon(w_s)\,\mathrm{d}x \\ \qquad +2\eta\displaystyle\int_\Omega \varepsilon(\partial_t u_s(t)) : \varepsilon(w_s)\,\mathrm{d}x - \int_\Omega \phi^2\, k_f^{-1}(v_f(t) - \partial_t u_s(t))\cdot w_s\,\mathrm{d}x \\ \qquad\qquad - \displaystyle\int_\Omega p(t)\,\mathrm{div}\,\big((\alpha-\phi)\,w_s\big)\,\mathrm{d}x = \int_\Omega \rho_s(1-\phi) f(t)\cdot w_s\,\mathrm{d}x, \qquad (1.37a) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \rho_f \phi\, v_f(t)\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f(t)) : \varepsilon(w_f)\,\mathrm{d}x \\ \qquad + \displaystyle\int_\Omega \phi^2\, k_f^{-1}(v_f(t) - \partial_t u_s(t))\cdot w_f\,\mathrm{d}x - \int_\Omega \theta(t)\, v_f(t)\cdot w_f\,\mathrm{d}x \\ \qquad\qquad - \displaystyle\int_\Omega p(t)\,\mathrm{div}\,(\phi\, w_f)\,\mathrm{d}x = \int_\Omega \rho_f \phi f(t)\cdot w_f\,\mathrm{d}x, \qquad (1.37b) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \frac{\alpha-\phi}{\kappa} p(t) q\,\mathrm{d}x + \int_\Omega \mathrm{div}\,\big((\alpha-\phi)\,\partial_t u_s(t) + \phi\, v_f(t)\big) q\,\mathrm{d}x \\ \qquad\qquad = \displaystyle\int_\Omega \frac{\theta(t)}{\rho_f} q\,\mathrm{d}x. \qquad (1.37c) \end{cases}$$

Furthermore, the energy estimate (1.11) *holds true and, if we assume that* $\partial_t\theta \in \mathrm{L}^\infty((0,T)\times\Omega))$, *this solution is unique.*

Proof. To show the existence of variational solutions verifying (1.37), we proceed as follows. First, we approximate the data by sequences of regular functions and we consider the sequence of strong solutions associated with these regular data. Then, we show that these strong solutions satisfy a variational formulation and we pass to the limit on this formulation after having established some *a priori* estimates and strong convergences of the sequences.

As A_η^κ is maximal, $D(A_\eta^\kappa)$ is dense in Z. Let z_0^n be a sequence of elements of $D(A_\eta^\kappa)$ converging towards z_0 strongly in Z. Let f^n denote a sequence of $\mathrm{H}^1(0,T;\mathrm{L}^2(\Omega))$ converging towards f in $\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$ and θ^n denote a sequence of $\mathrm{C}^1([0,T];\mathrm{L}^\infty(\Omega))$ converging towards θ in $\mathrm{C}^0([0,T]\times\Omega)$. From Theorem 1.7, we know that there exists a unique strong solution $z^n = (u_s^n, v_s^n, v_f^n, p^n) \in \mathrm{C}^1([0,T];Z)\cap\mathrm{C}^0([0,T];D(A_\eta^\kappa))$ to the problem

$$\begin{cases} \dot z^n(t) + A_\eta^\kappa z^n(t) + G^n(t)z^n(t) = g^n(t), \quad t\in[0,T], \\ z^n(0) = z_0^n. \end{cases} \tag{1.38}$$

Multiplying (1.38) by $y = (d_s, w_s, w_f, p) \in Y$, we see from (1.23) that $(A_\eta^\kappa z^n(t), y)_Z = a_\eta^\kappa(z^n(t), y)$. Hence z^n satisfies the following variational formulation:
for all $s\in[0,T]$,

$$(VF)^n \begin{cases} \forall (d_s, w_s, w_f, p) \in Y = [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d \times \mathrm{L}^2(\Omega), \\ \displaystyle\int_\Omega \sigma_s(\partial_t u_s^n(s)) : \varepsilon(d_s)\,\mathrm{d}x = \int_\Omega \sigma_s(v_s^n(s)) : \varepsilon(d_s)\,\mathrm{d}x, \\ \quad \displaystyle\int_\Omega \rho_s(1-\phi)\,\partial_t v_s^n(s)\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s(u_s^n(s)) : \varepsilon(w_s)\,\mathrm{d}x \\ \qquad +2\eta\displaystyle\int_\Omega \varepsilon(v_s^n(s)) : \varepsilon(w_s)\,\mathrm{d}x - \int_\Omega \phi^2\, k_f^{-1}(v_f^n(s) - v_s^n(s))\cdot w_s\,\mathrm{d}x \\ \qquad\quad -\displaystyle\int_\Omega p^n(s)\,\mathrm{div}\,((\alpha-\phi)\,w_s)\,\mathrm{d}x = \int_\Omega \rho_s(1-\phi) f^n(s)\cdot w_s\,\mathrm{d}x, \\ \displaystyle\int_\Omega \rho_f\phi\,\partial_t v_f^n(s)\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f^n(s)) : \varepsilon(w_f)\,\mathrm{d}x \\ \quad +\displaystyle\int_\Omega \phi^2\, k_f^{-1}(v_f^n(s) - v_s^n(s))\cdot w_f\,\mathrm{d}x - \int_\Omega \theta^n(s)\, v_f^n(s)\cdot w_f\,\mathrm{d}x \\ \qquad -\displaystyle\int_\Omega p^n(s)\,\mathrm{div}\,(\phi\, w_f)\,\mathrm{d}x = \int_\Omega \rho_f\phi f^n(s)\cdot w_f\,\mathrm{d}x, \\ \displaystyle\int_\Omega \frac{\alpha-\phi}{\kappa}\partial_t p^n(s)\, q\,\mathrm{d}x \\ \quad +\displaystyle\int_\Omega \mathrm{div}\,\big((\alpha-\phi)\,v_s^n(s) + \phi\, v_f^n(s)\big) q\,\mathrm{d}x = \int_\Omega \frac{\theta^n(s)}{\rho_f} q\,\mathrm{d}x. \end{cases}$$

Recalling that $z^n \in \mathrm{C}^0([0,T];D(A_\eta^\kappa)) \subset \mathrm{C}^0([0,T];Y)$, we can choose $d_s = u_s^n(s)$, $w_s = v_s^n(s)$, $w_f = v_f^n(s)$ and $q = p^n(s)$ as test functions. Integrating in time from 0 to t and applying Grönwall

Lemma like in Section 1.1, we get the energy inequality

$$\begin{aligned}
&\frac{1}{2}\int_\Omega \sigma_s(u_s^n(t)) : \varepsilon(u_s^n(t))\,\mathrm{d}x + \frac{\rho_s}{2}\int_\Omega (1-\phi)|v_s^n(t)|^2\,\mathrm{d}x + 2\eta\int_0^t\int_\Omega \varepsilon(v_s^n) : \varepsilon(v_s^n)\,\mathrm{d}x\,\mathrm{d}s \\
&\qquad + \frac{\rho_f}{2}\int_\Omega \phi|v_f^n(t)|^2\,\mathrm{d}x + \int_0^t\int_\Omega \phi\,\sigma_f(v_f^n) : \varepsilon(v_f^n)\,\mathrm{d}x\,\mathrm{d}s + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}|p^n(t)|^2\,\mathrm{d}x \\
&\quad \le \exp\left(\max\left(1, \frac{2\|\theta^n\|_{\mathrm{C}^0([0,T]\times\Omega)}}{\rho_f\phi_{\min}}\right)t\right)\left(\left(\frac{\rho_s}{2}(1-\phi_{\min}) + \frac{\rho_f}{2}\phi_{\max}\right)\int_0^t\int_\Omega |f^n|^2\,\mathrm{d}x\,\mathrm{d}s\right. \\
&\qquad + \frac{\kappa}{2\rho_f^2(\alpha-\phi)_{\min}}\int_0^t\int_\Omega |\theta^n|^2\,\mathrm{d}x\,\mathrm{d}s + \frac{1}{2}\int_\Omega \sigma_s(u_{s\,0}^n) : \varepsilon(u_{s\,0}^n)\,\mathrm{d}x \\
&\qquad\left. + \frac{\rho_s}{2}\int_\Omega (1-\phi)|v_{s\,0}^n|^2\,\mathrm{d}x + \frac{\rho_f}{2}\int_\Omega \phi|v_{f\,0}^n|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}|p_0^n|^2\,\mathrm{d}x\right). \quad (1.39)
\end{aligned}$$

Thanks to the assumptions done on the data, the right-hand side of the latter inequality is uniformly bounded with respect to n. Consequently, taking into account the assumptions $(h2)$ and $(h6)$ on ϕ and α, Korn inequality (1.12), the coercivity of the elastic and fluid forms (1.13) and (1.14), we deduce that u_s^n is uniformly bounded in $\mathrm{C}^0([0,T];[\mathrm{H}_0^1(\Omega)]^d)$, v_s^n in $\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, v_f^n in $\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$ and p^n in $\mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$.

Similarly, one can show that z^n is a Cauchy sequence in $\mathrm{C}^0([0,T];[\mathrm{H}_0^1(\Omega)]^d)\times\big(\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)\big)\times\big(\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)\big)\times\mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$. Indeed, denoting $z^{n,m}=z^n-z^m$, using the linearity of the coupled problem, the uniform bound we just obtained and taking into account the fact that $\|\theta^n\|_{L^\infty((0,T)\times\Omega)}$ is bounded uniformly in n, we obtain that there exists $C>0$ independent of n such that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\chi^{n,m}(t) \le C\chi^{n,m}(t) + \frac{\rho_s}{2}\int_\Omega (1-\phi)|f^{n,m}(t)|^2\,\mathrm{d}x \\
+ \frac{\rho_f}{2}\int_\Omega \phi|f^{n,m}(t)|^2\,\mathrm{d}x + C\|\theta^{n,m}\|_{L^\infty((0,T)\times\Omega)},
\end{aligned}$$

with

$$\begin{aligned}
\chi^{n,m}(t) = \frac{\rho_s}{2}\int_\Omega (1-\phi)|v_s^{n,m}(t)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \sigma_s(u_s^{n,m}(t)) : \varepsilon(u_s^{n,m}(t))\,\mathrm{d}x \\
+ \frac{\rho_f}{2}\int_\Omega \phi|v_f^{n,m}(t)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}\,|p^{n,m}(t)|^2\,\mathrm{d}x \\
+ 2\eta\int_0^t\int_\Omega \varepsilon(v_s^{n,m}) : \varepsilon(v_s^{n,m})\,\mathrm{d}x\,\mathrm{d}s + \int_0^t\int_\Omega \phi\,\sigma_f(v_f^{n,m}) : \varepsilon(v_f^{n,m})\,\mathrm{d}x\,\mathrm{d}s.
\end{aligned}$$

Grönwall Lemma and the fact that the sequences associated with the data are Cauchy sequences imply that

$$\forall\delta>0,\ \exists N\in\mathbb{N},\ \forall n\ge N,\ \forall m\ge N,\quad \chi^{n,m}(t)\le\delta\exp(Ct).$$

Consequently, there exists $u_s\in\mathrm{C}^0([0,T];[\mathrm{H}_0^1(\Omega)]^d)$, $v_s\in\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, $v_f\in\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$ and moreover, $p\in\mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$ such that

$$\begin{aligned}
&u_s^n\longrightarrow u_s \quad \text{in } \mathrm{C}^0([0,T];[\mathrm{H}_0^1(\Omega)]^d),\quad v_s^n\longrightarrow v_s \quad \text{in } \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d),\\
&v_f^n\longrightarrow v_f \quad \text{in } \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d),\quad p^n\longrightarrow p \quad \text{in } \mathrm{C}^0([0,T];\mathrm{L}^2(\Omega)),
\end{aligned}$$

while

$$v_s^n\longrightarrow v_s \quad \text{in } \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d),\quad v_f^n\longrightarrow v_f \quad \text{in } \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d).$$

Note that since $v_s^n = \partial_t u_s^n$, it also holds true in the limit and $\partial_t u_s = v_s \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$. These convergences enable to pass to the limit in $(VF)^n$, except for the inertial terms. Yet, these terms can be rewritten thanks to an integration by parts in time as follows: for $\psi \in \mathcal{D}(0,T)$, we have for example

$$\int_0^T \int_\Omega \rho_s(1-\phi)\,\partial_t v_s^n(t)\cdot w_s \psi(t)\,\mathrm{d}x\,\mathrm{d}t = -\int_0^T \int_\Omega \rho_s(1-\phi)\,v_s^n(t)\cdot w_s \dot{\psi}(t)\,\mathrm{d}x\,\mathrm{d}t \\ \underset{n\to\infty}{\longrightarrow} -\int_0^T \int_\Omega \rho_s(1-\phi)\,v_s(t)\cdot w_s \dot{\psi}(t)\,\mathrm{d}x\,\mathrm{d}t.$$

By similar arguments and thanks to the strong convergences, we get, in $\mathcal{D}'(0,T)$,

$$\begin{cases} \forall (d_s, w_s, w_f, p) \in Y, \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \sigma_s(u_s(t)) : \varepsilon(d_s)\,\mathrm{d}x = \int_\Omega \sigma_s(v_s(t)) : \varepsilon(d_s)\,\mathrm{d}x, & (1.40a) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \rho_s(1-\phi)\,v_s(t)\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s(u_s(t)) : \varepsilon(w_s)\,\mathrm{d}x \\ \qquad +2\eta\displaystyle\int_\Omega \varepsilon(v_s(t)) : \varepsilon(w_s)\,\mathrm{d}x - \int_\Omega \phi^2\, k_f^{-1}(v_f(t)-v_s(t))\cdot w_s\,\mathrm{d}x \\ \qquad -\displaystyle\int_\Omega p(t)\,\mathrm{div}\big((\alpha-\phi)\,w_s\big)\,\mathrm{d}x = \int_\Omega \rho_s(1-\phi) f(t)\cdot w_s\,\mathrm{d}x, & (1.40b) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \rho_f \phi\, v_f(t)\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f(t)) : \varepsilon(w_f)\,\mathrm{d}x \\ \qquad +\displaystyle\int_\Omega \phi^2\, k_f^{-1}(v_f(t)-v_s(t))\cdot w_f\,\mathrm{d}x - \int_\Omega \theta(t)\,v_f(t)\cdot w_f\,\mathrm{d}x \\ \qquad -\displaystyle\int_\Omega p(t)\,\mathrm{div}\,(\phi\, w_f)\,\mathrm{d}x = \int_\Omega \rho_f \phi f(t)\cdot w_f\,\mathrm{d}x, & (1.40c) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \frac{\alpha-\phi}{\kappa} p(t)\,q\,\mathrm{d}x + \int_\Omega \mathrm{div}\big((\alpha-\phi)\,v_s(t) + \phi\, v_f(t)\big) q\,\mathrm{d}x \\ \qquad = \displaystyle\int_\Omega \frac{\theta(t)}{\rho_f} q\,\mathrm{d}x. & (1.40d) \end{cases}$$

To obtain (1.37), it only remains to rewrite (1.40a) and (1.40b) as a second order equation in time, which holds true since $v_s = \partial_t u_s$ in $\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$. Lastly, we recover the initial conditions (1.36) by simply passing to the limit in the second line of (1.38).

To ensure uniqueness, we observe that every variational solution satisfying the energy estimate is unique. Indeed, for $f = 0$, $\theta = 0$ and $z_0 = 0$, we obtain $z = 0$ in virtue of (1.11). Therefore, it is sufficient to prove that every variational solution satisfying (1.37) verifies the energy identity (1.10) and thus the energy estimate. To do so, let us first derive a bound on $(\partial_t u_s, \partial_t v_s, \partial_t v_f, \partial_t p)$. From (1.40), we deduce

$$\forall y \in Y, \quad -\int_0^T \big(z(t),y\big)_Z \dot{\psi}(t)\,\mathrm{d}t + \int_0^T a_\eta^\kappa\big(z(t),y\big)\psi(t)\,\mathrm{d}t \\ + \int_0^T \big(G(t)z(t),y\big)_Z \psi(t)\,\mathrm{d}t = \int_0^T \big(g(t),y\big)_Z \psi(t)\,\mathrm{d}t. \quad (1.41)$$

Since $f \in \mathrm{L}^2(0,T;[\mathrm{L}^2(\Omega)]^d)$, $\theta \in \mathrm{C}^0((0,T)\times\Omega)$ and by continuity of the bilinear form a_η^κ over $Y \times Y$, we have

$$\forall y \in Y, \quad -\int_0^T \big(z(t),y\big)_Z \dot{\psi}(t)\,\mathrm{d}t = \int_0^T \big(h(t),y\big)_Z \psi(t)\,\mathrm{d}t \quad \text{with } h \in \mathrm{L}^2(0,T;Y').$$

Thus $\dot{z} \in \mathrm{L}^2(0,T;Y')$, namely $(\partial_t u_s, \partial_t v_s, \partial_t v_f, \partial_t p)$ belongs to

$$\mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d) \times \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d) \times \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d) \times \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega)).$$

Since $\partial_t u_s = v_s$, finally, it holds

$$\begin{aligned} \partial_t u_s \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d) \quad &\text{and} \quad \partial^2_{tt} u_s \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d), \\ v_f \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d) \quad &\text{and} \quad \partial_t v_f \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d), \\ p \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega)) \quad &\text{and} \quad \partial_t p \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega)). \end{aligned} \tag{1.42}$$

Using a standard result of functional analysis (see for instance [Dautray and Lions, 1992, Chapter XVIII, Proposition 7]), the previous regularities imply that the following relations hold in $\mathcal{D}'(0,T)$:

$$\begin{aligned} \forall w_s \in [\mathrm{H}^1_0(\Omega)]^d, \quad & \frac{\mathrm{d}^2}{\mathrm{d}t^2} \int_\Omega \rho_s(1-\phi)\, u_s(t) \cdot w_s \,\mathrm{d}x \\ & \qquad = \Big\langle \rho_s(1-\phi)\, \partial^2_{tt} u_s(t), w_s \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}^1_0(\Omega)]^d}, \\ \forall w_f \in [\mathrm{H}^1_0(\Omega)]^d, \quad & \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \rho_f \phi\, v_f(t) \cdot w_f \,\mathrm{d}x = \Big\langle \rho_f \phi\, \partial_t v_f(t), w_f \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}^1_0(\Omega)]^d}, \\ \forall q \in \mathrm{L}^2(\Omega), \quad & \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \frac{\alpha-\phi}{\kappa} p(t)\, q \,\mathrm{d}x = \int_\Omega \frac{\alpha-\phi}{\kappa} \partial_t p(t)\, q \,\mathrm{d}x. \end{aligned}$$

Moreover, since functions in $[\mathrm{H}^1_0(\Omega)]^d \otimes \mathcal{D}(0,T)$ and $\mathrm{L}^2(\Omega) \otimes \mathcal{D}(0,T)$ generate respectively the spaces $\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$ and $\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$, we obtain the space-time variational formulation

$$\left\{ \begin{aligned} &\forall (w_s, w_f, q) \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d) \times \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d) \times \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega)), \\ &\int_0^T \Big\langle \rho_s(1-\phi)\, \partial^2_{tt} u_s, w_s \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}^1_0(\Omega)]^d} \mathrm{d}t + \int_0^T \int_\Omega \sigma_s(u_s) : \varepsilon(w_s) \,\mathrm{d}x\,\mathrm{d}t \\ &\quad + 2\eta \int_0^T \int_\Omega \varepsilon(\partial_t u_s) : \varepsilon(w_s) \,\mathrm{d}x\,\mathrm{d}t - \int_0^T \int_\Omega \phi^2\, k_f^{-1} (v_f - \partial_t u_s) \cdot w_s \,\mathrm{d}x\,\mathrm{d}t \\ &\quad - \int_0^T \int_\Omega p \,\mathrm{div}\,\big((\alpha-\phi)\, w_s\big) \,\mathrm{d}x\,\mathrm{d}t = \int_0^T \int_\Omega \rho_s(1-\phi) f \cdot w_s \,\mathrm{d}x\,\mathrm{d}t, \\ &\int_0^T \Big\langle \rho_f \phi\, \partial_t v_f, w_f \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}^1_0(\Omega)]^d} \mathrm{d}t + \int_0^T \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(w_f) \,\mathrm{d}x\,\mathrm{d}t \\ &\quad + \int_0^T \int_\Omega \phi^2\, k_f^{-1} (v_f - \partial_t u_s) \cdot w_f \,\mathrm{d}x\,\mathrm{d}t v - \int_0^T \int_\Omega \theta\, v_f \cdot w_f \,\mathrm{d}x\,\mathrm{d}t \\ &\quad - \int_0^T \int_\Omega p \,\mathrm{div}\,(\phi\, w_f) \,\mathrm{d}x\,\mathrm{d}t = \int_0^T \int_\Omega \rho_f \phi f \cdot w_f \,\mathrm{d}x\,\mathrm{d}t, \\ &\int_0^T \int_\Omega \frac{\alpha-\phi}{\kappa} \partial_t p\, q \,\mathrm{d}x\,\mathrm{d}t \\ &\quad + \int_0^T \int_\Omega \mathrm{div}\,\big((\alpha-\phi)\, \partial_t u_s + \phi\, v_f\big) q \,\mathrm{d}x\,\mathrm{d}t = \int_0^T \int_\Omega \frac{\theta}{\rho_f} q \,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \right.$$

Now, since we know $\partial_t u_s \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$, $v_f \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$ and $p \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$, we can choose $(w_s, w_f, q) = (\partial_t u_s, v_f, p)$ as test functions in the above formulation, which provides the energy identity (1.10) and thus the energy estimate (1.11). □

Remark 1.11. The method used to prove Theorem 1.10 is standard and close to the Faedo-Galerkin method. The difference with Faedo-Galerkin method is that the approximated sequence is directly

recovered from the existence of strong solutions instead of being constructed on a suitable finite dimensional space. This allows us to obtain strong convergence for the whole sequence, whereas Faedo-Galerkin method provides only weak convergence of subsequences. In addition, it directly provides the continuity with respect to time of the solution and the strong convergence of the initial condition $z(0) = z_0$ in Z.

Remark 1.12. The variational solution could also be defined without assuming that it is continuous with respect to time, but only assuming that the regularities (1.42) are satisfied. The time continuity of the solution can then be recovered using the existence of a continuous and linear mapping of the space

$$W(0,T) = \big\{u \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d) \text{ such that } \partial_t u \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d)\big\},$$

into $\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$, see [Lions and Magenes, 1972, Chapter 1, Theorem 3.1].

The mild solution and the variational solution are two notions of solution whose existence and uniqueness here require the same hypotheses on the data. In fact, the following result states that these two types of solution are the same whenever $f \in L^2((0,T)\times\Omega)$ and $\theta \in \mathrm{C}^0([0,T]\times\Omega)$. Thus, they can be used indifferently depending on the context. For instance, the mild solution is widely used in control theory because of the practical aspects of Duhamel formula, whereas the variational solution formulation is usually the one implemented at the discrete level when considering finite element discretization.

Proposition 1.13. *If $f \in L^2((0,T)\times\Omega)$ and $\theta \in \mathrm{C}^0([0,T]\times\Omega)$, then the mild solution given by (1.30) and the variational solution satisfying (1.37) coincide.*

Proof. The mild solution and the variational solution are both unique. Hence, it is sufficient to show that the variational solution defined in Theorem 1.10 is also a mild solution, namely that it satisfies (1.28) and (1.29).

Let ψ be given in $\mathcal{D}(0,T)$. Since $v_s \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$, $v_f \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$, it holds that

$$a^\kappa_\eta\Big(\int_0^T z(t)\psi(t)\,\mathrm{d}t, y\Big) = \int_0^T a^\kappa_\eta\big(z(t),y\big)\psi(t)\,\mathrm{d}t,$$

for all $y \in Y$, so that we can rewrite (1.41) as

$$\forall y \in Y, \quad -\int_0^T \big(z(t),y\big)_Z\dot\psi(t)\,\mathrm{d}t + a^\kappa_\eta\Big(\int_0^T z(t)\psi(t)\,\mathrm{d}t, y\Big) + \int_0^T \big(G(t)z(t),y\big)_Z\psi(t)\,\mathrm{d}t = \int_0^T \big(g(t),y\big)_Z\psi(t)\,\mathrm{d}t.$$

From the definition of $D(A^\kappa_\eta)$, it follows that

$$\int_0^T z(t)\psi(t)\,\mathrm{d}t \in D(A^\kappa_\eta),$$

and thus

$$-\int_0^T \big(z(t),y\big)_Z\dot\psi(t)\,\mathrm{d}t + \Big(A^\kappa_\eta\Big(\int_0^T z(t)\psi(t)\,\mathrm{d}t\Big), y\Big)_Z + \int_0^T \big(G(t)z(t),y\big)_Z\psi(t)\,\mathrm{d}t = \int_0^T \big(g(t),y\big)_Z\psi(t)\,\mathrm{d}t, \quad (1.44)$$

for all $y \in Y$. As Y is dense in Z, (1.44) is also true for all $y \in Z$, which proves (1.29). □

Remark 1.14. Note that the mild and variational solutions coincide only under the assumption $f \in L^2((0,T)\times\Omega)$ and $\theta \in \mathrm{C}^0([0,T]\times\Omega)$. If $f \in L^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d)$, we can readily extend the existence and uniqueness of variational solutions proved in Theorem 1.10, but the existence of a mild solution is not guaranteed.

1.2.3 The case $\eta = 0$

Without solid viscosity, the solid formulation becomes hyperbolic. This hyperbolic-parabolic coupling was studied in [Barnafi et al., 2021], where existence of variational solutions is derived. In [Barnafi et al., 2021], the fluid mass input θ is supposed to be small enough, namely there exists $C_f > 0$ such that

$$\forall v_f \in [\mathrm{H}^1_0(\Omega)]^d, \quad \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(v_f)\, \mathrm{d}x - \int_\Omega \theta\, |v_f|^2 \, \mathrm{d}x \geq C_f \|v_f\|^2_{[\mathrm{H}^1_0(\Omega)]^d}.$$

Here, we are going to prove existence results of strong and mild solutions thanks to semigroup theory and deduce existence of variational solutions directly, without any smallness assumption on θ.

The main issue in this case is that the underlying bilinear form is not coercive. Indeed if $\eta = 0$, then the bilinear form introduced in the proof of Theorem 1.7 is no more coercive on Y in view of (1.33). Despite this lack of coercivity, we are going to show that Problem (1.32) is still well-posed when $\eta = 0$. To do so, we use the $\mathtt{T}$-coercivity approach [Ciarlet Jr, 2012; Chesnel and Ciarlet, 2013], which is a reformulation of Banach-Nečas-Babuška theory [Ern and Guermond, 2021a, Theorem 25.9] and that has been designed especially for non-coercive problems. For the sake of completeness, the definition and properties of $\mathtt{T}$-coercivity are recalled below.

Definition 1.15. [Chesnel and Ciarlet, 2013, Definition 3] Let V be an Hilbert space and let $a(\cdot,\cdot)$ be a continuous bilinear form over $V \times V$. We say that a is $\mathtt{T}$-coercive if there exists a bijective application $\mathtt{T} \in \mathcal{L}(V)$ and $\underline{\alpha} > 0$ such that

$$|a(z, \mathtt{T}z)| \geq \underline{\alpha} \|z\|^2_V, \quad z \in V.$$

Proposition 1.16. *[Chesnel and Ciarlet, 2013, Theorem 1] Let V be an Hilbert space. Let $\ell(\cdot)$ be a continuous linear form over V and $a(\cdot,\cdot)$ be a continuous bilinear form over $V \times V$. The problem*

$$\begin{cases} \text{Find } z \in V \quad \text{such that} \\ \forall y \in V, \quad a(z,y) = \ell(y), \end{cases}$$

is well-posed if and only if a is $\mathtt{T}$-coercive.

The following theorem states existence and uniqueness of solutions to Problem (1.19) in the case $\eta = 0$.

Theorem 1.17. *If $\eta = 0$, then the conclusions of Theorem 1.7 remain true.*

Proof. Let us show that A_0^κ is maximal by proving that the variational problem

$$\begin{cases} \text{Find } z \in Y \quad \text{such that} \\ \forall y \in Y, \quad a_0^\kappa(z,y) + \lambda_0 (z,y)_Z = (g,y)_Z, \end{cases}$$

is well-posed, where

$$\begin{aligned} a_0^\kappa(z,y) + \lambda_0 (z,y)_Z = & \int_\Omega \lambda_0\, \sigma_s(u_s) : \varepsilon(d_s)\, \mathrm{d}x - \int_\Omega \sigma_s(v_s) : \varepsilon(d_s)\, \mathrm{d}x + \int_\Omega \lambda_0 \rho_s (1-\phi)\, v_s \cdot w_s\, \mathrm{d}x \\ & + \int_\Omega \sigma_s(u_s) : \varepsilon(w_s)\, \mathrm{d}x + \int_\Omega \phi^2\, k_f^{-1} (v_f - v_s) \cdot (w_f - w_s)\, \mathrm{d}x \\ & + \int_\Omega \lambda_0 \rho_f \phi\, v_f \cdot w_f\, \mathrm{d}x + \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(w_f)\, \mathrm{d}x + \int_\Omega \lambda_0 \frac{\alpha - \phi}{\kappa}\, p\, q\, \mathrm{d}x \\ & - \int_\Omega p\, \mathrm{div}\left((\alpha - \phi)\, w_s + \phi\, w_f\right) \mathrm{d}x + \int_\Omega \mathrm{div}\left((\alpha - \phi)\, v_s + \phi\, v_f\right) q\, \mathrm{d}x, \end{aligned}$$

for any $z = (u_s, v_s, v_f, p)$ and $y = (d_s, w_s, w_f, q)$ in Y. From Proposition 1.16, it is sufficient to show that $a_0^\kappa(\cdot,\cdot) + \lambda_0(\cdot,\cdot)_Z$ is $\mathtt{T}$-coercive.

For a given z, we look for a y^* depending continuously on z such that $a_0^\kappa(z, y^*) + \lambda_0(z, y^*)_Z \geq \underline{\alpha}\|z\|_Y^2$ for some constant $\underline{\alpha} > 0$. Choosing $w_s^* = v_s$, $w_f^* = v_f$, $q^* = p$ and d_s^* in the form $\beta\, u_s + \gamma\, v_s$ yields

$$
\begin{aligned}
a_0^\kappa(z, y^*) + \lambda_0(z, y^*)_Z = & \int_\Omega \lambda_0 \beta\, \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x + \int_\Omega \lambda_0 \gamma\, \sigma_s(u_s) : \varepsilon(v_s)\,\mathrm{d}x \\
& - \int_\Omega \beta\, \sigma_s(v_s) : \varepsilon(u_s)\,\mathrm{d}x - \int_\Omega \gamma\, \sigma_s(v_s) : \varepsilon(v_s)\,\mathrm{d}x + \int_\Omega \lambda_0 \rho_s (1-\phi)\,|v_s|^2\,\mathrm{d}x \\
& + \int_\Omega \sigma_s(u_s) : \varepsilon(v_s)\,\mathrm{d}x + \int_\Omega \phi^2\, k_f^{-1}(v_f - v_s)\cdot(v_f - v_s)\,\mathrm{d}x \\
& + \int_\Omega \lambda_0 \rho_f \phi\,|v_f|^2\,\mathrm{d}x + \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x + \int_\Omega \lambda_0 \frac{\alpha - \phi}{\kappa}\, p^2\,\mathrm{d}x.
\end{aligned}
$$

By setting $\beta = \frac{1}{2}$ and $\gamma = -\frac{1}{2\lambda_0}$, the terms of the form $\int_\Omega \sigma_s(u_s) : \varepsilon(v_s)\,\mathrm{d}x$ vanish so that

$$
a_0^\kappa(z, y^*) + \lambda_0(z, y^*)_Z \geq \frac{\lambda_0}{2}\|u_s\|_{\mathrm{s}}^2 + \frac{1}{2\lambda_0}\|v_s\|_{\mathrm{s}}^2 + 2\mu_f \phi_{\min}\|\varepsilon(v_f)\|^2 + \lambda_0 \frac{(\alpha-\phi)_{\min}}{\kappa}\|p\|^2.
$$

Therefore, $a_0^\kappa(\cdot,\cdot) + \lambda_0(\cdot,\cdot)_Z$ is $\mathtt{T}$-coercive for the mapping $\mathtt{T}$ defined by

$$
\mathtt{T} : (u_s, v_s, v_f, p) \longmapsto \left(\frac{1}{2}u_s - \frac{1}{2\lambda_0}v_s, v_s, v_f, p\right), \tag{1.45}
$$

which is continuous and bijective on Y.

The remainder of the proof follows the very same lines as for the viscous case. □

Next we recover the existence of variational solutions from the existence of strong solutions. However, we obtain a variational formulation that slightly differs from (1.37) because of the hyperbolic-parabolic coupling between the solid and fluid equations.

Theorem 1.18. *Assume that $(h1)-(h6)$ hold true, that $\partial_t\theta \in \mathrm{L}^\infty((0,T)\times\Omega)$, and that $\eta = 0$. If $z_0 = (u_{s0}, v_{s0}, v_{f_0}, p_0) \in Z$, there exists a variational solution $u_s \in \mathrm{C}^0([0,T];[\mathrm{H}_0^1(\Omega)]^d)$, $\partial_t u_s \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$, $v_f \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$ and $p \in \mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$ such that*

$$
\big(u_s(0), \partial_t u_s(0), v_f(0), p(0)\big) = (u_{s0}, v_{s0}, v_{f_0}, p_0), \tag{1.46}
$$

and the following equations hold true, in $\mathcal{D}'(0,T)$,

$$\begin{cases} \forall (w_s, w_f, q) \in [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times \mathrm{L}^2(\Omega), \\ \dfrac{\mathrm{d}^2}{\mathrm{d}t^2} \displaystyle\int_\Omega \rho_s (1-\phi)\, u_s(t) \cdot w_s \,\mathrm{d}x + \int_\Omega \sigma_s(u_s(t)) : \varepsilon(w_s)\,\mathrm{d}x \\ \qquad - \displaystyle\int_\Omega \phi^2\, k_f^{-1} (v_f(t) - \partial_t u_s(t)) \cdot w_s \,\mathrm{d}x - \int_\Omega p(t) \,\mathrm{div}\left((\alpha-\phi)\, w_s\right) \mathrm{d}x \\ \qquad\qquad = \displaystyle\int_\Omega \rho_s (1-\phi) f(t) \cdot w_s \,\mathrm{d}x, \qquad (1.47a) \\ \dfrac{\mathrm{d}}{\mathrm{d}t} \displaystyle\int_\Omega \rho_f \phi\, v_f(t) \cdot w_f \,\mathrm{d}x + \int_\Omega \phi\, \sigma_f(v_f(t)) : \varepsilon(w_f)\,\mathrm{d}x \\ \qquad + \displaystyle\int_\Omega \phi^2\, k_f^{-1} (v_f(t) - \partial_t u_s(t)) \cdot w_f \,\mathrm{d}x - \int_\Omega \theta(t)\, v_f(t) \cdot w_f \,\mathrm{d}x \\ \qquad\qquad - \displaystyle\int_\Omega p(t) \,\mathrm{div}\,(\phi\, w_f)\,\mathrm{d}x = \int_\Omega \rho_f \phi f(t) \cdot w_f \,\mathrm{d}x, \qquad (1.47b) \\ \dfrac{\mathrm{d}}{\mathrm{d}t} \displaystyle\int_\Omega \frac{\alpha-\phi}{\kappa} p(t) q \,\mathrm{d}x + \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \mathrm{div}\left((\alpha-\phi)\, u_s(t)\right) q \,\mathrm{d}x \\ \qquad\qquad + \displaystyle\int_\Omega \mathrm{div}\,(\phi\, v_f(t)) q \,\mathrm{d}x = \int_\Omega \frac{\theta(t)}{\rho_f} q \,\mathrm{d}x. \qquad (1.47c) \end{cases}$$

This variational solution is unique, and coincides with the mild solution. Furthermore, the energy estimate (1.11) *with* $\eta = 0$ *holds true.*

Remark 1.19. Theorem 1.18 sheds light on the influence of solid viscosity on the model. Since $\eta = 0$, $\partial_t u_s$ does not belong to $\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$ but only to $\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$. For this reason, equations (1.37c) and (1.47c) are not similar because, when $\eta = 0$, the term $\mathrm{div}\left((1-\phi)\,\partial_t u_s(t)\right)$ is not in $\mathrm{L}^2(\Omega)$ in the space variable. One has only $\mathrm{div}\left((1-\phi)\,\partial_t u_s(t)\right) \in \mathrm{C}^0([0,T];\mathrm{H}^{-1}(\Omega))$. This confirms that viscoelastic effects have an impact on the regularity of the solution, as it was already observed for other linear or non-linear poroelastic models [Showalter, 2000; Bociu et al., 2016; Verri et al., 2018].

Proof. We follow the same steps as for the proof of Theorem 1.10. The input data are approximated by regular functions, *a priori* estimates are established for the approximated solutions and we pass to the limit on the variational formulation $(VF)^n$ with $\eta = 0$.

The estimate (1.39) still holds true even if $\eta = 0$ because $z^n \in \mathrm{C}^0([0,T];D(A_0^\kappa)) \subset \mathrm{C}^0([0,T];Y)$, in particular $v_s^n \in \mathrm{C}^0([0,T];[\mathrm{H}^1_0(\Omega)]^d)$, which justifies that z^n is regular enough to reproduce the formal calculations made in Section 1.1. As previously, this estimate implies that z^n is a Cauchy sequence in $\mathrm{C}^0([0,T];Z)$. However, since $\eta = 0$, estimate (1.39) only implies that v_f^n is a Cauchy sequence in $\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$. Hence, the convergence

$$v_f^n \longrightarrow v_f \quad \text{in } \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d),$$

is still valid but now $v_s^n = \partial_t u_s^n$ does not converge in $\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$ but only in $\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$. This changes the way to pass to the limit on $(VF)^n$ and in particular in the first equation. Let ψ be an element of $\mathcal{D}(0,T)$. For any $d_s \in [\mathrm{L}^2(\Omega)]^d$, we consider the unique solution $\eta_s \in [\mathrm{H}^1_0(\Omega)]^d$ of $-\mathrm{div}\left(\sigma_s(\eta_s)\right) = d_s$ as a test function, so that

$$\begin{aligned} \int_0^T \int_\Omega \sigma_s(v_s^n(t)) : \varepsilon(\eta_s)\, \psi(t) \,\mathrm{d}x\,\mathrm{d}t &= \int_0^T \int_\Omega \varepsilon(v_s^n(t)) : \sigma_s(\eta_s)\, \psi(t) \,\mathrm{d}x\,\mathrm{d}t \\ &= \int_0^T \int_\Omega v_s^n(t) \cdot d_s\, \psi(t) \,\mathrm{d}x\,\mathrm{d}t \underset{n\to\infty}{\longrightarrow} \int_0^T \int_\Omega v_s(t) \cdot d_s\, \psi(t) \,\mathrm{d}x\,\mathrm{d}t. \end{aligned}$$

This proves that for all $d_s \in [\mathrm{L}^2(\Omega)]^d$, we have

$$\frac{\mathrm{d}}{\mathrm{dt}} \int_\Omega u_s(t) \cdot d_s \,\mathrm{d}x - \int_\Omega v_s(t) \cdot d_s \,\mathrm{d}x = 0, \tag{1.48}$$

and we recover that $v_s = \partial_t u_s$ in $\mathcal{D}'((0,T)\times\Omega)$. In particular, it holds that $\partial_t u_s \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$.

We can obtain (1.47a) and (1.47b) in a similar way as for the viscous case. Finally, to get (1.47c), we observe that

$$\begin{aligned}\int_0^T \int_\Omega \mathrm{div}\left((\alpha-\phi)\, v_s^n(t)\right) q\, \psi(t) \,\mathrm{d}x \,\mathrm{d}t &= -\int_0^T \int_\Omega \mathrm{div}\left((\alpha-\phi)\, u_s^n(t)\right) q\, \dot\psi(t) \,\mathrm{d}x \,\mathrm{d}t \\ &\underset{n\to\infty}{\longrightarrow} -\int_0^T \int_\Omega \mathrm{div}\left((\alpha-\phi)\, u_s(t)\right) q\, \dot\psi(t) \,\mathrm{d}x \,\mathrm{d}t,\end{aligned}$$

where we have integrated by parts in time and then used that u_s^n converges in $\mathrm{C}^0([0,T];[\mathrm{H}^1_0(\Omega)]^d)$.

In the same way as for the viscous case, we notice that the weak formulation (1.47) provides some regularity on the time derivative of the solution. For instance, the fluid equation (1.47b) implies that for any $\psi \in \mathcal{D}(0,T)$ and $w_f \in [\mathrm{H}^1_0(\Omega)]^d$, we have

$$-\int_0^T \int_\Omega \rho_f \phi\, v_f(t) \cdot w_f\, \dot\psi(t) \,\mathrm{d}x \,\mathrm{d}t = \int_0^T \int_\Omega F(t) \cdot w_f\, \psi(t) \,\mathrm{d}x \,\mathrm{d}t,$$

with $F = \rho_f \phi\, f + \mathrm{div}\left(\phi\, \sigma_f(v_f)\right) - \phi^2\, k_f^{-1}(v_f - v_s) + \theta\, v_f - \phi\, \nabla p \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d)$. Since functions in $[\mathrm{H}^1_0(\Omega)]^d \otimes \mathcal{D}(0,T)$ generate $\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$, it follows that $\partial_t v_f \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d)$ and that, for any test function $w_f \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$,

$$\begin{aligned}\int_0^T \Big\langle \rho_f \phi\, \partial_t v_f, w_f \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}^1_0(\Omega)]^d} \mathrm{d}t + \int_0^T \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(w_f) \,\mathrm{d}x \,\mathrm{d}t \\ + \int_0^T \int_\Omega \phi^2\, k_f^{-1}(v_f - \partial_t u_s) \cdot w_f \,\mathrm{d}x \,\mathrm{d}t - \int_0^T \int_\Omega \theta\, v_f \cdot w_f \,\mathrm{d}x \,\mathrm{d}t \\ - \int_0^T \int_\Omega p\, \mathrm{div}\,(\phi\, w_f) \,\mathrm{d}x \,\mathrm{d}t = \int_0^T \int_\Omega \rho_f \phi f \cdot w_f \,\mathrm{d}x \,\mathrm{d}t.\end{aligned} \tag{1.49}$$

Similarly, we infer from (1.47a) and (1.47c) that $\partial_{tt}^2 u_s \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d)$ and $\frac{\alpha-\phi}{\kappa}\partial_t p + \mathrm{div}\left((\alpha-\phi)\partial_t u_s\right) \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$, and that for any $w_s \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$

$$\begin{aligned}\int_0^T \Big\langle \rho_s(1-\phi)\, \partial_{tt}^2 u_s, w_s \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}^1_0(\Omega)]^d} \mathrm{d}t + \int_0^T \int_\Omega \sigma_s(u_s) : \varepsilon(w_s) \,\mathrm{d}x \,\mathrm{d}t \\ - \int_0^T \int_\Omega \phi^2\, k_f^{-1}(v_f - \partial_t u_s) \cdot w_s \,\mathrm{d}x \,\mathrm{d}t - \int_0^T \int_\Omega p\, \mathrm{div}\left((\alpha-\phi)\, w_s\right) \mathrm{d}x \,\mathrm{d}t \\ = \int_0^T \int_\Omega \rho_s(1-\phi) f \cdot w_s \,\mathrm{d}x \,\mathrm{d}t,\end{aligned} \tag{1.50}$$

and for any $q \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$

$$\int_0^T \int_\Omega \Big(\frac{\alpha-\phi}{\kappa}\partial_t p + \mathrm{div}\left((\alpha-\phi)\partial_t u_s\right)\Big) q \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \int_\Omega \mathrm{div}\,(\phi\, v_f)\, q \,\mathrm{d}x \,\mathrm{d}t = 0. \tag{1.51}$$

Note that the main difference compared to the viscous case is that $\partial_t p$ is not in $\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$ any more since the structure velocity $\partial_t u_s$ does not belong to $\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$. Yet $\partial_t\Big(\frac{\alpha-\phi}{\kappa}p + \mathrm{div}\left((\alpha-\phi)u_s\right)\Big) \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$ and $\partial_t p \in \mathrm{L}^2(0,T;\mathrm{H}^{-1}(\Omega))$.

Let us now prove that the weak solution is unique. Let (u_s, v_f, p) be a solution to (1.47) with zero initial conditions and source terms, and let τ be given in $(0,T)$. Contrary to the viscous case, we cannot take $w_s = \partial_t u_s$ as test function in (1.50) because $\partial_t u_s \notin \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$. To overcome this lack of regularity, we consider the so-called Ladyzhenskaya test functions [Ladyženskaja et al., 1968]. For the solid and pressure equations, we use the same test functions that were considered in [Barucq et al., 2004, Theorem 3] and [Saint-Macary, 2004, Section 4.2.2.] for Biot's consolidation model, namely

$$\psi_s(t) = \begin{cases} -\displaystyle\int_t^\tau u_s(\sigma)\,\mathrm{d}\sigma & \text{if } \tau \geq t \\ 0 \quad \text{if } \tau \leq t \end{cases} \quad \text{and } \psi_p(t) = \begin{cases} -\displaystyle\int_t^\tau \int_0^v p(\sigma)\,\mathrm{d}\sigma\,\mathrm{d}v & \text{if } \tau \geq t \\ 0 \quad \text{if } \tau \leq t. \end{cases}$$

To these test functions, we have to add a fluid test function which is built in the very same manner and corresponds to the fluid counterpart of the previous structure and pressure test functions. Therefore we consider

$$\psi_f(t) = \begin{cases} -\displaystyle\int_t^\tau \int_0^v v_f(\sigma)\,\mathrm{d}\sigma\,\mathrm{d}v & \text{if } \tau \geq t \\ 0 \quad \text{if } \tau \leq t. \end{cases}$$

The functions ψ_s belongs to $\mathrm{C}^0([0,T];[\mathrm{H}^1_0(\Omega)]^d)$, ψ_f belongs to $\mathrm{C}^1([0,T];[\mathrm{H}^1_0(\Omega)]^d)$ and ψ_p belongs to $\mathrm{C}^1([0,T];\mathrm{L}^2(\Omega))$ and they are admissible test functions. For $t \leq \tau$, remembering that the considered solution is associated with zero initial conditions, they satisfy

$$\psi_s(\tau) = 0, \qquad \partial_t \psi_s(t) = u_s(t), \qquad \partial_t \psi_s(0) = 0. \tag{1.52}$$

$$\psi_p(\tau) = 0, \qquad \partial_t \psi_p(t) = \int_0^t p(\sigma)\,\mathrm{d}\sigma, \qquad \partial^2_{tt} \psi_p(t) = p(t), \qquad \partial^2_{tt} \psi_p(0) = 0, \tag{1.53}$$

and

$$\psi_f(\tau) = 0, \qquad \partial_t \psi_f(t) = \int_0^t v_f(\sigma)\,\mathrm{d}\sigma, \quad \partial^2_{tt} \psi_f(t) = v_f(t), \quad \partial^2_{tt} \psi_f(0) = 0. \tag{1.54}$$

Taking ψ_s as a test function in (1.50) we compute the different terms. Due to (1.52), we have in a standard way (see [Lions and Magenes, 1972] in the case of an abstract second order equation or [Saint-Macary, 2004] for the Biot's consolidation model)

$$\int_0^\tau \left\langle \rho_s(1-\phi)\,\partial^2_{tt} u_s, \psi_s \right\rangle_{[\mathrm{H}^{-1}(\Omega)]^d,[\mathrm{H}^1_0(\Omega)]^d} \mathrm{d}t = -\frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|u_s(\tau)|^2\,\mathrm{d}x,$$

$$\int_0^\tau \int_\Omega \sigma_s(u_s) : \varepsilon(\psi_s)\,\mathrm{d}x\,\mathrm{d}t = -\frac{1}{2}\int_\Omega \sigma_s(\psi_s(0)) : \varepsilon(\psi_s(0))\,\mathrm{d}x,$$

Moreover, since $v_f(t) = \partial^2_{tt}\psi_f(t)$, $\partial_t u_s(t) = \partial^2_{tt}\psi_s(t)$, $\partial_t \psi_f(0) = \partial_t \psi_s(0) = 0$ and $\psi_s(\tau) = 0$, the friction term writes, after integration by parts in time,

$$-\int_0^\tau \int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot \psi_s\,\mathrm{d}x\,\mathrm{d}t = \int_0^\tau \int_\Omega \phi^2 k_f^{-1}(\partial_t \psi_f - \partial_t \psi_s)\cdot \partial_t \psi_s\,\mathrm{d}x\,\mathrm{d}t.$$

Finally we obtain the following identity

$$\begin{aligned} -\frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|u_s(\tau)|^2\,\mathrm{d}x - \frac{1}{2}\int_\Omega \sigma_s(\psi_s(0)) : \varepsilon(\psi_s(0))\,\mathrm{d}x \\ + \int_0^\tau \int_\Omega \phi^2 k_f^{-1}(\partial_t \psi_f - \partial_t \psi_s)\cdot \partial_t \psi_s\,\mathrm{d}x\,\mathrm{d}t \\ - \int_0^\tau \int_\Omega p\,\mathrm{div}\left((\alpha-\phi)\psi_s\right)\mathrm{d}x\,\mathrm{d}t = 0. \end{aligned} \tag{1.55}$$

Let us now focus on the fluid equation. We take ψ_f as a test fonction in (1.49). Due to the properties (1.54), the fluid inertial and viscous terms become respectively

$$\int_0^\tau \big\langle \rho_f \phi\, \partial_t v_f, \psi_f \big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d,[\mathrm{H}^1_0(\Omega)]^d} \,\mathrm{d}t = -\int_0^\tau \int_\Omega \rho_f \phi\, v_f \cdot \partial_t \psi_f \,\mathrm{d}x\,\mathrm{d}t$$
$$= -\frac{1}{2}\int_\Omega \rho_f \phi\, |\partial_t \psi_f(\tau)|^2 \,\mathrm{d}x,$$

and

$$\int_0^\tau \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(\psi_f)\,\mathrm{d}x\,\mathrm{d}t = -\int_0^\tau \int_\Omega \phi\, \sigma_f(\partial_t \psi_f) : \varepsilon(\partial_t \psi_f)\,\mathrm{d}x\,\mathrm{d}t.$$

Once again the friction term can be transformed as follows

$$\int_0^\tau \int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot \psi_f \,\mathrm{d}x\,\mathrm{d}t = -\int_0^\tau \int_\Omega \phi^2 k_f^{-1}(\partial_t \psi_f - \partial_t \psi_s)\cdot \partial_t \psi_f \,\mathrm{d}x\,\mathrm{d}t,$$

and, thanks to (1.54), (1.53), an integration by parts in time in the pressure term yields

$$-\int_0^\tau \int_\Omega p\,\mathrm{div}\,(\phi\,\psi_f)\,\mathrm{d}x\,\mathrm{d}t = \int_0^\tau \int_\Omega \partial_t \psi_p\,\mathrm{div}\,(\phi\,\partial_t \psi_f)\,\mathrm{d}x\,\mathrm{d}t.$$

The last term, involving θ, writes

$$-\int_0^\tau \int_\Omega \theta\, v_f \cdot \psi_f \,\mathrm{d}x\,\mathrm{d}t = \int_0^\tau \int_\Omega \theta\, |\partial_t \psi_f|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^\tau \int_\Omega \partial_t \theta\, \psi_f \cdot \partial_t \psi_f \,\mathrm{d}x\,\mathrm{d}t.$$

Summing up all these contributions implies

$$-\frac{1}{2}\int_\Omega \rho_f \phi\, |\partial_t \psi_f(\tau)|^2 \,\mathrm{d}x - \int_0^\tau \int_\Omega \phi\, \sigma_f(\partial_t \psi_f) : \varepsilon(\partial_t \psi_f)\,\mathrm{d}x\,\mathrm{d}t$$
$$-\int_0^\tau \int_\Omega \phi^2 k_f^{-1}(\partial_t \psi_f - \partial_t \psi_s)\cdot \partial_t \psi_f \,\mathrm{d}x\,\mathrm{d}t + \int_0^\tau \int_\Omega \partial_t \psi_p\,\mathrm{div}\,(\phi\,\partial_t \psi_f)\,\mathrm{d}x\,\mathrm{d}t$$
$$= -\int_0^\tau \int_\Omega \theta\, |\partial_t \psi_f|^2 \,\mathrm{d}x\,\mathrm{d}t - \int_0^\tau \int_\Omega \partial_t \theta\, \psi_f \cdot \partial_t \psi_f \,\mathrm{d}x\,\mathrm{d}t. \quad (1.56)$$

Next we take ψ_p as a test function in (1.51). As in [Saint-Macary, 2004] we have

$$\int_0^\tau \int_\Omega \Big(\frac{\alpha-\phi}{\kappa}\partial_t p + \mathrm{div}\,\big((\alpha-\phi)\partial_t u_s\big)\Big)\psi_p \,\mathrm{d}x\,\mathrm{d}t$$
$$= -\frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}\,|\partial_t \psi_p(\tau)|^2 \,\mathrm{d}x + \int_0^\tau \int_\Omega p\,\mathrm{div}\,\big((\alpha-\phi)\psi_s\big)\,\mathrm{d}x\,\mathrm{d}t.$$

Moreover

$$\int_0^\tau \int_\Omega \mathrm{div}\,(\phi\, v_f)\psi_p \,\mathrm{d}x\,\mathrm{d}t = -\int_0^\tau \int_\Omega \mathrm{div}\,(\phi\,\partial_t \psi_f)\partial_t \psi_p \,\mathrm{d}x\,\mathrm{d}t,$$

and thus we obtain the following identity

$$-\frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}\,|\partial_t \psi_p(\tau)|^2 \,\mathrm{d}x + \int_0^\tau \int_\Omega p\,\mathrm{div}\,\big((\alpha-\phi)\psi_s\big)\,\mathrm{d}x\,\mathrm{d}t$$
$$-\int_0^\tau \int_\Omega \mathrm{div}\,(\phi\,\partial_t \psi_f)\partial_t \psi_p \,\mathrm{d}x\,\mathrm{d}t = 0. \quad (1.57)$$

Summing up (1.55), (1.56) and (1.57), we obtain

$$\frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|u_s(\tau)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,|\partial_t\psi_f(\tau)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}\,|\partial_t\psi_p(\tau)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \sigma_s(\psi_s(0)) : \varepsilon(\psi_s(0))\,\mathrm{d}x + \int_0^\tau\int_\Omega \phi\,\sigma_f(\partial_t\psi_f) : \varepsilon(\partial_t\psi_f)\,\mathrm{d}x\,\mathrm{d}t + \int_0^\tau\int_\Omega \phi^2 k_f^{-1}(\partial_t\psi_f - \partial_t\psi_s)^2\,\mathrm{d}x\,\mathrm{d}t = \int_0^\tau\int_\Omega \theta\,|\partial_t\psi_f|^2\,\mathrm{d}x\,\mathrm{d}t + \int_0^\tau\int_\Omega \partial_t\theta\,\psi_f\cdot\partial_t\psi_f\,\mathrm{d}x\,\mathrm{d}t. \quad (1.58)$$

To conclude, we observe that

$$\int_0^\tau\int_\Omega \theta\,|\partial_t\psi_f|^2\,\mathrm{d}x\,\mathrm{d}t \le (\rho_f\phi_{\min})^{-1}\|\theta\|_{L^\infty((0,T)\times\Omega)}\int_0^\tau\int_\Omega \rho_f\phi\,|\partial_t\psi_f|^2\,\mathrm{d}x\,\mathrm{d}t$$

and that, since $\psi_f(t) = -\int_t^\tau \partial_t\psi_f(\sigma)\mathrm{d}\sigma$, we can estimate the last term of (1.58) as follows

$$\begin{aligned}
&\int_0^\tau\int_\Omega \partial_t\theta\,\psi_f\cdot\partial_t\psi_f\,\mathrm{d}x\,\mathrm{d}t\\
&\quad\le \|\partial_t\theta\|_{L^\infty((0,T)\times\Omega)}\int_0^\tau\int_\Omega \left|\int_t^\tau \partial_t\psi_f(\sigma)\,\mathrm{d}\sigma\right|\,|\partial_t\psi_f(t)|\,\mathrm{d}x\,\mathrm{d}t\\
&\quad\le \|\partial_t\theta\|_{L^\infty((0,T)\times\Omega)}\int_\Omega\int_0^\tau\int_0^\tau |\partial_t\psi_f(\sigma)|\,|\partial_t\psi_f(t)|\,\mathrm{d}\sigma\,\mathrm{d}t\,\mathrm{d}x\\
&\quad\le T(\rho_f\phi_{\min})^{-1}\|\partial_t\theta\|_{L^\infty((0,T)\times\Omega)}\int_0^\tau\int_\Omega \rho_f\phi\,|\partial_t\psi_f|^2\,\mathrm{d}x\,\mathrm{d}t.
\end{aligned}$$

Consequently, using Grönwall Lemma, we deduce that $u_s = \partial_t\psi_f = \partial_t\psi_p = 0$. Hence $u_s = v_f = p = 0$, which proves the uniqueness of the variational solution.

Now that we know that the variational solution is unique, it follows that it is necessarily the one obtained by the approximation process built from $(VF)^n$. Since this approximation process is based on the energy estimate (1.39) with $\eta = 0$, we can pass to the limit in this estimation to get (1.11).

In particular, to show that the mild solution is equal to the variational solution, it is sufficient to prove that it also derives from this approximation process. Let us denote by z the mild solution given by Theorem 1.17, and remind the notation $G^n(t)(u_s, v_s, v_f, p) = \big(0, 0, -(\rho_f\phi)^{-1}\theta^n(t)v_f\big)$, with $\theta^n \in \mathrm{C}^1([0,T];\mathrm{L}^\infty(\Omega))$ converging towards θ in $\mathrm{C}^0([0,T]\times\Omega)$. From Duhamel formula, it holds that

$$z(t) - z^n(t) = \Phi_0^\kappa(t)(z_0 - z_0^n) + \int_0^t \Phi_0^\kappa(t-s)\big(g(s) - g^n(s)\big)\,\mathrm{d}s - \int_0^t \Phi_0^\kappa(t-s)\big(G(s)z(s) - G^n(s)z(s)\big)\,\mathrm{d}s.$$

Writing $G(s)z(s) - G^n(s)z^n(s) = (G(s) - G^n(s))\,z^n(s) + G(s)\,(z(s) - z^n(s))$ and recalling that Φ_0^κ is a C^0-semigroup of contraction, we infer

$$\|z(t) - z^n(t)\|_Z \le \|z_0 - z_0^n\|_Z + \int_0^t \|g(s) - g^n(s)\|_Z\,\mathrm{d}s + (\rho_f\phi_{\min})^{-1}\|\theta - \theta^n\|_{\mathrm{C}^0([0,T]\times\Omega)}\int_0^t \|z^n(s)\|_Z\,\mathrm{d}s + \omega\int_0^t \|z(s) - z^n(s)\|_Z\,\mathrm{d}s,$$

where ω is defined in (1.35). Thus, for any $\delta > 0$, we can find n large enough such that

$$\|z(t) - z^n(t)\|_Z \leq \delta + \omega \int_0^t \|z(s) - z^n(s)\|_Z \ \mathrm{d}s.$$

Using Grönwall Lemma, we conclude that $\|z(t) - z^n(t)\|_Z \leq \delta e^{\omega T}$, and hence $z^n \longrightarrow z$ in $\mathrm{C}^0([0,T]; Z)$. □

Remark 1.20. In the previous proof, we took advantage of the semigroup framework to show the existence of the variational solution. Note that it could also be shown by regularization of the viscous case, see [Barucq et al., 2004, Theorem 2] where such a regularization is performed on Biot's consolidation model.

Remark 1.21. As in the viscous case, we could also define the variational solution without assuming that it is continuous with respect to time, but rather by seeking for

$$\begin{aligned} u_s \in \mathrm{L}^\infty(0,T;[\mathrm{H}^1_0(\Omega)]^d),\ \partial_t u_s \in \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d) \text{ and } \partial^2_{tt} u_s \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d),\\ v_f \in \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d) \cap \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d) \text{ and } \partial_t v_f \in \mathrm{L}^2(0,T;[\mathrm{H}^{-1}(\Omega)]^d),\\ p \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega)) \text{ and } \partial_t\Big(\frac{\alpha-\phi}{\kappa} p + \operatorname{div}\big((\alpha-\phi)u_s\big)\Big) \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega)), \end{aligned}$$

such that (1.46) and (1.47) are verified. With this definition, the continuity in time of the solution can then be recovered using, for instance, a parabolic regularization, while it is obtained directly in the above proof.

In the next section, we analyze the poromechanics problem for an incompressible elastic skeleton, modeled by the assumption $\kappa = +\infty$. This assumption is crucial for targeting biomedical applications since the tissues in our body are mostly composed of water, and thus are close to being incompressible.

1.3 Existence of solutions for an incompressible skeleton $\kappa = +\infty$

When $\kappa = +\infty$ – and thus $\alpha = 1$, see the thesis introduction – the system of equations (1.18) reads

$$\begin{cases} \partial_t u_s - v_s = 0, & \text{(1.59a)}\\ \rho_s(1-\phi)\,\partial_t v_s - \operatorname{div}\big(\sigma_s^{\mathrm{vis}}(v_s)\big) - \operatorname{div}\big(\sigma_s(u_s)\big) & \\ \qquad\qquad - \phi^2\, k_f^{-1}(v_f - v_s) + (1-\phi)\,\nabla p = \rho_s(1-\phi)\, f, & \text{(1.59b)}\\ \rho_f \phi\,\partial_t v_f - \operatorname{div}\big(\phi\,\sigma_f(v_f)\big) + \phi^2\, k_f^{-1}(v_f - v_s) - \theta\, v_f + \phi\,\nabla p = \rho_f \phi\, f, & \text{(1.59c)}\\ \operatorname{div}\big((1-\phi)\, v_s + \phi\, v_f\big) = \dfrac{\theta}{\rho_f}. & \text{(1.59d)} \end{cases}$$

It has to be completed with boundary conditions (1.3) and initial conditions (1.4). Note that, in the present case, there is no initial condition for the pressure anymore.

Equation (1.59d) traduces the mixture's incompressibility, which comes from the assumption that the solid and the fluid phases are both incompressible. It takes the form of a constraint on the divergence of the mixture's velocity.

But, as already noticed in Section 1.1, it is sufficient to consider the case

$$\operatorname{div}\big((1-\phi)\, v_s + \phi\, v_f\big) = 0, \ \text{ in } \Omega \times (0,T). \tag{1.60}$$

Indeed, assuming for instance that $\theta \in \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega)) \cap \mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ and that the compatibility condition $\int_\Omega \theta(t)\,\mathrm{d}x = 0$ is satisfied for all $t \in [0,T]$, we consider the Bogovskii's operator [Bogovskii, 1979] and we build $v_\theta \in \mathrm{H}^1(0,T;[\mathrm{H}^1_0(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}^2(\Omega)]^d)$ such that

$$\operatorname{div} v_\theta = \frac{\theta}{\rho_f}. \tag{1.61}$$

The change of variables $\hat{v}_s = v_s - v_\theta$ and $\hat{v}_f = v_f - v_\theta$ gives $\operatorname{div}\big((1-\phi)\,\hat{v}_s + \phi\,\hat{v}_f\big) = 0$ by construction. Furthermore, $\Big(u_s - \int_0^t v_\theta(s)\,\mathrm{d}s, \hat{v}_s, \hat{v}_f, p\Big)$ verifies (1.59a), (1.59b) and (1.59c) with right-hand sides that are different but still regular since $v_\theta \in \mathrm{H}^1(0,T;[\mathrm{H}^1_0(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}^2(\Omega)]^d)$.

The first part of this section is devoted to the functional analysis of the coupling constraint (1.60).

1.3.1 Functional framework

We consider the space

$$V_\phi = \Big\{(v_s, v_f) \in [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \,:\, \operatorname{div}\big((1-\phi)\,v_s + \phi\,v_f\big) = 0 \quad \text{in}\,\Omega\Big\}$$

of functions in $[\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d$ satisfying the constraint (1.60). Let us also define the space H_ϕ as the closure of V_ϕ in $[\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d$.

Then, we introduce the mixture's divergence operator defined by

$$\begin{array}{rccc} B: & [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d & \longrightarrow & \mathrm{L}^2_0(\Omega) \\ & (v_s, v_f) & \longmapsto & \operatorname{div}\big((1-\phi)\,v_s + \phi\,v_f\big). \end{array}$$

The bounded operator B satisfies the following inf-sup condition.

Proposition 1.22. *Assume that $\phi \in \mathrm{H}^{d/2+r}(\Omega)$ with $r > 0$. There exists $\underline{\beta} > 0$ such that, for all $p \in \mathrm{L}^2_0(\Omega)$,*

$$\sup_{(v_s,v_f)\in[\mathrm{H}^1_0(\Omega)]^d\times[\mathrm{H}^1_0(\Omega)]^d} \frac{\displaystyle\int_\Omega \operatorname{div}\big((1-\phi)\,v_s + \phi\,v_f\big)\,p\,\mathrm{d}x}{\|(v_s,v_f)\|_{[\mathrm{H}^1_0(\Omega)]^d\times[\mathrm{H}^1_0(\Omega)]^d}} \geq \underline{\beta}\|p\|. \tag{1.62}$$

Proof. There exists $C_{\mathrm{div}} > 0$ such that for any $p \in \mathrm{L}^2_0(\Omega)$, there exists $v_p \in [\mathrm{H}^1_0(\Omega)]^d$ satisfying

$$\operatorname{div} v_p = p \quad \text{and} \quad \|\nabla v_p\| \leq C_{\mathrm{div}}\|p\|. \tag{1.63}$$

Setting $v = (v_p, v_p)$, we have $Bv = p$ by construction and $\|v\|_{[\mathrm{H}^1_0(\Omega)]^d\times[\mathrm{H}^1_0(\Omega)]^d} \leq C\|p\|$ from the above inequality. □

Remark 1.23. Note that the constant $\underline{\beta}$ of the above inf-sup condition does not depend on the porosity ϕ and is also valid for $\phi = 0$ (resp. $\phi = 1$) which are the limit cases for which there is no fluid (resp. no structure).

This inf-sup condition allows us to state the following result, which is a generalization of De Rham Theorem [Temam, 2001; Girault and Raviart, 1986; Boyer and Fabrie, 2012]. It is a key ingredient to show the existence of pressure in the incompressible case.

Theorem 1.24. *Assume that $\phi \in \mathrm{H}^{d/2+r}(\Omega)$ with $r > 0$. If $f = (f_s, f_f) \in [\mathrm{H}^{-1}(\Omega)]^d \times [\mathrm{H}^{-1}(\Omega)]^d$ satisfies*

$$\langle f, w\rangle = \langle f_s, w_s\rangle_{[\mathrm{H}^{-1}(\Omega)]^d,[\mathrm{H}^1_0(\Omega)]^d} + \langle f_f, w_f\rangle_{[\mathrm{H}^{-1}(\Omega)]^d,[\mathrm{H}^1_0(\Omega)]^d} = 0, \quad \forall w = (w_s, w_f) \in V_\phi,$$

then there exists a unique $p \in \mathrm{L}^2_0(\Omega)$ such that $f_s = -(1-\phi)\,\nabla p$ and $f_f = -\phi\,\nabla p$.

Proof. The proof follows standard arguments and is based on the Closed Range Theorem. Note that $V_\phi = \mathrm{Ker} B$. Let us characterize the adjoint of B. Since $\phi \in \mathrm{H}^{d/2+r}(\Omega)$ with $r > 0$, ϕ is a multiplier of $[\mathrm{H}^1_0(\Omega)]^d$, namely

$$\forall v \in [\mathrm{H}^1_0(\Omega)]^d, \quad \phi\, v \in [\mathrm{H}^1_0(\Omega)]^d \quad \text{and} \quad \|\nabla(\phi\, v)\| \leq C_\phi \|\nabla v\|$$

for some positive constant C_ϕ. Therefore, for all $p \in \mathrm{L}^2_0(\Omega)$, $(1-\phi)\,\nabla p$ and $\phi\,\nabla p$ belong to $[\mathrm{H}^{-1}(\Omega)]^d$ so that we can define the adjoint operator as

$$\begin{array}{rccc} B^* : & \mathrm{L}^2_0(\Omega) & \longrightarrow & [\mathrm{H}^{-1}(\Omega)]^d \times [\mathrm{H}^{-1}(\Omega)]^d \\ & p & \longmapsto & \big(-(1-\phi)\,\nabla p, -\phi\,\nabla p\big). \end{array}$$

Thanks to Proposition 1.22, the Closed Range Theorem implies that $(\mathrm{Ker} B)^\circ = \mathrm{Rg}\, B^*$. Consequently, for any $f \in (\mathrm{Ker} B)^\circ = (V_\phi)^\circ$, namely for any $f = (f_s, f_f) \in [\mathrm{H}^{-1}(\Omega)]^d \times [\mathrm{H}^{-1}(\Omega)]^d$ satisfying

$$\langle f, w \rangle = 0, \quad \forall w \in V_\phi,$$

there exists a unique $p \in \mathrm{L}^2_0(\Omega)$ such that $f_s = -(1-\phi)\,\nabla p$ and $f_f = -\phi\,\nabla p$. □

Theorem 1.24 allows us to characterize the space H_ϕ in the following way.

Proposition 1.25. *The space H_ϕ can be expressed as*

$$H_\phi = \big\{(v_s, v_f) \in [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \,:\, \mathrm{div}\,\big((1-\phi)\,v_s + \phi\, v_f\big) = 0 \quad \text{in}\, \mathcal{D}'(\Omega) \\ \textit{and}\ \big((1-\phi)\,v_s + \phi\, v_f\big)\cdot n = 0 \quad \text{on}\, \partial\Omega\big\}.$$

Proof. We denote by $\mathcal{H}$ the space

$$\big\{(v_s, v_f) \in [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \,:\, \mathrm{div}\,\big((1-\phi)\,v_s + \phi\, v_f\big) = 0 \quad \text{in}\, \mathcal{D}'(\Omega) \\ \text{and}\ \big((1-\phi)\,v_s + \phi\, v_f\big)\cdot n = 0 \quad \text{on}\, \partial\Omega\big\}.$$

Let $v = (v_s, v_f)$ be an element of H_ϕ. By definition, H_ϕ is the closure of V_ϕ in $[\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d$, so there exists a sequence (v_s^n, v_f^n) belonging to V_ϕ that converges towards v in $[\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d$. Since $\mathrm{div}\,\big((1-\phi)\,v_s^n + \phi\, v_f^n\big) = 0$, the equality $\mathrm{div}\,\big((1-\phi)\,v_s + \phi\, v_f\big) = 0$ holds true in the limit. Further, $(1-\phi)\,v_s + \phi\, v_f \in \mathrm{H}_{\mathrm{div}}(\Omega)$. The continuity of the normal trace operator then implies that

$$\big\|\big((1-\phi)\,v_s + \phi\, v_f\big)\cdot n - \big((1-\phi)\,v_s^n + \phi\, v_f^n\big)\cdot n\big\|_{H^{-1/2}(\partial\Omega)} \underset{n\to\infty}{\longrightarrow} 0,$$

which implies that $\big((1-\phi)\,v_s + \phi\, v_f\big)\cdot n = 0$, since $\big((1-\phi)\,v_s^n + \phi\, v_f^n\big)\cdot n = 0$. Hence, $H_\phi \subset \mathcal{H}$.

Now, let us prove the other inclusion. Let denote by $\mathcal{H}^*$ the orthogonal complement of H_ϕ into $\mathcal{H}$ and let $f = (f_s, f_f)$ be an element of $\mathcal{H}^*$. Noting that $\mathcal{H}^* \subset H_\phi^\perp$ and $V_\phi \subset H_\phi$, it follows from Theorem 1.24 that there exists a pressure $p \in \mathrm{L}^2_0(\Omega)$ such that $f_s = -(1-\phi)\,\nabla p$ and $f_f = -\phi\,\nabla p$. Moreover since $\nabla p = -(f_s + f_f)$ and $(f_s, f_f) \in [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d$, we get $p \in \mathrm{H}^1(\Omega)$. Since f belongs to $\mathcal{H}$, we have $\mathrm{div}\,\big(((1-\phi)^2 + \phi^2)\nabla p\big) = \mathrm{div}\,\big((1-\phi)^2\,\nabla p + \phi^2\,\nabla p\big) = 0$ in $\mathcal{D}'(\Omega)$ and $\big((1-\phi)^2\,\nabla p + \phi^2\,\nabla p\big)\cdot n = 0$. Thus p is equal to zero (up to a constant) as the unique solution of an elliptic Neumann problem, so $\nabla p = 0$ and $f = 0$. To conclude, $\mathcal{H}^* = \{0\}$, which proves that $H_\phi = \mathcal{H}$. □

Remark 1.26. If $\phi \in \mathrm{C}^\infty(\Omega)$, one can show that V_ϕ and H_ϕ correspond to the closures of the space $\mathcal{V}_\phi$ in $[\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d$ and $[\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d$, where

$$\mathcal{V}_\phi = \big\{(v_s, v_f) \in [\mathcal{D}(\Omega)]^d \times [\mathcal{D}(\Omega)]^d \,:\, \mathrm{div}\,\big((1-\phi)\,v_s + \phi\, v_f\big) = 0\big\}.$$

We are now going to combine this functional framework adapted to the constraint (1.60) with the semigroup approach in order to study Problem (1.59). Here again, we investigate the cases $\eta > 0$ and $\eta = 0$ separately. To simplify the proof we consider that θ, which appears now only in the term $-\theta\, v_f$ in the fluid equation, does not depend on time: $\theta \in \mathrm{L}^\infty(\Omega)$. It simplifies the proof, but it could be easily modified to include the time-dependent case as in the proofs of Section 1.2.

1.3.2 The case $\eta > 0$

We formulate the problem in the functional framework established previously. We seek for a solution $z = (u_s, v_s, v_f)$ in the energy space $H = [\mathrm{H}^1_0(\Omega)]^d \times H_\phi$ endowed with the scalar product

$$(z, y)_H = \int_\Omega \sigma_s(u_s) : \varepsilon(d_s) + \int_\Omega \rho_s(1-\phi)\, v_s \cdot w_s \,\mathrm{d}x + \int_\Omega \rho_f \phi\, v_f \cdot w_f \,\mathrm{d}x,$$

for any $z = (u_s, v_s, v_f)$, $y = (d_s, w_s, w_f)$ belonging to H, and with the corresponding norm

$$\|z\|_H^2 = \|u_s\|_{\mathrm{s}}^2 + \int_\Omega \rho_s(1-\phi)\,|v_s|^2 \,\mathrm{d}x + \int_\Omega \rho_f \phi\,|v_f|^2 \,\mathrm{d}x.$$

Setting $V = [\mathrm{H}^1_0(\Omega)]^d \times V_\phi$, we consider the bilinear form

$$\begin{aligned} a_\eta^\infty(z, y) = &- \int_\Omega \sigma_s(v_s) : \varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s) : \varepsilon(w_s)\,\mathrm{d}x \\ &+ 2\eta \int_\Omega \varepsilon(v_s) : \varepsilon(w_s)\,\mathrm{d}x + \int_\Omega \phi^2\, k_f^{-1}(v_f - v_s)\cdot(w_f - w_s)\,\mathrm{d}x \\ &+ \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(w_f)\,\mathrm{d}x - \int_\Omega \theta v_f \cdot w_f \,\mathrm{d}x, \end{aligned} \tag{1.64}$$

defined for all $z = (u_s, v_s, v_f)$ and $y = (d_s, w_s, w_f)$ in V. This bilinear form is the same as (1.22) but without the terms involving the pressure because of the test functions in V. Note that here, since we have assumed that θ does not depend on time, we can include the term $\int_\Omega \theta v_f \cdot w_f$ in the definition of the bilinear form associated with our coupled problem. When θ depends on time, one cannot and we have to introduce the operator $G(t)$ which is a bounded perturbation, see (1.25). As in (1.23) and (1.24), we define the unbounded operator $\left(A_\eta^\infty, D(A_\eta^\infty)\right)$ associated with the bilinear form (1.64) by

$$(A_\eta^\infty z, y)_H = a_\eta^\infty(z, y), \quad \forall z \in D(A_\eta^\infty), \forall y \in V,$$

in the domain

$$D(A_\eta^\infty) = \{z \in V : \exists g \in H,\, a_\eta^\infty(z, y) = (g, y)_H, \quad y \in V\}.$$

The above definitions are quite abstract, in particular because they rely on test functions in the constrained space V_ϕ. In the next proposition, we recover a more explicit expression of A_η^∞ and $D(A_\eta^\infty)$ thanks to the generalization of De Rham's Theorem established previously.

Proposition 1.27. *The operator's domain can be characterized as*

$$D(A_\eta^\infty) = \left\{ \begin{array}{l} u_s, v_s, v_f \in [\mathrm{H}^1_0(\Omega)]^d \text{ such that } \exists!\, p \in \mathrm{L}^2_0(\Omega) \text{ such that} \\ \quad -\operatorname{div}(\sigma_s(u_s)) - \operatorname{div}(\sigma_s^{vis}(v_s)) + (1-\phi)\,\nabla p \in [\mathrm{L}^2(\Omega)]^d, \\ \quad -\operatorname{div}(\phi\,\sigma_f(v_f)) + \phi\,\nabla p \in [\mathrm{L}^2(\Omega)]^d, \\ \quad \operatorname{div}\left((1-\phi)v_s + \phi\, v_f\right) = 0 \end{array} \right\}. \tag{1.65}$$

In addition, for all $z = (u_s, v_s, v_f) \in D(A_\eta^\infty)$ and $g = (g_u, g_s, g_f) \in H$, we have

$$A_\eta^\infty z = g \Leftrightarrow \exists!\, p \in \mathrm{L}^2_0(\Omega) \text{ such that,}$$

$$\left\{ \begin{array}{l} g_u = -v_s, \\ \rho_s(1-\phi)\, g_s = -\operatorname{div}\left(\sigma_s(u_s)\right) - \operatorname{div}\left(\sigma_s^{vis}(v_s)\right) \\ \qquad\qquad\qquad -\phi^2\, k_f^{-1}(v_f - v_s) + (1-\phi)\,\nabla p, \\ \rho_f \phi\, g_f = -\operatorname{div}\left(\phi\,\sigma_f(v_f)\right) + \phi^2\, k_f^{-1}(v_f - v_s) - \theta\, v_f + \phi\,\nabla p. \end{array} \right. \tag{1.66}$$

Proof. Let $z = (u_s, v_s, v_f)$ be an element of $D(A_\eta^\infty)$. By definition, there exists $g = (g_u, g_s, g_f) \in H$ such that $a_\eta^\infty(z, y) = (g, y)_H$ for all $y = (d_s, w_s, w_f) \in V$, namely

$$\int_\Omega \sigma_s(g_u) : \varepsilon(d_s)\,\mathrm{d}x = -\int_\Omega \sigma_s(v_s) : \varepsilon(d_s)\,\mathrm{d}x, \quad \forall d_s \in [\mathrm{H}_0^1(\Omega)]^d, \tag{1.67}$$

and

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\, g_s \cdot w_s\,\mathrm{d}x + \int_\Omega \rho_f \phi\, g_f \cdot w_f\,\mathrm{d}x &= \int_\Omega \sigma_s(u_s) : \varepsilon(w_s)\,\mathrm{d}x \\ &+ 2\eta \int_\Omega \varepsilon(v_s) : \varepsilon(w_s)\,\mathrm{d}x + \int_\Omega \phi^2\, k_f^{-1}(v_f - v_s)\cdot(w_f - w_s)\,\mathrm{d}x \\ &+ \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(w_f)\,\mathrm{d}x - \int_\Omega \theta v_f \cdot w_f\,\mathrm{d}x, \quad \forall (w_s, w_f) \in V_\phi. \end{aligned} \tag{1.68}$$

The relation (1.67) implies that $g_u = -v_s$ in $[\mathrm{H}_0^1(\Omega)]^d$. From (1.68), we deduce that

$$\begin{aligned}\forall (w_s, w_f) \in V_\phi, \quad \big((g_s, g_f), (w_s, w_f)\big)_{H_\phi} &= \\ \Big\langle -\mathrm{div}\,\big(\sigma_s(u_s)\big) - \mathrm{div}\,\big(\sigma_s^{\mathrm{vis}}(v_s)\big) - \phi^2\, k_f^{-1}(v_f - v_s), w_s \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}_0^1(\Omega)]^d} \\ + \Big\langle -\mathrm{div}\,\big(\phi\,\sigma_f(v_f)\big) + \phi^2\, k_f^{-1}(v_f - v_s) - \theta\, v_f, w_f \Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d, [\mathrm{H}_0^1(\Omega)]^d}.\end{aligned}$$

Applying Theorem 1.24, we get the existence of a pressure $p \in \mathrm{L}_0^2(\Omega)$ such that

$$\begin{cases} \rho_s(1-\phi)\, g_s = -\mathrm{div}\,\big(\sigma_s(u_s)\big) - \mathrm{div}\,\big(\sigma_s^{\mathrm{vis}}(v_s)\big) \\ \qquad\qquad\qquad\qquad -\phi^2\, k_f^{-1}(v_f - v_s) + (1-\phi)\,\nabla p \quad \text{in } [\mathrm{H}^{-1}(\Omega)]^d, \\ \rho_f \phi\, g_f = -\mathrm{div}\,\big(\phi\,\sigma_f(v_f)\big) + \phi^2\, k_f^{-1}(v_f - v_s) - \theta\, v_f + \phi\,\nabla p \quad \text{in } [\mathrm{H}^{-1}(\Omega)]^d, \end{cases}$$

which proves (1.66). Since g_s and g_f belong to $[\mathrm{L}^2(\Omega)]^d$, the above relation essentially holds true in $[\mathrm{L}^2(\Omega)]^d$, which yields (1.65). □

Remark 1.28. The characterization of the Lagrange multiplior p associated with the constraint on the mixture velocity as the weak solution of an elliptic problem, as done in [Avalos and Triggiani, 2009] in the context of fluid-structure interaction problems, is not straightforward precisely because the constraint involves the mixture velocity which is not a natural unknown of our coupled problem.

Lastly, we set $g = (0, f, f)$ and we denote by Π the Leray projection operator from $[\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d$ into $H = [\mathrm{H}_0^1(\Omega)]^d \times H_\phi$. We are now ready to state the following existence and uniqueness result.

Theorem 1.29. *Assume that* $(h1) - (h3)$ *hold true, that* $\eta > 0$ *and that* $\int_\Omega \theta\,\mathrm{d}x = 0$.

(i) If $z_0 \in D(A_\eta^\infty)$ *and* $f \in \mathrm{H}^1(0,T;[\mathrm{L}^2(\Omega)]^d)$ *so that* $\Pi g \in \mathrm{H}^1(0,T;H)$, *then there exists a unique strong solution* $z \in \mathrm{C}^1([0,T];H) \cap \mathrm{C}^0([0,T];D(A_\eta^\infty))$ *satisfying*

$$\begin{cases} \dot z(t) + A_\eta^\infty z(t) = \Pi g(t), \quad t \in [0,T], \\ z(0) = z_0. \end{cases} \tag{1.69}$$

(ii) If $z_0 \in H$ *and* $f \in \mathrm{L}^2(0,T;[\mathrm{L}^2(\Omega)]^d)$ *so that* $\Pi g \in \mathrm{L}^2(0,T;H)$, *then Problem* (1.69) *has a unique mild solution* $z \in \mathrm{C}^0([0,T];H)$ *such that* $z(0) = z_0$ *and*

$$\int_0^T z(t)\psi(t)\,\mathrm{d}t \in D(A_\eta^\infty),$$

$$-\int_0^T z(t)\dot\psi(t)\,\mathrm{d}t + A_\eta^\infty\left(\int_0^T z(t)\psi(t)\,\mathrm{d}t\right) = \int_0^T \Pi g(t)\psi(t)\,\mathrm{d}t,$$

for all $\psi \in \mathrm{C}^1_c([0,T];\mathbb{R})$. *Moreover,* z *is given by the Duhamel formula*

$$z(t) = \Phi_\eta^\infty(t) z_0 + \int_0^t \Phi_\eta^\infty(t-s)\Pi g(s)\,\mathrm{d}s, \tag{1.70}$$

where Φ_η^∞ *denotes the continuous semigroup generated by* A_η^∞ *in the sense that*

$$A_\eta^\infty x = -\frac{\mathrm{d}}{\mathrm{d}t}\big(\Phi_\eta^\infty(t)x\big)_{|t=0^+}, \quad x \in H. \tag{1.71}$$

Proof. The proof of this result is almost similar to the proof of Theorem 1.7, replacing Z by H and Y by V. The only difference is that the term $-\theta\, v_f$ is treated within the operator A_η^∞ instead of being considered as a perturbation.

For any $z = (u_s, v_s, v_f) \in D(A_\eta^\infty)$, we observe that

$$a_\eta^\infty(z,z) = 2\eta \int_\Omega |\varepsilon(v_s)|^2\,\mathrm{d}x + \int_\Omega \phi^2\, k_f^{-1}(v_f - v_s)\cdot(v_f - v_s)\,\mathrm{d}x \\ + \int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x - \int_\Omega \theta\,|v_f|^2\,\mathrm{d}x.$$

Thus $(A_\eta^\infty z, z)_H \geq -\omega \|z\|_H^2$ with $\omega = (\rho_f \phi_{\min})^{-1}\|\theta\|_{\mathrm{L}^\infty(\Omega)}$.

Moreover, for all $\lambda_0 > \omega$, the operator $A_\eta^\infty + \lambda_0 I$ is surjective from $D(A_\eta^\infty)$ to H because

$$a_\eta^\infty(z,z) + \lambda_0 (z,z)_H \\ \geq \lambda_0 \|u_s\|_{\mathrm{s}}^2 + 2\eta\|\varepsilon(v_s)\|^2 + 2\mu_f\phi_{\min}\|\varepsilon(v_f)\|^2 + \big(\lambda_0\rho_f\phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)}\big)\|v_f\|^2.$$

From Lumer-Phillips theorem, we deduce that A_η^∞ is the generator – in the sense of (1.71) – of a strongly continuous semigroup and the conclusion follows from [Bensoussan et al., 2007, Part II, Chapter 1, Propositions 3.1–3.3] and [Burq and Gérard, 2002, Corollary 2.25]. □

Remark 1.30. By reproducing the proof of Theorem 1.7, we can extend the result of Theorem 1.29 for a time-dependent θ satisfying the compatibility condition $\int_\Omega \theta(t)\,\mathrm{d}x = 0$ for all $t \in [0,T]$. The existence of a mild solution then requires that $\theta \in \mathrm{C}^0([0,T]\times\Omega)$ and the existence of a strong solution is guaranteed under the assumption $\theta \in \mathrm{C}^1([0,T];\mathrm{L}^\infty(\Omega))$. However, because of the lifting (1.61), more regularity on θ is needed for the original problem to be well-posed. More precisely, when performing the change of variables $(u_s, v_f, p) \longmapsto \Big(u_s - \int_0^t v_\theta(s)\,\mathrm{d}s, v_f - v_\theta, p\Big)$, the right-hand sides of (1.59b) and (1.59c) become respectively

$$\rho_s(1-\phi)f - \rho_s(1-\phi)\partial_t v_\theta + \operatorname{div}\left(\sigma_s\left(\int_0^t v_\theta\,\mathrm{d}s\right)\right) + \operatorname{div}\big(\sigma_s^{\mathrm{vis}}(v_\theta)\big),$$

and

$$\rho_f\phi f - \rho_f\phi\partial_t v_\theta + \operatorname{div}\big(\phi\sigma_f(v_\theta)\big) + \theta v_\theta.$$

The existence of a mild solution requires that all terms belong to $\mathrm{L}^2(0,T;[\mathrm{L}^2(\Omega)]^d)$ and thus that the lifting $v_\theta \in \mathrm{H}^1(0,T;[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}^2(\Omega)]^d)$, which is ensured if $\theta \in \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega)) \cap \mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$. Similarly, the existence of a strong solution is guaranteed under the assumption $\theta \in \mathrm{H}^2(0,T;\mathrm{L}^2(\Omega)) \cap \mathrm{H}^1(0,T;\mathrm{H}^1(\Omega))$.

The previous theorem only involves the displacement and velocity fields. The existence of pressure and the relation between (1.69) and the original system (1.59) are precised below.

Corollary 1.31. *Assume that* $z_0 \in D(A_\eta^\infty)$, $f \in \mathrm{H}^1(0,T;[\mathrm{L}^2(\Omega)]^d)$ *and let* $z = (u_s, v_s, v_f) \in \mathrm{C}^1([0,T];H) \cap \mathrm{C}^0([0,T];D(A_\eta^\infty))$ *be the strong solution of* (1.69). *There exists a unique pressure* $p \in \mathrm{C}^0([0,T];\mathrm{L}^2_0(\Omega))$ *such that* (z,p) *satisfies* (1.59) *pointwise almost everywhere.*

Proof. Let $z = (u_s, v_s, v_f)$ be the solution of (1.69). Since $z \in \mathrm{C}^1([0,T];H) \cap \mathrm{C}^0([0,T];D(A_\eta^\infty))$, for almost every $t \in (0,T)$, the equation

$$A_\eta^\infty z(t) = \Pi g(t) - \dot{z}(t),$$

holds true in the energy space H, where we recall that $g = (0, f, f)$. Thus for almost every $t \in (0,T)$, Proposition 1.27 ensures the existence of a pressure $p(t) \in \mathrm{L}^2_0(\Omega)$ such that

$$\begin{cases} -\partial_t u_s = -v_s, \\ \rho_s(1-\phi)\, f - \rho_s(1-\phi)\, \partial_t v_s = -\mathrm{div}\left(\sigma_s^{\mathrm{vis}}(v_s)\right) - \mathrm{div}\left(\sigma_s(u_s)\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad - \phi^2\, k_f^{-1}(v_f - v_s) + (1-\phi)\,\nabla p, \\ \rho_f \phi\, f - \rho_f \phi\, \partial_t v_f = -\mathrm{div}\left(\phi\, \sigma_f(v_f)\right) + \phi^2\, k_f^{-1}(v_f - v_s) + \phi\, \nabla p. \end{cases}$$

In virtue of (1.65), the two last lines of the above system are verified at least in $[\mathrm{L}^2(\Omega)]^d$ (the first one is satisfied in $[\mathrm{H}^1_0(\Omega)]^d$). Hence (1.59) is satisfied almost everywhere. Moreover, since $z \in \mathrm{C}^0([0,T];V)$, we find that $\nabla p \in \mathrm{C}^0([0,T];H^{-1}(\Omega))$, which implies that $p \in \mathrm{C}^0([0,T];\mathrm{L}^2_0(\Omega))$ in virtue of Nečas Lemma. □

If the input data are less regular, we get the existence of displacement, velocities and pressure in the following weak sense.

Theorem 1.32. *Assume $(h1)-(h4)$ are satisfied, $\eta > 0$, $\int_\Omega \theta \,\mathrm{d}x = 0$ and $z_0 = (u_{s0}, v_{s0}, v_{f_0}) \in H$. Then there exists a unique variational solution $u_s \in \mathrm{C}^0([0,T];[\mathrm{H}^1_0(\Omega)]^d)$, $\partial_t u_s \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$ and $v_f \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$ with $(\partial_t u_s, v_f) \in \mathrm{L}^2(0,T;V_\phi)$ such that*

$$\big(u_s(0), \partial_t u_s(0), v_f(0)\big) = (u_{s0}, v_{s0}, v_{f_0}),$$

and, for all $(w_s, w_f) \in V_\phi$, the following equation holds in $\mathcal{D}'(0,T)$:

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}\int_\Omega \rho_s(1-\phi)\, u_s(t)\cdot w_s \,\mathrm{d}x + \int_\Omega \sigma_s(u_s(t)) : \varepsilon(w_s)\,\mathrm{d}x + 2\eta\int_\Omega \varepsilon(\partial_t u_s(t)) : \varepsilon(w_s)\,\mathrm{d}x \\ + \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \rho_f \phi\, v_f(t)\cdot w_f \,\mathrm{d}x + \int_\Omega \phi\, \sigma_f(v_f(t)) : \varepsilon(w_f)\,\mathrm{d}x \\ + \int_\Omega \phi^2\, k_f^{-1}(v_f(t) - \partial_t u_s(t))\cdot(w_f - w_s)\,\mathrm{d}x - \int_\Omega \theta\, v_f(t)\cdot w_f \,\mathrm{d}x \\ = \int_\Omega \rho_s(1-\phi) f(t)\cdot w_s \,\mathrm{d}x + \int_\Omega \rho_f \phi f(t)\cdot w_f \,\mathrm{d}x. \end{aligned} \tag{1.72}$$

The energy estimate (1.17) holds true and the variational solution coincides with the mild solution given by (1.70). Furthermore, there exists a unique pressure $p \in \mathcal{D}'\big((0,T)\times\Omega\big)$ such that (u_s, v_s, v_f, p) satisfies (1.59) in the distribution sense, with $v_s = \partial_t u_s$.

Remark 1.33. The incompressibility constraint is satisfied since $(\partial_t u_s, v_f)$ belongs to $\mathrm{L}^2(0,T;V_\phi)$. It can be written in variational form as

$$\forall q \in \mathrm{L}^2(\Omega), \quad \int_\Omega \mathrm{div}\left((1-\phi)\,\partial_t u_s(t) + \phi\, v_f(t)\right) q \,\mathrm{d}x = 0.$$

Proof. The proof of existence of (u_s, v_f) follows exactly the same lines as in the compressible case. We build a sequence of strong solutions for smooth data. These solutions $(u_s^n, v_s^n, v_f^n)_n$ satisfy the energy estimate (1.17) and constitute a Cauchy sequence in $\mathrm{C}^0([0,T];H)$. Moreover (v_s^n, v_f^n) is a

Cauchy sequence in $\mathrm{L}^2(0,T;V_\phi)$ and we can pass to the limit as we did in the proof of Theorem 1.10. We get the first order system:

$$\begin{cases} \forall t\in[0,T], \forall (d_s,w_s,w_f)\in V=[\mathrm{H}^1_0(\Omega)]^d\times V_\phi, \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \sigma_s(u_s(t)):\varepsilon(d_s)\,\mathrm{d}x-\int_\Omega \sigma_s(v_s(t)):\varepsilon(d_s)\,\mathrm{d}x=0, & (1.73a) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \rho_s(1-\phi)\,v_s(t)\cdot w_s\,\mathrm{d}x+\int_\Omega \sigma_s(u_s(t)):\varepsilon(w_s)\,\mathrm{d}x \\ \quad +2\eta\displaystyle\int_\Omega \varepsilon(v_s(t)):\varepsilon(w_s)\,\mathrm{d}x+\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \rho_f\phi\,v_f(t)\cdot w_f\,\mathrm{d}x-\int_\Omega \theta\,v_f(t)\cdot w_f\,\mathrm{d}x \\ \quad +\displaystyle\int_\Omega \phi\,\sigma_f(v_f(t)):\varepsilon(w_f)\,\mathrm{d}x+\int_\Omega \phi^2\,k_f^{-1}(v_f(t)-v_s(t))\cdot(w_f-w_s)\,\mathrm{d}x \\ \quad =\displaystyle\int_\Omega \rho_s(1-\phi)f(t)\cdot w_s\,\mathrm{d}x+\int_\Omega \rho_f\phi f(t)\cdot w_f\,\mathrm{d}x, & (1.73b) \end{cases}$$

which can be rewritten in second order to obtain (1.72).

Apart for the regularity provided by the energy estimate (in particular $(v_s,v_f)\in\mathrm{L}^2(0,T;V_\phi)$), like in the compressible regime, the previous system provides some regularity on the time derivatives of the solution. The first equation (1.73a) states that $\partial_t u_s=v_s$ in $\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$ and the second equation (1.73b) implies that

$$(\partial^2_{tt}u_s,\partial_t v_f)\in\mathrm{L}^2(0,T;V_\phi').$$

These regularities, together with the density of V_ϕ in H_ϕ, yield

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\int_\Omega \rho_s(1-\phi)\,u_s(t)\cdot w_s\,\mathrm{d}x+\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \rho_f\phi\,v_f(t)\cdot w_f\,\mathrm{d}x \\ =\Big\langle\big(\rho_s(1-\phi)\,\partial^2_{tt}u_s(t),\rho_f\phi\,\partial_t v_f(t)\big),(w_s,w_f)\Big\rangle_{V_\phi',V_\phi}$$

in $\mathcal{D}'(0,T)$, for all $(w_s,w_f)\in V_\phi$. Since functions in $V_\phi\otimes\mathcal{D}(0,T)$ generate the space $\mathrm{L}^2(0,T;V_\phi)$, we get

$$\begin{cases} \forall (w_s,w_f)\in\mathrm{L}^2(0,T;V_\phi), \\ \displaystyle\int_0^T\Big\langle\big(\rho_s(1-\phi)\,\partial^2_{tt}u_s,\rho_f\phi\,\partial_t v_f\big),(w_s,w_f)\Big\rangle_{V_\phi',V_\phi}\mathrm{d}t+\int_0^T\int_\Omega \sigma_s(u_s):\varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t \\ \quad +2\eta\displaystyle\int_0^T\int_\Omega \varepsilon(v_s):\varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t+\int_0^T\int_\Omega \phi\,\sigma_f(v_f):\varepsilon(w_f)\,\mathrm{d}x\,\mathrm{d}t \\ \quad +\displaystyle\int_0^T\int_\Omega \phi^2\,k_f^{-1}(v_f-v_s)\cdot(w_f-w_s)\,\mathrm{d}x\,\mathrm{d}t-\int_0^T\int_\Omega \theta\,v_f\cdot w_f\,\mathrm{d}x\,\mathrm{d}t \\ \quad =\displaystyle\int_0^T\int_\Omega \rho_s(1-\phi)f\cdot w_s\,\mathrm{d}x\,\mathrm{d}t+\int_0^T\int_\Omega \rho_f\phi f\cdot w_f\,\mathrm{d}x\,\mathrm{d}t. \end{cases}$$

The energy estimate (1.17) then follows by choosing $(w_s,w_f)=(\partial_t u_s,v_f)$ above and guarantees uniqueness.

The equivalence between the variational and mild solutions can be proved in the same way as in the compressible case (see Proposition 1.13). We only need to notice that V is dense in H.

To show the existence of pressure, we integrate (1.73b) in time over $(0,t)$ (see for instance [Temam, 2001] for a similar argument for the Stokes system). Let us define

$$U_s(t)=\int_0^t u_s(s)\,\mathrm{d}s,\ V_s(t)=\int_0^t v_s(s)\,\mathrm{d}s,\ V_f(t)=\int_0^t v_f(s)\,\mathrm{d}s \text{ and } F(t)=\int_0^t f(s)\,\mathrm{d}s,$$

it follows that

$$\Big\langle \rho_s(1-\phi)\,(v_s(t)-v_{s0}) - \operatorname{div}\big(\sigma_s^{\mathrm{vis}}(V_s(t))\big) - \operatorname{div}\big(\sigma_s(U_s(t))\big) \\ - \phi^2\,k_f^{-1}(V_f(t)-V_s(t)) - \rho_s(1-\phi)\,F(t), w_s\Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d,[\mathrm{H}_0^1(\Omega)]^d} \\ + \Big\langle \rho_f\phi\,(v_f(t)-v_{f_0}) - \operatorname{div}\big(\phi\,\sigma_f(V_f(t))\big) + \phi^2\,k_f^{-1}(V_f(t)-V_s(t)) - \rho_f\phi\,F(t), w_f\Big\rangle_{[\mathrm{H}^{-1}(\Omega)]^d,[\mathrm{H}_0^1(\Omega)]^d} = 0,$$

for all $(w_s, w_f) \in V_\phi$. Combining Theorem 1.24 and Nečas Lemma provides the existence of $P \in \mathrm{C}^0([0,T];\mathrm{L}^2_0(\Omega))$ such that

$$\rho_s(1-\phi)\,(v_s - v_{s0}) - \operatorname{div}\big(\sigma_s^{\mathrm{vis}}(V_s)\big) - \operatorname{div}\big(\sigma_s(U_s)\big) - \phi^2\,k_f^{-1}(V_f - V_s) - \rho_s(1-\phi)\,F = -(1-\phi)\,\nabla P,$$
$$\rho_f\phi\,(v_f - v_{f_0}) - \operatorname{div}\big(\phi\,\sigma_f(V_f)\big) + \phi^2\,k_f^{-1}(V_f - V_s) - \rho_f\phi\,F = -\phi\,\nabla P.$$

As a consequence, $p = \partial_t P$ satisfies (1.59) in the distribution sense. □

Remark 1.34. We could also define the variational solution without assuming time continuity. Time continuity would then follow from the continuous injection of the space

$$W_\phi(0,T) = \Big\{(v_s, v_f) \in \mathrm{L}^2(0,T;V_\phi) \text{ such that } (\partial_t v_s, \partial_t v_f) \in \mathrm{L}^2(0,T;V_\phi')\Big\},$$

into $\mathrm{C}^0([0,T];H_\phi)$.

1.3.3 The case $\eta = 0$

This case combines the two difficulties encountered earlier: the incompressibility constraint and the absence of solid viscosity. To handle this case, an option is to combine the functional framework adapted to the incompressibility constraint with the T-coercivity approach used in Section 1.2.3. This method provides the following result.

Theorem 1.35. *If $\eta = 0$, then the conclusions of Theorem 1.29 and Corollary 1.31 remain true.*

Proof. To prove that the operator $A_0^\infty + \lambda_0 I$ is surjective from $D(A_0^\infty)$ to H for all $\lambda_0 > \omega$, we show that $a_0^\infty(\cdot,\cdot) + \lambda_0(\cdot,\cdot)_H$ is T-coercive for the mapping $\mathtt{T} : (u_s, v_s, v_f) \longmapsto (\frac{1}{2}u_s - \frac{1}{2\lambda_0}v_s, v_s, v_f)$ defined by (1.45). To do so, we reproduce exactly the same calculations as in the compressible case (see the proof of Theorem 1.17), but replacing Y and Z by V and H respectively. This mapping is bijective from V into itself because it does not affect the velocity components and thus the mixture's divergence constraint. The rest of the proof is the same as in Theorem 1.29 and Corollary 1.31. □

Yet, we present here another approach to prove that $A_0^\infty + \lambda_0 I$ is surjective. This proof is based on a mixed formulation of the problem and is more suitable for numerical approximation. Indeed, the formulation (1.64) involves a constrained space V_ϕ that we would like to relax for numerical purpose, most numerical strategies relying on mixed formulations. Note moreover that the space V_ϕ depends on the porosity and thus on a specific data set. The mixed problem we would like to solve writes

$$\begin{cases} \text{Find } z \in Y_0 \quad \text{such that} \\ \forall y \in Y_0, \quad a_{\lambda_0}(z,y) = (g,y)_{Z_0}, \end{cases} \tag{1.75}$$

with

$$\begin{aligned}
a_{\lambda_0}(z,y) &= \lambda_0 \int_\Omega \sigma_s(u_s) : \varepsilon(d_s)\,\mathrm{d}x - \int_\Omega \sigma_s(v_s) : \varepsilon(d_s)\,\mathrm{d}x \\
&+ \lambda_0 \int_\Omega \rho_s(1-\phi)\, v_s \cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s(u_s) : \varepsilon(w_s)\,\mathrm{d}x \\
&+ \int_\Omega \phi^2\, k_f^{-1}(v_f - v_s)\cdot(w_f - w_s)\,\mathrm{d}x + \lambda_0 \int_\Omega \rho_f \phi\, v_f \cdot w_f\,\mathrm{d}x \\
&+ \int_\Omega \phi\, \sigma_f(v_f) : \varepsilon(w_f)\,\mathrm{d}x - \int_\Omega \theta v_f \cdot w_f\,\mathrm{d}x \\
&- \int_\Omega p\,\mathrm{div}\left((1-\phi)\, w_s + \phi\, w_f\right)\mathrm{d}x + \int_\Omega \mathrm{div}\left((1-\phi)\, v_s + \phi\, v_f\right) q\,\mathrm{d}x,
\end{aligned}$$

for any $z = (u_s, v_s, v_f, p)$ and $y = (d_s, w_s, w_f, q)$ in Y_0, where

$$Z_0 = [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times \mathrm{L}^2_0(\Omega)$$

and

$$Y_0 = [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times \mathrm{L}^2_0(\Omega).$$

The spaces Z_0 and Y_0 are almost similar to Z and Y but include the additional condition $\int_\Omega p\,\mathrm{d}x = 0$ that is required to ensure pressure uniqueness.

Proposition 1.36. *Let $g \in Z_0$. If $\lambda_0 > (\rho_f \phi_{\min})^{-1} \|\theta\|_{\mathrm{L}^\infty(\Omega)}$, then Problem (1.75) is well-posed in Y_0.*

Proof. According to Proposition 1.16, it is sufficient to find y^* depending continuously on z such that the inequality $a_{\lambda_0}(z, y^*) \geq \underline{\alpha} \|z\|^2_{Y_0}$ is satisfied for any $z \in Z_0$, with $\underline{\alpha} > 0$.

From the properties of the divergence operator, we know, as already stated in (1.63), that for any $p \in \mathrm{L}^2_0(\Omega)$, there exists $v_p \in [\mathrm{H}^1_0(\Omega)]^d$ and $C_{\mathrm{div}} > 0$ such that

$$\mathrm{div}\, v_p = p \quad \text{and} \quad \|\nabla v_p\|^2 \leq C_{\mathrm{div}} \|p\|^2.$$

For some constants α, β and γ to be adjusted, we choose $d_s^* = \beta\, u_s + \gamma\, v_s$, $w_s^* = \alpha\, v_s - v_p$, $w_f^* = \alpha\, v_f - v_p$ and $q^* = \alpha\, p$. Thus

$$\begin{aligned}
a_{\lambda_0}(z, y^*) &= \lambda_0 \int_\Omega \beta\, \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x + \lambda_0 \int_\Omega \gamma\, \sigma_s(u_s) : \varepsilon(v_s)\,\mathrm{d}x \\
&- \int_\Omega \beta \sigma_s(v_s) : \varepsilon(u_s)\,\mathrm{d}x - \int_\Omega \gamma\, \sigma_s(v_s) : \varepsilon(v_s)\,\mathrm{d}x - \int_\Omega \sigma_s(u_s) : \varepsilon(v_p)\,\mathrm{d}x \\
&+ \lambda_0 \int_\Omega \rho_s(1-\phi)\left(\alpha\,|v_s|^2 - v_s \cdot v_p\right)\mathrm{d}x + \int_\Omega \alpha\, \sigma_s(u_s) : \varepsilon(v_s)\,\mathrm{d}x \\
&+ \int_\Omega \alpha\, \phi^2\, k_f^{-1}(v_f - v_s)\cdot(v_f - v_s)\,\mathrm{d}x + \int_\Omega \left(\lambda_0 \rho_f \phi - \theta\right)\left(\alpha\,|v_f|^2 - v_f \cdot v_p\right)\mathrm{d}x \\
&+ \int_\Omega \phi\left(\alpha\, \sigma_f(v_f) : \varepsilon(v_f) - \sigma_f(v_f) : \varepsilon(v_p)\right)\mathrm{d}x + \int_\Omega p\,\mathrm{div}\left((1-\phi)\, v_p + \phi\, v_p\right)\mathrm{d}x.
\end{aligned}$$

Note that the term $\int_\Omega p\,\mathrm{div}\left((1-\phi)\, v_p + \phi\, v_p\right)\mathrm{d}x$ is equal to $\|p\|^2$ thanks to the choice of v_p. As in the proof of Theorem 1.17, we set $\beta = \frac{\alpha}{2}$ and $\gamma = -\frac{\alpha}{2\lambda_0}$ in order to remove the terms in the form

$\int_\Omega \sigma_s(u_s) : \varepsilon(v_s)\,\mathrm{d}x$. Consequently, we have

$$\begin{aligned} a_{\lambda_0}(z, y^*) \geq \frac{\lambda_0\alpha}{2}\int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x - \int_\Omega \sigma_s(u_s) : \varepsilon(v_p)\,\mathrm{d}x \\ + \frac{\alpha}{2\lambda_0}\int_\Omega \sigma_s(v_s) : \varepsilon(v_s)\,\mathrm{d}x + \lambda_0\rho_s(1-\phi_{\max})\int_\Omega \big(\alpha\,|v_s|^2 - v_s\cdot v_p\big)\,\mathrm{d}x \\ + \big(\lambda_0\rho_f\phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)}\big)\int_\Omega \big(\alpha\,|v_f|^2 - v_f\cdot v_p\big)\,\mathrm{d}x \\ + \phi_{\min}\int_\Omega \big(\alpha\,\sigma_f(v_f) : \varepsilon(v_f) - \sigma_f(v_f) : \varepsilon(v_p)\big)\,\mathrm{d}x + \int_\Omega p^2\,\mathrm{d}x. \end{aligned} \tag{1.76}$$

We choose λ_0 such that $\lambda_0\rho_f\phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)} > 0$. Next, for all $\delta > 0$, Young inequality yields

$$\begin{aligned} -\int_\Omega \sigma_s(u_s) : \varepsilon(v_p)\,\mathrm{d}x \geq -\frac{\delta}{2}\int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega \sigma_s(v_p) : \varepsilon(v_p)\,\mathrm{d}x, \\ -\int_\Omega \sigma_f(v_f) : \varepsilon(v_p)\,\mathrm{d}x \geq -\frac{\delta}{2}\int_\Omega \sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega \sigma_f(v_p) : \varepsilon(v_p)\,\mathrm{d}x, \\ -\int_\Omega v_s\cdot v_p\,\mathrm{d}x \geq -\frac{\delta}{2}\int_\Omega |v_s|^2\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega |v_p|^2\,\mathrm{d}x, \\ -\int_\Omega v_f\cdot v_p\,\mathrm{d}x \geq -\frac{\delta}{2}\int_\Omega |v_f|^2\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega |v_p|^2\,\mathrm{d}x. \end{aligned} \tag{1.77}$$

Furthermore, it holds

$$\begin{aligned} \|v_p\|^2 \leq C_p\|\nabla v_p\|^2 \leq C_pC_{\mathrm{div}}\|p\|^2, \\ \int_\Omega \sigma_f(v_p) : \varepsilon(v_p)\,\mathrm{d}x = \lambda_f\|\mathrm{div}\,v_p\|^2 + 2\mu_f\|\varepsilon(v_p)\|^2 \leq (\lambda_f + 2\mu_fC_{\mathrm{div}})\|p\|^2, \\ \int_\Omega \sigma_s(v_p) : \varepsilon(v_p)\,\mathrm{d}x = \lambda\|\mathrm{div}\,v_p\|^2 + 2\mu\|\varepsilon(v_p)\|^2 \leq (\lambda + 2\mu C_{\mathrm{div}})\|p\|^2, \end{aligned} \tag{1.78}$$

where C_p denotes the constant of Poincaré inequality.

Using (1.77) and (1.78) to bound from below the right-hand side of (1.76) and rearranging terms, we obtain

$$\begin{aligned} a_{\lambda_0}(z, y^*) \geq \Big(\frac{\lambda_0\alpha}{2} - \frac{\delta}{2}\Big)\|u_s\|_{\mathrm{s}}^2 + \frac{\alpha}{2\lambda_0}\|v_s\|_{\mathrm{s}}^2 + \lambda_0\rho_s(1-\phi_{\max})\Big(\alpha - \frac{\delta}{2}\Big)\|v_s\|^2 \\ + \big(\lambda_0\rho_f\phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)}\big)\Big(\alpha - \frac{\delta}{2}\Big)\|v_f\|^2 + 2\mu_f\phi_{\min}\Big(\alpha - \frac{\delta}{2}\Big)\|\varepsilon(v_f)\|^2 + \Big(1 - \frac{\delta^*}{2\delta}\Big)\|p\|^2, \end{aligned}$$

where $\delta^* = \lambda + 2\mu C_{\mathrm{div}} + \lambda_0\rho_s(1-\phi_{\max})C_pC_{\mathrm{div}} + (\lambda_0\rho_f\phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)})C_pC_{\mathrm{div}} + \phi_{\min}(\lambda_f + 2\mu_fC_{\mathrm{div}})$, $\delta^* > 0$.

Hence, setting $\delta = \delta^*$ and $\alpha = \alpha^* = \max\big(\delta^*, \frac{2\delta^*}{\lambda_0}\big)$, we get

$$a_{\lambda_0}(z, y^*) \geq \frac{\delta^*}{2}\|u_s\|_{\mathrm{s}}^2 + \frac{\alpha^*}{2\lambda_0}\|v_s\|_{\mathrm{s}}^2 + \mu_f\phi_{\min}\delta^*\|\varepsilon(v_f)\|^2 + \frac{1}{2}\|p\|^2.$$

Finally, we infer that a_{λ_0} is $\mathtt{T}$-coercive for the mapping

$$\mathtt{T} : (u_s, v_s, v_f, p) \longmapsto \Big(\frac{\alpha^*}{2}u_s - \frac{\alpha^*}{2\lambda_0}v_s, \alpha^*v_s - v_p, \alpha^*v_f - v_p, \alpha^*p\Big), \tag{1.79}$$

which is bijective since $p \longmapsto v_p$ is a bijection. □

Remark 1.37. This mixed formulation is also applicable to the case $\kappa = +\infty$ and $\eta > 0$. In that case, the proof can be simplified by considering the mapping $\mathtt{T} : (u_s, v_s, v_f, p) \longmapsto (u_s, \alpha^* v_s - v_p, \alpha^* v_f - v_p, \alpha^* p)$.

Remark 1.38. The mixed formulation is equivalent to the constrained formulation thanks to the inf-sup property proved in Proposition 1.22. Note moreover that, as for the proof of Proposition 1.22, the T-coercivity only relies on the standard inf-sup condition for the divergence operator and therefore is independent of the porosity ϕ.

Finally, as for the compressible inviscid case, we can prove the existence of a variational solution.

Theorem 1.39. *Assume $(h1)-(h4)$ are satisfied, $\eta = 0$, $\int_\Omega \theta \,\mathrm{d}x = 0$ and $z_0 = (u_{s0}, v_{s0}, v_{f_0}) \in H$. Then there exists a unique variational solution $u_s \in \mathrm{C}^0([0,T]; [\mathrm{H}^1_0(\Omega)]^d)$ and $(\partial_t u_s, v_f) \in \mathrm{C}^0([0,T]; H_\phi)$ with $v_f \in \mathrm{L}^2(0,T; [\mathrm{H}^1_0(\Omega)]^d)$such that*

$$\big(u_s(0), \partial_t u_s(0), v_f(0)\big) = (u_{s0}, v_{s0}, v_{f_0})$$

and the following equations hold, in $\mathcal{D}'(0,T)$,

$$\begin{cases} \dfrac{\mathrm{d}^2}{\mathrm{d}t^2}\displaystyle\int_\Omega \rho_s(1-\phi)\, u_s(t)\cdot w_s \,\mathrm{d}x + \int_\Omega \sigma_s(u_s(t)) : \varepsilon(w_s)\,\mathrm{d}x \\ \quad + \dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int_\Omega \rho_f \phi\, v_f(t)\cdot w_f \,\mathrm{d}x + \int_\Omega \phi\, \sigma_f(v_f(t)) : \varepsilon(w_f)\,\mathrm{d}x \\ \quad + \displaystyle\int_\Omega \phi^2\, k_f^{-1}(v_f(t) - \partial_t u_s(t))\cdot(w_f - w_s)\,\mathrm{d}x - \int_\Omega \theta\, v_f(t)\cdot w_f \,\mathrm{d}x \\ \quad = \displaystyle\int_\Omega \rho_s(1-\phi) f(t)\cdot w_s \,\mathrm{d}x + \int_\Omega \rho_f \phi f(t)\cdot w_f \,\mathrm{d}x, \quad \forall (w_s, w_f) \in V_\phi. \end{cases} \tag{1.80}$$

The energy estimate (1.17) holds true (with $\eta = 0$) and the variational solution coincides with the mild solution. Furthermore, there exists a unique pressure $p \in \mathcal{D}'\big((0,T)\times\Omega\big)$ such that (u_s, v_s, v_f, p) satisfies (1.59) in the distribution sense, with $v_s = \partial_t u_s$.

Proof. We follow the same steps as in Theorem 1.18, but within the functional framework adapted to the incompressibility constraint. Existence of solutions is obtained by an approximated sequence of strong solutions $z^n = (u_s^n, v_s^n, v_f^n) \in \mathrm{C}^1([0,T]; H) \cap \mathrm{C}^0([0,T]; D(A_0^\infty))$ verifying

$$\begin{cases} \dot z^n(t) + A_0^\infty z^n(t) = \Pi g^n(t), \quad t \in [0,T], \\ z^n(0) = z_0^n, \end{cases}$$

where $\Pi g^n = \Pi(0, f^n, f^n) \in \mathrm{H}^1(0,T;H)$ and $z_0^n \in D(A_0^\infty)$ denote respectively an approximation of source terms and initial conditions. This sequence of solution satisfies the variational formulation

$$(VF)^n_\infty \begin{cases} \forall t \in [0,T], \forall (d_s, w_s, w_f) \in V = [\mathrm{H}^1_0(\Omega)]^d \times V_\phi, \\ \displaystyle\int_\Omega \sigma_s(\partial_t u_s^n(t)) : \varepsilon(d_s)\,\mathrm{d}x = \int_\Omega \sigma_s(v_s^n(t)) : \varepsilon(d_s)\,\mathrm{d}x, & (1.81a) \\ \displaystyle\int_\Omega \rho_s(1-\phi)\,\partial_t v_s^n(t)\cdot w_s \,\mathrm{d}x + \int_\Omega \sigma_s(u_s^n(t)) : \varepsilon(w_s)\,\mathrm{d}x \\ \quad + \displaystyle\int_\Omega \rho_f \phi\,\partial_t v_f^n(t)\cdot w_f \,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f^n(t)) : \varepsilon(w_f)\,\mathrm{d}x \\ \quad + \displaystyle\int_\Omega \phi^2\, k_f^{-1}(v_f^n(t) - v_s^n(t))\cdot(w_f - w_s)\,\mathrm{d}x - \int_\Omega \theta^n\, v_f^n(t)\cdot w_f \,\mathrm{d}x \\ \quad = \displaystyle\int_\Omega \rho_s(1-\phi) f^n(t)\cdot w_s \,\mathrm{d}x + \int_\Omega \rho_f \phi f^n(t)\cdot w_f \,\mathrm{d}x. & (1.81b) \end{cases}$$

Moreover, by taking $(u_s^n, v_s^n, v_f^n) \in \mathrm{C}^0([0,T];V)$ as test functions in (1.81a), we get that it satisfies the energy estimate (1.17) with $\eta = 0$. Hence z^n is a Cauchy sequence in $\mathrm{C}^0([0,T];H)$ and v_f^n is a Cauchy sequence in $\mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$, which allows us to pass to the limit in (1.81b). For a given $d_s \in [\mathrm{L}^2(\Omega)]^d$, we choose the unique solution $\eta_s \in [\mathrm{H}_0^1(\Omega)]^d$ of $-\mathrm{div}\left(\sigma_s(\eta_s)\right) = d_s$ as a test function in (1.81a), which yields

$$\forall t \in [0,T], \forall d_s \in [\mathrm{L}^2(\Omega)]^d, \quad \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u_s(t)\cdot d_s\,\mathrm{d}x - \int_\Omega v_s(t)\cdot d_s\,\mathrm{d}x = 0,$$

after passing to the limit. Putting these two limit formulations together gives (1.80). As for the viscous case the equation (1.80) implies that

$$(\partial_{tt}^2 u_s, \partial_t v_f) \in \mathrm{L}^2(0,T;V_\phi'),$$

and for any test functions (w_s, w_f) in V_ϕ

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\int_\Omega \rho_s(1-\phi)\,u_s\cdot w_s\,\mathrm{d}x + \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \rho_f\phi\,v_f\cdot w_f\,\mathrm{d}x = \left\langle(\rho_s(1-\phi)\,\partial_{tt}^2 u_s, \rho_f\phi\,\partial_t v_f), (w_s,w_f)\right\rangle_{V_\phi',V_\phi}.$$

To show uniqueness, we are going to use the same Ladyzhenskaya test functions as in the compressible case. The difficulty then lies in justifying that the calculations done in the compressible case remain valid in the constrained functional setting. Let (u_s, v_f) be a solution to (1.80) with zero initial conditions and source terms, and let τ be given in $(0,T)$. We first write the weak space-time variational formulation satisfied by (u_s, v_f). In a standard way, by multiplying the weak formulation (1.80) by a $\psi \in H^1(0,T)$ such that $\psi(T) = 0$ and integrating over $(0,T)$ and by parts in time we obtain

$$\begin{aligned}
&-\int_0^T\int_\Omega \rho_s(1-\phi)\,\partial_t u_s\cdot\partial_t\psi(t)w_s\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\Omega \sigma_s(u_s):\varepsilon(w_s)\psi(t)\,\mathrm{d}x\,\mathrm{d}t\\
&\qquad -\int_0^T\int_\Omega \rho_f\phi\,v_f\cdot\partial_t\psi(t)w_f\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\Omega \phi\,\sigma_f(v_f):\varepsilon(w_f)\psi(t)\,\mathrm{d}x\,\mathrm{d}t\\
&\qquad +\int_0^T\int_\Omega \phi^2\,k_f^{-1}(v_f-\partial_t u_s)\cdot(w_f-w_s)\psi(t)\,\mathrm{d}x\,\mathrm{d}t - \int_0^T\int_\Omega \theta\,v_f\cdot w_f\psi(t)\,\mathrm{d}x\,\mathrm{d}t = 0.
\end{aligned}$$

Since linear combinaisons of fonctions of the type $(\psi(t)w_s, \psi(t)w_f)$ with $\psi \in H^1(0,T)$ such that $\psi(T) = 0$ and $(w_s,w_f) \in V_\phi$ are dense in the space of functions $\boldsymbol{w}$ of $H^1(0,T;V_\phi)$ such that $\boldsymbol{w}(T) = 0$, we obtain

$$\begin{cases}
-\displaystyle\int_0^T\int_\Omega \rho_s(1-\phi)\,\partial_t u_s\cdot\partial_t w_s\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\Omega \sigma_s(u_s):\varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t\\
\quad -\displaystyle\int_0^T\int_\Omega \rho_f\phi\,v_f\cdot\partial_t\psi(t)w_f\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\Omega \phi\,\sigma_f(v_f):\varepsilon(w_f)\,\mathrm{d}x\,\mathrm{d}t\\
+\displaystyle\int_0^T\int_\Omega \phi^2\,k_f^{-1}(v_f-\partial_t u_s)\cdot(w_f-w_s)\,\mathrm{d}x\,\mathrm{d}t - \int_0^T\int_\Omega \theta\,v_f\cdot w_f\,\mathrm{d}x\,\mathrm{d}t = 0,\\
\quad \forall (w_s,w_f) \in H^1(0,T;V_\phi) \text{ such that } w_s(T) = w_f(T) = 0.
\end{cases} \tag{1.82}$$

Then, we consider the same test functions as in the compressible case, namely

$$\psi_s(t) = \begin{cases} -\displaystyle\int_t^\tau u_s(\sigma)\,\mathrm{d}\sigma, \text{ if } \tau \geq t\\ 0, \text{ if } \tau \leq t\end{cases} \text{ and } \psi_f(t) = \begin{cases} -\displaystyle\int_t^\tau\int_0^v v_f(\sigma)\,\mathrm{d}\sigma\,\mathrm{d}v, \text{ if } \tau \geq t\\ 0, \text{ if } \tau \leq t.\end{cases}$$

These test functions are still admissible here. Indeed, we know that $(\partial_t u_s, v_f) \in \mathrm{C}^0([0,T];H_\phi)$. Recalling the characterization of the space H_ϕ established in Proposition 1.25, it follows that

$$\operatorname{div}\big((1-\phi)\,\partial_t u_s + \phi\, v_f\big) = 0, \quad \text{in } \mathrm{C}^0([0,T];\mathcal{D}'(\Omega)).$$

Note that, as in the inviscid compressible case, $\operatorname{div}\big((1-\phi)\,\partial_t u_s+\phi\, v_f\big)$ belongs to $\mathrm{L}^2((0,T);\mathrm{H}^{-1}(\Omega))$. Next, by integrating two times in time, we obtain

$$\operatorname{div}\left((1-\phi)\Big(-\int_t^\tau\int_0^v \partial_t u_s(\sigma)\,\mathrm{d}\sigma\,\mathrm{d}v\Big) + \phi\Big(-\int_t^\tau\int_0^v v_f(\sigma)\,\mathrm{d}\sigma\,\mathrm{d}v\Big)\right) = 0.$$

Since $u_s(0)=0$, we conclude that $(\psi_s,\psi_f)\in \mathrm{C}^1([0,T];V_\phi)$.

Choosing (ψ_s,ψ_f) as test functions, the calculations are exactly the same as in the compressible case – see (1.58) – but without the pressure terms, and with θ independent of time. We get

$$\begin{aligned}\frac12\int_\Omega \rho_s(1-\phi)\,|u_s(\tau)|^2\,\mathrm{d}x + \frac12\int_\Omega \rho_f\phi\,|\partial_t\psi_f(\tau)|^2\,\mathrm{d}x + \frac12\int_\Omega \sigma_s(\psi_s(0)):\varepsilon(\psi_s(0))\,\mathrm{d}x\\ + \int_0^\tau\int_\Omega \phi\,\sigma_f(\partial_t\psi_f):\varepsilon(\partial_t\psi_f)\,\mathrm{d}x\,\mathrm{d}t + \int_0^\tau\int_\Omega \phi^2 k_f^{-1}(\partial_t\psi_f-\partial_t\psi_s)^2\,\mathrm{d}x\,\mathrm{d}t\\ = \int_0^\tau\int_\Omega \theta\,|\partial_t\psi_f|^2\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Estimating the right-hand side as in the viscous case shows that $u_s = v_f = 0$ by an application of Grönwall Lemma.

Using Duhamel formula (1.70), one shows exactly as in the compressible case that the sequence z^n also converges towards the mild solution in $\mathrm{C}^0([0,T];H)$, so that the mild solution coincides with the variational solution built from the same approximation process. Finally, the existence of a pressure p such that (u_s,v_s,v_f,p) satisfies (1.59) in the distribution sense is obtained by combining Theorem 1.24 and Nečas Lemma, like in the proof of Theorem 1.32. □

1.4 Incompressible limit

In this section, we show how to pass to the limit in the weak formulation for $\kappa < +\infty$ as κ goes to infinity and obtain the incompressible system. Similar incompressible limits were considered for Biot's consolidation model, both in linear [Showalter, 2000] or non-linear [Bociu and Webster, 2021] regimes, and the influence of compressibility was analyzed in the 1D linear case [Bociu et al., 2019].

For this purpose, we need to get an energy estimate independent of κ in the compressible case. This can be achieved by lifting the right-hand side of the pressure equation. We consider $v_{\theta,\alpha} = \frac{1}{\alpha} v_\theta$ where v_θ is defined by (1.61), so that

$$\operatorname{div}(\alpha\, v_{\theta,\alpha}) = \rho_f^{-1}\theta. \tag{1.83}$$

Note that to ensure that $v_{\theta,\alpha}$ is in $[\mathrm{H}^2(\Omega)]^d$, we need more assumptions and more regularity on the Biot-Willis coefficient α. Therefore, $(h6)$ becomes

$$(h6)_{\text{bis}} \begin{cases} \alpha \in \mathrm{H}^{d/2+r}(\Omega) \text{ with } d/2+r \geq 2,\\ \forall x\in\Omega, \quad 0 < (\alpha-\phi)_{\min} \leq \alpha(x)-\phi(x) \leq (\alpha-\phi)_{\max} < 1. \end{cases}$$

As already noticed in Remark 1.30, in order to have the adequate regularity for the right-hand side of the equation verified by the new unknowns, such a lifting requires additional assumptions on the fluid mass input θ, namely

$$(h5)_{\text{bis}} \begin{cases} \theta \in \mathrm{C}^0([0,T]\times\Omega)\cap \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega))\cap \mathrm{L}^2(0,T;\mathrm{H}^1(\Omega)),\\ \forall t\in[0,T], \quad \displaystyle\int_\Omega \theta(t)\,\mathrm{d}x = 0. \end{cases}$$

Consequently, under $(h5)_{\rm bis}$ and $(h6)_{\rm bis}$, the lifting $v_{\theta,\alpha}$ satisfying (1.83) belongs to the space $\mathrm{H}^1(0,T;[\mathrm{H}^1_0(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}^2(\Omega)]^d)$, and we make the change of variables $(u_s,v_f,p)\longmapsto\Big(u_s-\int_0^t v_{\theta,\alpha}(s)\,\mathrm{d}s, v_f-v_{\theta,\alpha},p\Big)$ so that the right-hand side of the pressure equation reduces to zero.

In order to recover the initial conditions in the incompressible limit, we are not going to pass to the limit in (1.37) and (1.47) which are written in $\mathcal{D}'(0,T)$, but rather in the following weak formulation: for any $(u_{s0},v_{s0},v_{f_0},p_0)\in Z$ and $f\in\mathrm{L}^2(0,T;[\mathrm{L}^2(\Omega)]^d)$, find $u_s^\kappa\in\mathrm{C}^0([0,T];[\mathrm{H}^1_0(\Omega)]^d)$, $\partial_t u_s^\kappa\in\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$ if $\eta>0$ or $\partial_t u_s^\kappa\in\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$ if $\eta=0$, and $v_f^\kappa\in\mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$, and $p^\kappa\in\mathrm{C}^0([0,T];\mathrm{L}^2(\Omega))$ such that

$$\begin{cases}
\displaystyle\int_0^T\!\!\int_\Omega \rho_s(1-\phi)\,u_s^\kappa\cdot\partial_{tt}^2 w_s\,\mathrm{d}x\,\mathrm{d}t+\int_0^T\!\!\int_\Omega\sigma_s(u_s^\kappa):\varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t\\
\quad+2\eta\displaystyle\int_0^T\!\!\int_\Omega\varepsilon(\partial_t u_s^\kappa):\varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t-\int_0^T\!\!\int_\Omega\phi^2\,k_f^{-1}(v_f^\kappa-\partial_t u_s^\kappa)\cdot w_s\,\mathrm{d}x\,\mathrm{d}t\\
\qquad-\displaystyle\int_0^T\!\!\int_\Omega p^\kappa\,\mathrm{div}\,\big((\alpha-\phi)\,w_s\big)\,\mathrm{d}x\,\mathrm{d}t=\int_0^T\!\!\int_\Omega\rho_s(1-\phi)f\cdot w_s\,\mathrm{d}x\,\mathrm{d}t\\
\qquad+\displaystyle\int_\Omega\rho_s(1-\phi)\,v_{s0}\cdot w_s(0)\,\mathrm{d}x-\int_\Omega\rho_s(1-\phi)\,u_{s0}\cdot\partial_t w_s(0)\,\mathrm{d}x,
\end{cases}\tag{1.84a}$$

$$\begin{cases}
-\displaystyle\int_0^T\!\!\int_\Omega\rho_f\phi\,v_f^\kappa\cdot\partial_t w_f\,\mathrm{d}x\,\mathrm{d}t+\int_0^T\!\!\int_\Omega\phi\,\sigma_f(v_f^\kappa):\varepsilon(w_f)\,\mathrm{d}x\,\mathrm{d}t\\
\quad+\displaystyle\int_0^T\!\!\int_\Omega\phi^2\,k_f^{-1}(v_f^\kappa-\partial_t u_s^\kappa)\cdot w_f\,\mathrm{d}x\,\mathrm{d}t-\int_0^T\!\!\int_\Omega\theta\,v_f^\kappa\cdot w_f\,\mathrm{d}x\,\mathrm{d}t\\
\qquad-\displaystyle\int_0^T\!\!\int_\Omega p^\kappa\,\mathrm{div}\,(\phi\,w_f)\,\mathrm{d}x\,\mathrm{d}t=\int_0^T\!\!\int_\Omega\rho_f\phi f\cdot w_f\,\mathrm{d}x\,\mathrm{d}t\\
\qquad\qquad+\displaystyle\int_\Omega\rho_f\phi\,v_{f_0}\cdot w_f(0)\,\mathrm{d}x,
\end{cases}\tag{1.84b}$$

and

$$\begin{cases}
-\displaystyle\int_0^T\!\!\int_\Omega\frac{\alpha-\phi}{\kappa}p^\kappa\,\partial_t q\,\mathrm{d}x\,\mathrm{d}t+\int_0^T\!\!\int_\Omega\mathrm{div}\,\big((\alpha-\phi)\,\partial_t u_s^\kappa\big)q\,\mathrm{d}x\,\mathrm{d}t\\
\quad+\displaystyle\int_0^T\!\!\int_\Omega\mathrm{div}\,(\phi\,v_f^\kappa)q\,\mathrm{d}x\,\mathrm{d}t=\int_\Omega\frac{\alpha-\phi}{\kappa}p_0q(0)\,\mathrm{d}x,\quad\text{if }\eta>0,
\end{cases}\tag{1.85a}$$

$$\begin{cases}
-\displaystyle\int_0^T\!\!\int_\Omega\frac{\alpha-\phi}{\kappa}p^\kappa\,\partial_t q\,\mathrm{d}x\,\mathrm{d}t-\int_0^T\!\!\int_\Omega\mathrm{div}\,\big((\alpha-\phi)\,u_s^\kappa\big)\partial_t q\,\mathrm{d}x\,\mathrm{d}t\\
\quad+\displaystyle\int_0^T\!\!\int_\Omega\mathrm{div}\,(\phi\,v_f^\kappa)q\,\mathrm{d}x\,\mathrm{d}t=\int_\Omega\frac{\alpha-\phi}{\kappa}p_0q(0)\,\mathrm{d}x\\
\qquad\qquad+\displaystyle\int_\Omega\mathrm{div}\,\big((\alpha-\phi)\,u_{s0}\big)q(0)\,\mathrm{d}x,\quad\text{if }\eta=0,
\end{cases}\tag{1.85b}$$

for all admissible test functions

$$\begin{cases}
w_s\in\mathrm{H}^2(0,T;[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d),\\
w_f\in\mathrm{H}^1(0,T;[\mathrm{L}^2(\Omega)]^d)\cap\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d),\\
q\in\mathrm{H}^1(0,T;\mathrm{L}^2(\Omega)),\\
w_s(T)=\partial_t w_s(T)=w_f(T)=q(T)=0.
\end{cases}\tag{1.86}$$

The main difference between the weak formulation (1.84) – (1.85) and (1.37) or (1.47) is that the test functions depend on space but also on time. Besides, the initial conditions are weakly

imposed in (1.84) – (1.85), while they are strongly imposed in (1.36) or (1.46). This space-time weak formulation can be obtained from (1.37) or (1.47) with the same arguments used to derive (1.82).

Remark 1.40. By choosing $(w_s, w_f, q) = \big(\hat{w}_s(x), \hat{w}_f(x), \hat{q}(x)\big)\,\psi(t)$ with $(\hat{w}_s, \hat{w}_f, \hat{q}) \in [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times \mathrm{L}^2(\Omega)$ and $\psi \in \mathcal{D}(0,T)$, we see that the weak formulation (1.84) – (1.85) implies the variational formulation (1.37) or (1.47). Hence, from the uniqueness of the variational solution, the solutions of these two formulations coincide.

We are now ready to establish how the solution in the compressible case converges towards the solution in the incompressible regime as κ goes to infinity.

Theorem 1.41. *Assume that* $(h1)-(h4)$, $(h5)_{\mathrm{bis}}$ *and* $(h6)_{\mathrm{bis}}$ *are satisfied. For* $z_0 = (u_{s0}, v_{s0}, v_{f\,0}) \in Z$, *let* $(u_s^\kappa, v_f^\kappa, p^\kappa)$ *be the solution of* (1.84) – (1.85). *As* κ *goes to infinity,* $(u_s^\kappa, \partial_t u_s^\kappa, v_f^\kappa)$ *converge weakly towards the solution of the following formulation: find* $u_s \in \mathrm{C}^0([0,T];[\mathrm{H}^1_0(\Omega)]^d)$, $\partial_t u_s \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d)$ *and* $v_f \in \mathrm{C}^0([0,T];[\mathrm{L}^2(\Omega)]^d) \cap \mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)$ *or* $(\partial_t u_s, v_f) \in \mathrm{L}^2(0,T;V_\phi)$ *if* $\eta > 0$, *such that*

$$
\begin{cases}
\displaystyle\int_0^T\!\!\int_\Omega \rho_s(1-\phi)\,u_s\cdot\partial_{tt}^2 w_s\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\!\!\int_\Omega \sigma_s(u_s):\varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle +2\eta\int_0^T\!\!\int_\Omega \varepsilon(\partial_t u_s):\varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t - \int_0^T\!\!\int_\Omega \rho_f\phi\,v_f\cdot\partial_t w_f\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle +\int_0^T\!\!\int_\Omega \phi\,\sigma_f(v_f):\varepsilon(w_f)\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\!\!\int_\Omega \phi^2\,k_f^{-1}(v_f-\partial_t u_s)\cdot(w_f-w_s)\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle \quad -\int_0^T\!\!\int_\Omega \theta\,v_f\cdot w_f\,\mathrm{d}x\,\mathrm{d}t = \int_0^T\!\!\int_\Omega \rho_s(1-\phi)f\cdot w_s\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle \quad\quad +\int_0^T\!\!\int_\Omega \rho_f\phi f\cdot w_f\,\mathrm{d}x\,\mathrm{d}t + \int_\Omega \rho_s(1-\phi)\,v_{s0}\cdot w_s(0)\,\mathrm{d}x \\
\displaystyle \quad\quad\quad -\int_\Omega \rho_s(1-\phi)\,u_{s0}\cdot\partial_t w_s(0)\,\mathrm{d}x + \int_\Omega \rho_f\phi\,v_{f_0}\cdot w_f(0)\,\mathrm{d}x,
\end{cases}
\tag{1.87}
$$

and

$$
\begin{cases}
\displaystyle\int_0^T\!\!\int_\Omega \mathrm{div}\,\big((1-\phi)\partial_t u_s + \phi\,v_f\big)q\,\mathrm{d}x\,\mathrm{d}t = 0, \quad \text{if } \eta > 0, & \text{(1.88a)} \\
\displaystyle -\int_0^T\!\!\int_\Omega \mathrm{div}\,\big((1-\phi)\,u_s\big)\partial_t q\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle \quad\quad +\int_0^T\!\!\int_\Omega \mathrm{div}\,(\phi\,v_f)q\,\mathrm{d}x\,\mathrm{d}t = \int_\Omega \mathrm{div}\,\big((1-\phi)\,u_{s0}\big)q(0)\,\mathrm{d}x, \quad \text{if } \eta = 0, & \text{(1.88b)}
\end{cases}
$$

for all admissible test functions

$$
\begin{cases}
w_s \in \mathrm{H}^2(0,T;[\mathrm{L}^2(\Omega)]^d), \\
w_f \in \mathrm{H}^1(0,T;[\mathrm{L}^2(\Omega)]^d), \\
(w_s, w_f) \in \mathrm{L}^2(0,T;V_\phi), \\
q \in \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega)), \\
w_s(T) = \partial_t w_s(T) = w_f(T) = q(T) = 0.
\end{cases}
\tag{1.89}
$$

Proof. Let us prove this result in the inviscid case $\eta = 0$, the viscous case being similar. Since we can lift the mixture's divergence constraint as in (1.83), let us consider the case where the right-hand

side of the pressure equation is equal to zero. The resulting energy estimate reads

$$\begin{aligned}
&\frac{\rho_s}{2}\int_\Omega (1-\phi)|\partial_t u_s^\kappa(t)|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \sigma_s(u_s^\kappa(t)) : \varepsilon(u_s^\kappa(t))\,\mathrm{d}x \\
&\qquad + 2\eta\int_0^t\int_\Omega \varepsilon(\partial_t u_s^\kappa) : \varepsilon(\partial_t u_s^\kappa)\,\mathrm{d}x\,\mathrm{d}s + \frac{\rho_f}{2}\int_\Omega \phi|v_f^\kappa(t)|^2\,\mathrm{d}x \\
&\quad + \int_0^t\int_\Omega \phi\,\sigma_f(v_f^\kappa) : \varepsilon(v_f^\kappa)\,\mathrm{d}x\,\mathrm{d}s + \int_0^t\int_\Omega \phi^2 k_f^{-1}(v_f^\kappa - \partial_t u_s^\kappa)\cdot(v_f^\kappa - \partial_t u_s^\kappa)\,\mathrm{d}x\,\mathrm{d}s \\
&\qquad + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}|p^\kappa(t)|^2\,\mathrm{d}x \le \exp\left(\max\left(1, \frac{2\|\theta\|_{\mathrm{C}^0([0,T]\times\Omega)}}{\rho_f\phi_{\min}}\right)t\right) \\
&\quad \times\left(\left(\frac{\rho_s}{2}(1-\phi_{\min}) + \frac{\rho_f}{2}\phi_{\max}\right)\int_0^t\int_\Omega |f|^2\,\mathrm{d}x\,\mathrm{d}s + \frac{\rho_s}{2}\int_\Omega (1-\phi)|v_{s0}|^2\,\mathrm{d}x\right. \\
&\qquad \left. + \frac{1}{2}\int_\Omega \sigma_s(u_{s0}) : \varepsilon(u_{s0})\,\mathrm{d}x + \frac{\rho_f}{2}\int_\Omega \phi|v_{f_0}|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\alpha-\phi}{\kappa}|p_0|^2\,\mathrm{d}x\right). \qquad (1.90)
\end{aligned}$$

We deduce that, up to subsequences, the following weak convergences hold true as κ goes to infinity:

$$\begin{aligned}
u_s^\kappa &\rightharpoonup u_s^\infty \text{ weakly star in } \mathrm{L}^\infty(0,T;[\mathrm{H}_0^1(\Omega)]^d),\\
\partial_t u_s^\kappa &\rightharpoonup \partial_t u_s^\infty \text{ weakly star in } \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d),\\
v_f^\kappa &\rightharpoonup v_f^\infty \text{ weakly star in } \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d),\\
v_f^\kappa &\rightharpoonup v_f^\infty \text{ weakly in } \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d),
\end{aligned}$$

for some elements $u_s^\infty \in \mathrm{L}^\infty(0,T;[\mathrm{H}_0^1(\Omega)]^d)$ with $\partial_t u_s^\kappa \in \mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d)$ and v_f^∞ belonging to the space $\mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d)\cap \mathrm{L}^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)$.

We have no bound on the pressure p^κ but (1.90) implies that $\frac{p^\kappa}{\sqrt{\kappa}}$ is bounded in $\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$. Hence, we can select a subsequence (still denoted by p^κ) such that $\frac{p^\kappa}{\sqrt{\kappa}}$ converges in the weak–* topology of $\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$.

By adding (1.84a) to (1.84b) and by restricting the velocities test functions (w_s, w_f) to functions in $\mathrm{L}^2(0,T;V_\phi)$, it follows that $(u_s^\kappa, v_f^\kappa, p^\kappa)$ satisfies

$$\begin{cases}
\displaystyle\int_0^T\int_\Omega \rho_s(1-\phi)\,u_s^\kappa\cdot\partial_{tt}^2 w_s\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\Omega \sigma_s(u_s^\kappa) : \varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle + 2\eta\int_0^T\int_\Omega \varepsilon(\partial_t u_s^\kappa) : \varepsilon(w_s)\,\mathrm{d}x\,\mathrm{d}t - \int_0^T\int_\Omega \rho_f\phi\,v_f^\kappa\cdot\partial_t w_f\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle + \int_0^T\int_\Omega \phi\,\sigma_f(v_f^\kappa) : \varepsilon(w_f)\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\Omega \phi^2\,k_f^{-1}(v_f^\kappa - \partial_t u_s^\kappa)\cdot(w_f - w_s)\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle - \int_0^T\int_\Omega \theta\,v_f^\kappa\cdot w_f\,\mathrm{d}x\,\mathrm{d}t = \int_0^T\int_\Omega \rho_s(1-\phi)f\cdot w_s\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\Omega \rho_f\phi f\cdot w_f\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle \qquad + \int_\Omega \rho_s(1-\phi)\,v_{s0}\cdot w_s(0)\,\mathrm{d}x - \int_\Omega \rho_s(1-\phi)\,u_{s0}\cdot\partial_t w_s(0)\,\mathrm{d}x \\
\displaystyle \qquad + \int_\Omega \rho_f\phi\,v_{f_0}\cdot w_f(0)\,\mathrm{d}x, \\
\displaystyle - \int_0^T\int_\Omega \frac{\alpha-\phi}{\kappa}p^\kappa\,\partial_t q\,\mathrm{d}x\,\mathrm{d}t - \int_0^T\int_\Omega \mathrm{div}\left((\alpha-\phi)\,u_s^\kappa\right)\partial_t q\,\mathrm{d}x\,\mathrm{d}t \\
\displaystyle + \int_0^T\int_\Omega \mathrm{div}\,(\phi\,v_f^\kappa)q\,\mathrm{d}x\,\mathrm{d}t = \int_\Omega \frac{\alpha-\phi}{\kappa}p_0 q(0)\,\mathrm{d}x + \int_\Omega \mathrm{div}\left((\alpha-\phi)\,u_{s0}\right)q(0)\,\mathrm{d}x,
\end{cases}$$

for all admissible test functions verifying (1.89).

Thus we can pass to the weak limit in this formulation by noting that $\alpha - \phi \longrightarrow 1 - \phi$,

$$\int_0^T \int_\Omega \frac{\alpha - \phi}{\kappa} p^\kappa \, \partial_t q \, \mathrm{d}x \, \mathrm{d}t = \frac{1}{\sqrt{\kappa}} \int_0^T \int_\Omega (\alpha - \phi) \frac{p^\kappa}{\sqrt{\kappa}} \, \partial_t q \, \mathrm{d}x \, \mathrm{d}t \longrightarrow 0,$$

and that

$$\int_\Omega \frac{\alpha - \phi}{\kappa} p_0 q(0) \, \mathrm{d}x \longrightarrow 0,$$

as κ goes to infinity.

To conclude, (u_s^∞, v_f^∞) satisfies exactly (1.87) and (1.88b) in the incompressible limit. Moreover u_s^∞, $\partial_t u_s^\infty$ and v_f^∞ are continuous functions in time because they also satisfy (1.80) and hence coincide with the mild solution. Indeed, (1.80) can be recovered from (1.87) – (1.88b) by taking admissible test functions of the form $(w_s, w_f, q) = \big(\hat{w}_s(x), \hat{w}_f(x), \hat{q}(x)\big)\,\psi(t)$ with $(\hat{w}_s, \hat{w}_f, \hat{q}) \in V_\phi \times \mathrm{L}^2(\Omega)$ and $\psi \in \mathcal{D}(0,T)$. □

Remark 1.42. In the case where the right-hand side of the pressure equation is not equal to zero, we need to perform a lifting. Note that without this lifting step the energy estimate does not provide a uniform bound in κ as κ goes to infinity because of the coefficient $\frac{\kappa}{2\rho_f^2(\alpha-\phi)_{\min}}$ appearing in the right-hand side of (1.11). Moreover once the lifting is performed under assumptions $(h5)_{\text{bis}}$ and $(h6)_{\text{bis}}$, the new right-hand sides of the structure and fluid equations depend on α. Yet, it is easy to verify that they converge strongly in the proper spaces ensuring the convergence of the right-hand sides as α goes to one in $\mathrm{H}^{d/2+r}(\Omega)$ with $d/2 + r \geq 2$.

Remark 1.43. Theorem 1.41 provides the weak convergence of the displacement and velocities in the incompressible limit. If the incompressible regime solution is more regular, we can also obtain the pressure convergence and recover strong convergence for the displacement and velocities. More precisely, following the same guidelines as in [Ern and Guermond, 2021b, Lemma 75.1], we can show that

$$\begin{aligned}
\big\|u_s^\kappa - u_s^\infty\big\|^2_{\mathrm{L}^\infty(0,T;[\mathrm{H}^1_0(\Omega)]^d)} + \big\|\partial_t u_s^\kappa - \partial_t u_s^\infty\big\|^2_{\mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d)} \\
+ \big\|v_f^\kappa - v_f^\infty\big\|^2_{\mathrm{L}^\infty(0,T;[\mathrm{L}^2(\Omega)]^d)} + \frac{1}{\kappa}\big\|p^\kappa - p^\infty\big\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))} \\
+ \eta\big\|\partial_t u_s^\kappa - \partial_t u_s^\infty\big\|^2_{\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)} + \big\|v_f^\kappa - v_f^\infty\big\|^2_{\mathrm{L}^2(0,T;[\mathrm{H}^1_0(\Omega)]^d)} \\
\lesssim \frac{1}{\kappa^2}\big\|\partial_t p^\infty\big\|^2_{\mathrm{H}^1(0,T;\mathrm{L}^2(\Omega))}.
\end{aligned}$$

Thus, if $\partial_t p^\infty \in \mathrm{H}^1(0,T;\mathrm{L}^2(\Omega))$, then the above error estimate specifies the convergence speed of $(u_s^\kappa, \partial_t u_s^\kappa, v_f^\kappa, p^\kappa)$ towards $(u_s^\infty, \partial_t u_s^\infty, v_f^\infty, p^\infty)$ as κ goes to infinity.

1.5 Numerical experiments

In this section, we present some numerical examples to illustrate the theoretical results presented earlier. In particular, we numerically investigate the regularity of the solutions to the static and time-dependent problems. Note that there is an extensive literature on the numerical approximation of Biot-type systems, see [Russell and Wheeler, 1983; Zienkiewicz and Shiomi, 1984; Wheeler and Yotov, 2006; Phillips and Wheeler, 2008; Markert et al., 2009; Mikelić and Wheeler, 2013; Wheeler et al., 2014; Oyarzúa and Ruiz-Baier, 2016; Yi, 2017; Both et al., 2017; Lee, 2018; Storvik et al., 2019] and references therein. The numerical analysis of the specific model (1.2) presented in this work was performed in [Burtschell et al., 2019; Barnafi et al., 2021], where the time discretization

is performed with a monolithic backward Euler scheme. Moreover, an alternating minimization splitting scheme was proposed in [Both et al., 2022], which leads to a solver closely related to the undrained and fixed-stress splits of Biot's equations. We follow here the monolithic scheme of [Burtschell et al., 2019; Barnafi et al., 2021]. In addition, all simulations in this section were performed using the FEniCS finite element library [Logg et al., 2012; Alnæs et al., 2015].

1.5.1 Spatial discretization

For small values of bulk modulus, our equations can be discretized with standard finite elements. However, when the coefficient κ becomes large, we have to take into account the saddle-point structure of the problem involving the mixture's divergence constraint and to chose finite element spaces that satisfy the inf-sup condition (1.62) at the discrete level.

In the incompressible or nearly incompressible case, the expression of the mapping

$$\mathtt{T} : (u_s, v_s, v_f, p) \longmapsto \Big(\frac{\alpha^*}{2} u_s - \frac{\alpha^*}{2\lambda_0} v_s, \alpha^* v_s - v_p, \alpha^* v_f - v_p, \alpha^* p\Big),$$

defined in (1.79) suggests us how to select convenient finite element spaces in order to discretize the problem. Indeed, to get a stable discretization, it is sufficient to reproduce the construction of v_p at the discrete level. This is possible by choosing finite elements that are stable (in the Brezzi [Boffi et al., 2013] sense) for Stokes equations.

More precisely, let us suppose that we use a *conforming* approximation of the space $Y = [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times \mathrm{L}^2(\Omega)$ by a finite dimensional space

$$Y_h = V_{s,h} \times V_{s,h} \times V_{f,h} \times Q_h \subset Y,$$

where $V_{s,h}$, $V_{f,h}$ and Q_h denote respectively the finite element spaces chosen to discretize the solid part, the fluid part and the pressure of the mixture. Assume further that $(V_{s,h}, Q_h)$ and $(V_{f,h}, Q_h)$ are two inf-sup stable pairs associated with the standard Stokes problem and verify Fortin Lemma [Boffi et al., 2013, Proposition 5.4.2], *i.e.* there exists two operators $\Pi_{s,h} : [\mathrm{H}^1_0(\Omega)]^d \mapsto V_{s,h}$ and $\Pi_{f,h} : [\mathrm{H}^1_0(\Omega)]^d \mapsto V_{f,h}$ satisfying, for each $v \in [\mathrm{H}^1_0(\Omega)]^d$,

- For all $q_h \in Q_h$,
$$\int_\Omega \operatorname{div} v\, q_h \,\mathrm{d}x = \int_\Omega \operatorname{div}\big(\Pi_{i,h}(v)\big)\, q_h \,\mathrm{d}x, \quad i \in \{s, f\}; \tag{1.91}$$
- There exists a constant $C_{i,\pi} > 0$ independent of h such that
$$\|\nabla\big(\Pi_{i,h}(v)\big)\| \leq C_{i,\pi} \|\nabla v\|, \quad i \in \{s, f\}. \tag{1.92}$$

Under these hypotheses, we claim that the bilinear form a_{λ_0} defined in the mixed formulation (1.75) is uniformly $\mathtt{T}_h$-coercive. Namely, it holds

$$\forall h > 0, \exists \mathtt{T}_h \in \mathcal{L}(Y_h), \forall z_h \in Y_h, \quad |a_{\lambda_0}(z_h, \mathtt{T}_h z_h)| \geq \underline{\alpha} \|z_h\|^2_{Y_h} \quad \text{and} \quad |||\mathtt{T}_h||| \leq C, \tag{1.93}$$

for some constants $\underline{\alpha} > 0$ and $C > 0$ independent of h.

Indeed, setting

$$\mathtt{T}_h : (u_{s,h}, v_{s,h}, v_{f,h}, p_h) \longmapsto \Big(\frac{\alpha^*}{2} u_{s,h} - \frac{\alpha^*}{2\lambda_0} v_{s,h}, \alpha^* v_{s,h} - \Pi_{s,h}(v_{p_h}), \alpha^* v_{f,h} - \Pi_{f,h}(v_{p_h}), \alpha^* p_h\Big), \tag{1.94}$$

where v_{p_h} is defined by (1.63), the property (1.91) enables us to reproduce the calculations from the proof of Proposition 1.36 at the discrete level, so that the first condition of (1.93) holds true. The second condition then follows from (1.92) since

$$\|\nabla\big(\Pi_{i,h}(v_{p_h})\big)\| \leq C_{i,\pi} \|\nabla v_{p_h}\| \leq C_{i,\pi} C_{\mathrm{div}} \|p_h\|,$$

for each $i \in \{s, f\}$.

Therefore, a stable discretization of the incompressible system is offered by standard inf-sup stable conforming finite elements associated with the Stokes system. For instance, as it was observed in [Barnafi et al., 2021], one can use Taylor-Hood elements $[\mathbb{P}_{k+1}]^d - \mathbb{P}_k$ $(k \geq 1)$ for the pairs $(V_{s,h}, Q_h)$ and $(V_{f,h}, Q_h)$. More broadly, the previous $\mathtt{T}_h$-coercivity argument implies the stability of the MINI element $\mathbb{P}_1^b - \mathbb{P}_1$, the $\mathbb{P}_2 - \mathbb{P}_0$ element, or also Scott-Vogelius elements $[\mathbb{P}_k]^d - \mathbb{P}_{k-1}^{-1}$ with $k \geq 4$ and $d = 2$.

Finally, note that the mapping (1.94) is independent of the porosity ϕ and that the obtention of (1.93) does not require any assumption on the size of the permeability tensor k_f, as it was assumed in [Barnafi et al., 2021]. Hence, our approach provides a robust discretization regardless of porosity and permeability.

1.5.2 Regularity of the operator's domain

In both compressible and incompressible cases, we proved the existence and uniqueness of a strong solution in $\mathrm{C}^0([0,T]; D(A_\eta^\kappa))$, with $\eta \geq 0$ and $0 < \kappa \leq +\infty$. The operator's domain $D(A_\eta^\kappa)$ was defined by extension from a continuous bilinear form (see *e.g.* (1.23) and (1.24)) but we did not express it as a standard Sobolev space. In what follows, we give some numerical evidences that the operator's domain is not regular, namely

$$D(A_{\eta\geq 0}^\kappa) \neq [\mathrm{H}^2(\Omega)]^d \cap [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}^2(\Omega)]^d \cap [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}^2(\Omega)]^d \cap [\mathrm{H}_0^1(\Omega)]^d \times \mathrm{H}^1(\Omega)$$

and

$$D(A_{\eta\geq 0}^\infty) \neq [\mathrm{H}^2(\Omega)]^d \cap [\mathrm{H}_0^1(\Omega)]^d \times \big([\mathrm{H}^2(\Omega)]^d \times [\mathrm{H}^2(\Omega)]^d \cap V_\phi\big) \times \mathrm{H}^1(\Omega).$$

To do so, we compute numerically the solution of the static problem $z + A_\eta^\kappa z = g$ with $g \in Z$ or $g \in H$, viz.

$$\begin{cases} u_s - v_s = g_u, \\ \rho_s(1-\phi)\, v_s - \operatorname{div}\big(\sigma_s(u_s)\big) - \operatorname{div}\big(\sigma_s^{\mathrm{vis}}(v_s)\big) + (\alpha - \phi)\nabla p = \rho_s(1-\phi)\, g_s, \\ \rho_f \phi\, v_f - \operatorname{div}\big(\phi\, \sigma_f(v_f)\big) + \phi\, \nabla p = \rho_f \phi\, g_f, \\ \dfrac{\alpha - \phi}{\kappa} p + \operatorname{div}\big((\alpha - \phi)\, v_s + \phi\, v_f\big) = \begin{cases} \dfrac{\alpha - \phi}{\kappa}\, g_p & \text{if } \kappa < +\infty, \\ 0 & \text{if } \kappa = +\infty, \end{cases} \end{cases} \tag{1.95}$$

with $(g_u, g_s, g_f, g_p) \in [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times \mathrm{L}^2(\Omega)$ and where we have assumed that $k_f = 0$ and $\theta = 0$ without loss of generality.

We consider $\Omega = \{x \in \mathbb{R}^2,\ |x| \leq 1\}$ a very smooth domain and $(\mathcal{T}_h)_h$ a regular family of meshes of $\bar{\Omega}$, made of triangles. The coarsest mesh size H corresponds to a uniform mesh constructed with 8 subdivisions along each axis direction. Setting $\rho_s = \rho_f = \lambda_f = \mu_f = \eta = \lambda = \mu = 1$ and taking a constant porosity $\phi = 0.5$, we compute the error in L^2-norm between the approximated solution $(u_{s,h}, v_{s,h}, v_{f,h}, p_h)$ of (1.95) and a reference solution computed on a very refined mesh.

The resulting convergence graphs are presented in Figure 1.1 for $\kappa = 1$ and smooth data g_s, g_f, g_p. The convergence rates depend on the regularity of the solid displacement data g_u. If g_u is smooth, we obtain optimal orders of convergence, as expected by the theory. However, if $g_u \in [\mathrm{H}_0^1(\Omega)]^d \backslash [\mathrm{H}^2(\Omega)]^d$, Figure 1.1 (right) exhibits suboptimal convergence rates, thus indicating that the solution of (1.95) is not in H^2. The same occurs in the incompressible case, as shown in Figure 1.2.

Remark 1.44. In most applications, we have $g_u = 0$ so that the solution of the static problem (1.95) may me more regular than suggested by Figures 1.1 and 1.2. Yet, if $\kappa < +\infty$ or if $\eta = 0$, we only know that $\operatorname{div}\big((1-\phi)v_s\big) \in \mathrm{L}^2(\Omega)$, while the usual regularity results for Stokes-like systems

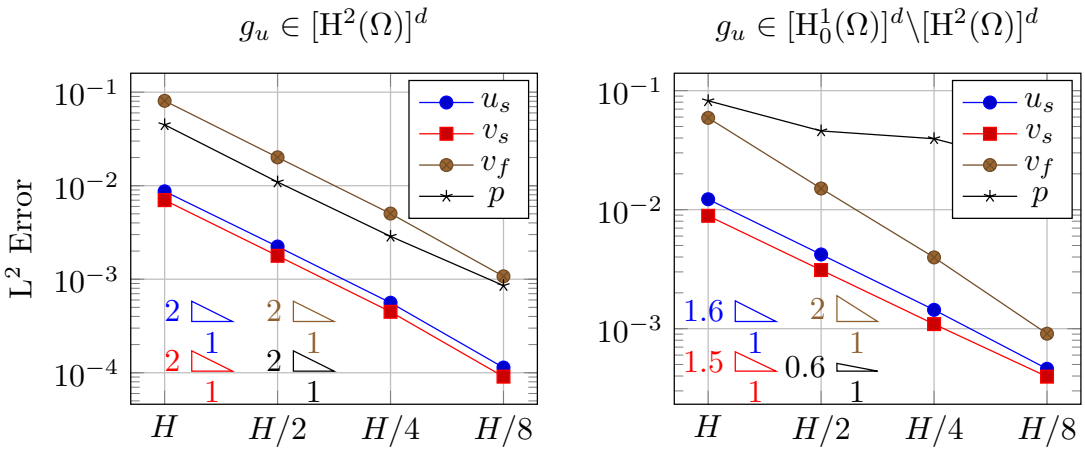


Figure 1.1 – Approximation errors and computed convergence rates for the discretization of the compressible ($\kappa = 1$) steady-state problem (1.95) with $[\mathbb{P}_1]^2 \times [\mathbb{P}_1]^2 \times [\mathbb{P}_1]^2 \times \mathbb{P}_1$ elements.

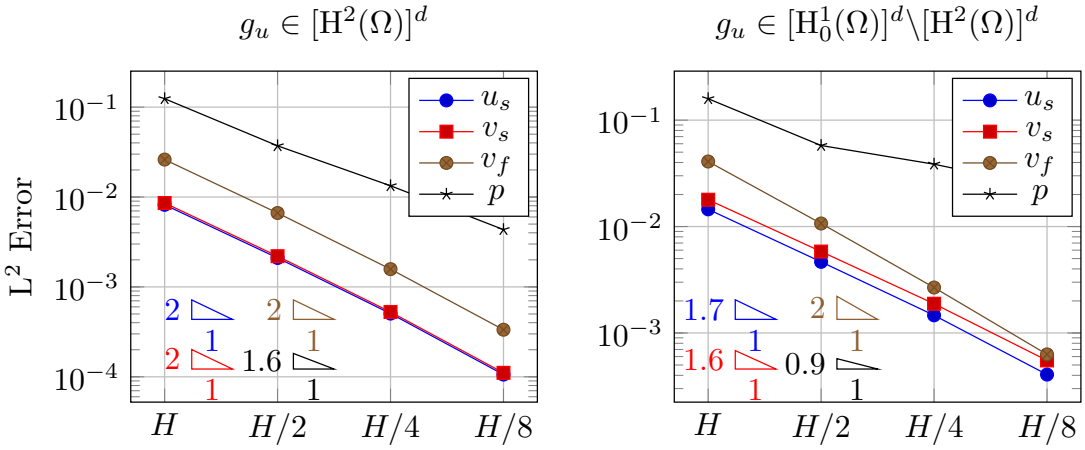


Figure 1.2 – Approximation errors and computed convergence rates for the discretization of the incompressible ($\kappa = +\infty$) steady-state problem (1.95) with $\mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1$ elements.

require a divergence right-hand side in $\mathrm{H}^1(\Omega)$. Therefore, our conjecture is that the only case in which the full regularity $[\mathrm{H}^2(\Omega)]^d \cap [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}^2(\Omega)]^d \cap [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}^2(\Omega)]^d \cap [\mathrm{H}_0^1(\Omega)]^d \times \mathrm{H}^1(\Omega)$ is recovered is the viscous incompressible case $\eta > 0$ and $\kappa = +\infty$, provided that the porosity ϕ is smooth enough.

1.5.3 Regularity of solutions

Even if the operator's domain is not regular, we are now going to shed light on the regularizing effect for the solution of the unsteady problem. For this purpose, we use the time discretization introduced and fully analyzed in [Burtschell et al., 2019], which consists of a midpoint scheme for the solid fields and an implicit backward Euler scheme for the fluid and the pressure. The major interest of this scheme is that it preserves energy balance at the discrete level.

As before, we perform the simulation on the smooth domain $\Omega = \{x \in \mathbb{R}^2, |x| \leq 1\}$ meshed by a regular family of triangulations $(\mathcal{T}_h)_h$. We set $\rho_s = \rho_f = \lambda_f = \mu_f = \lambda = \mu = 1$, $k_f = 0$, $\theta = 0$ and $\phi = 0.5$. We take $\kappa = 10^{10}$, but smallest values of κ would lead to comparable results. All the simulations are run during a hundred of time iterations, with a time step $\Delta t = 10^{-2}$ and up to the

final time $T = 1$.

Figure 1.3 illustrates the possible regularizing effect of time. Indeed, although the initial velocities and the applied exterior body force belong to $[\mathrm{H}_0^1(\Omega)]^d \backslash [\mathrm{H}^2(\Omega)]^d$, we recover optimal convergence rates in $\mathrm{L}^\infty(0,T;Z)$-norm between the approximated and reference solutions. Indeed, when using $\mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1$ elements for spatial discretization, the optimal convergence rate in this norm is equal to 1 for the displacement and to 2 for the other quantities since $Z = [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times \mathrm{L}^2(\Omega)$. Note that putting $\eta = 0$ slightly degrades the convergence order of the solid velocity, but does not affect the regularity of solid displacement, fluid velocity and pressure.

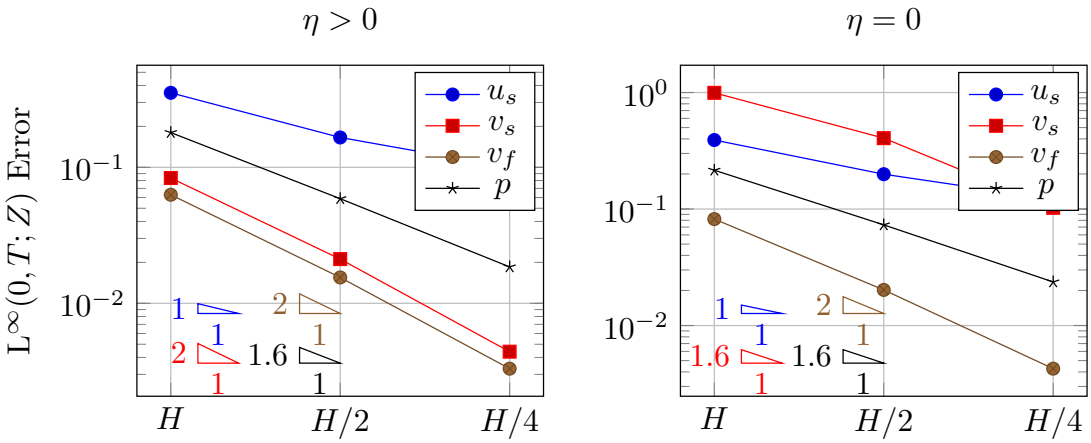


Figure 1.3 – Approximation errors and convergence rates for the discretization of the time-dependent problem (1.59) with $\mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1$ elements and v_{s0}, v_{f0}, f in $[\mathrm{H}_0^1(\Omega)]^d \backslash [\mathrm{H}^2(\Omega)]^d$.

If the initial conditions and the right-hand side belong strictly to the energy space, Figure 1.4 highlights that the convergence rates are considerably diminished, which confirms the regularity found in Theorems 1.32 and 1.39. Moreover, the solution is less regular when $\eta = 0$, as predicted by our theoretical results.

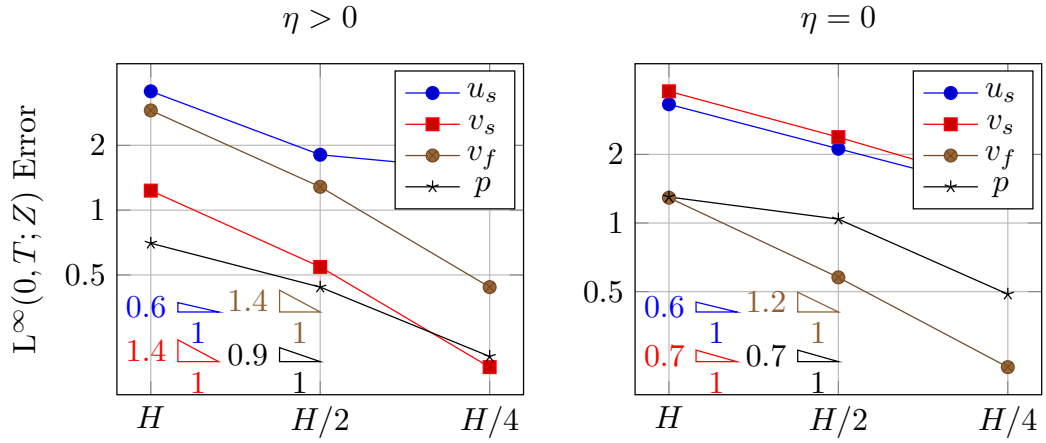


Figure 1.4 – Approximation errors and convergence rates for the discretization of the time-dependent problem (1.59) with $\mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1^b \times \mathbb{P}_1$ elements and v_{s0}, v_{f0}, f in $[\mathrm{L}^2(\Omega)]^d \backslash [\mathrm{H}_0^1(\Omega)]^d$.

Conclusion

In this work, we study the well-posedness for a fully unsteady and strongly coupled poromechanics model. Using an original combination of semigroup theory and T-coercivity, we demonstrated the existence and uniqueness of strong solutions in the compressible and incompressible cases, with or without solid viscosity. By unifying semigroup and variational techniques, we also recovered the existence and uniqueness of weak solutions. From the methodological point of view, this unified approach enabled us to take benefit of the best of both theories depending on the hyperbolic-parabolic or parabolic-parabolic nature of the coupling, in particular to prove uniqueness. To handle the incompressible case, we developed a functional framework and an extension of De Rham Theorem adapted to the mixture's divergence constraint. When the incompressible regime is reached, our analysis offers a spatial discretization of the problem with conforming finite elements, including Taylor-Hood elements but also any Stokes-stable elements such as the MINI element. Moreover, this choice of finite element discretization spaces is robust with respect to porosity and permeability. Finally, our theoretical results are corroborated by numerical experiments. We have shown numerically that the operator's domain is not regular, and illustrate the possible regularizing effect on the unsteady problem.

CHAPTER 2

The T-coercivity method for mixed problems

This chapter reproduces results obtained in collaboration with Patrick Ciarlet (ENSTA Paris). The corresponding article is currently under review. In Chapter 1, the notion of T-coercivity was in particular used to deal with the incompressibility mixture divergence constraint. The goal of this chapter is to extend this concept to general saddle-point and perturbed saddle-point problems and to apply it to electromagnetism, nearly-incompressible elasticity and neutron diffusion.

Contents

The T-coercivity approach for mixed problems

Mathieu Barré[1,2] and Patrick Ciarlet[3]

[1] Inria, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
[2] LMS, École Polytechnique, CNRS, Institut Polytechnique de Paris
Route de Saclay, 91120 Palaiseau, France
[3] POEMS, CNRS, Inria, ENSTA Paris, Institut Polytechnique de Paris
828 Boulevard des Maréchaux, 91120 Palaiseau, France

Submitted.

Abstract

Classically, the well-posedness of variational formulations of mixed linear problems is achieved through the inf-sup condition on the constraint. In this note, we propose an alternative framework to study such problems by using the T-coercivity approach to derive a global inf-sup condition. This is a constructive approach that leads to the design of suitable approximations in a simple way. In general, the derivation of the uniform discrete inf-sup condition for the approximate problems stems straightforwardly from the study of the original problem. To support our view, we solve a series of classical mixed problems with the T-coercivity approach. Among others, the celebrated Fortin's Lemma appears naturally in the numerical analysis of the approximate problems.

Keywords — T-coercivity, inf-sup condition, Fortin's Lemma, perturbed saddle-point problems.
Mathematics Subject Classification (2020) — 65N30, 35J57, 76D07, 78M10.

Introduction

Traditionally, the well-posedness of variational formulations of mixed linear problems is achieved through the inf-sup condition, also called stability condition [Ladyzhenskaya, 1969; Babuška, 1973; Brezzi, 1974]. As a matter of fact, proving this condition allows to derive existence and uniqueness of the solution, and continuous dependence with respect to the data. On the other hand, the way this condition is established depends on the problem to be solved. The analysis of such problems can be performed either following a *monolithic* approach, namely studying the *all-in-one* bilinear form incorporating the constraint, or by studying the constrained part of the problem separately.

In this note, we focus on the monolithic approach and investigate the mixed problem's well-posedness based on the T-coercivity framework. The principle of this framework is to find an *explicit realization* of the inf-sup condition for the all-in-one bilinear form. Of equal importance, in the T-coercivity framework, is the *design of suitable approximations* of the original problem. Indeed, with the help of the explicit realization of the condition for the original problem, one can get useful insight on how to derive the so-called uniform discrete inf-sup condition for the approximate, or discrete, problems set in finite-dimensional vector spaces. Thus, convergence of the approximate solutions to the exact one follows under well-known principles in numerical analysis, such as Céa's Lemma (or a variant), and a basic approximability property of elements of the original space of solutions. To summarize, although the T-coercivity approach may not bring new result to the theory of variational formulations, it proposes a compact way to study them theoretically and also on how to approximate them.

So far, the T-coercivity approach has been mainly applied to two categories of linear problems. First, for problems involving an invertible operator and a compact perturbation, see eg. [Hiptmair, 2002; Buffa et al., 2002; Buffa and Christiansen, 2003; Buffa, 2005; Ciarlet Jr, 2012; Sayas et al., 2019]. Then, for problems with sign-changing coefficients, cf. [Bonnet-Ben Dhia et al., 2010b; Nicaise and Venel, 2011; Bonnet-Ben Dhia et al., 2012; Chesnel and Ciarlet, 2013; Bonnet-Ben Dhia et al., 2013; Bonnet-Ben Dhia et al., 2018, 2014b,a; Bunoiu and Ramdani, 2016; Chesnel, 2016; Bunoiu et al., 2020; Ciarlet Jr, 2020; Halla, 2021; Bunoiu et al., 2021; Ciarlet Jr, 2022]. For the second category, we observe that well-posedness and (efficient) approximation of the variational formulations has actually been achieved with the help of the T-coercivity approach. Up to the authors' knowledge, this approach was only applied to mixed problems in [Jamelot and Ciarlet Jr, 2013; Hong et al., 2023]: in the first reference, it is applied to the specific case of neutron diffusion, whereas the second one focuses on perturbed saddle-point problems.

In this note, we apply the T-coercivity approach to general mixed problems, including unperturbed and perturbed saddle-point problems. In particular, we will explain the connections with the classical theory, for which we use [Boffi et al., 2013] as the reference textbook. Among those connections, we note that the celebrated Fortin's Lemma will appear naturally in the (numerical) analysis of the discrete problems.

Let us introduce some notation. Given a Hilbert space V, we denote by $(\cdot,\cdot)_V$ and $\|\cdot\|_V$ the inner product and the norm on V, and by V' its dual space. In a product space $V\times W$, we use the norm

$$\|(v,w)\|_{V\times W}=\big(\|v\|_V^2+\|w\|_W^2\big)^{1/2},$$

and similarly for the inner product. Vector-valued function spaces are written in boldface character. A connected, bounded, open subset of $\mathbb{R}^d$ with a Lipschitz boundary is called a *domain*.

Let Ω be a domain with boundary $\partial\Omega$. We denote by $\boldsymbol{n}$ the unit outward normal vector field to $\partial\Omega$. Let $\mathrm{L}^2(\Omega)$ and $\boldsymbol{L}^2(\Omega)$ be the set of square-integrable real-valued and $\mathbb{R}^d$-valued functions on Ω. The natural norm in $L^2(\Omega)$ or $\boldsymbol{L}^2(\Omega)$ is denoted by $\|\cdot\|$, and we let

$$L^2_0(\Omega)=\left\{v\in L^2(\Omega)\,,\ \int_\Omega v\,\mathrm{d}x=0\right\}.$$

In what follows, unless otherwise stated, the standard Sobolev space $H^1_0(\Omega)$ is endowed with the norm $v\mapsto\|\nabla v\|$, that defines a norm that is equivalent to $\|\cdot\|_{H^1(\Omega)}$ thanks to Poincaré's inequality. The dual space of $H^1_0(\Omega)$ is denoted by $H^{-1}(\Omega)$. Similarly, $\boldsymbol{H}^1_0(\Omega)$ is endowed with the norm $\boldsymbol{v}\mapsto\big(\sum_{i=1,d}\|\nabla v_i\|^2\big)^{1/2}$, that defines a norm that is equivalent to $\|\cdot\|_{[\mathrm{H}^1(\Omega)]^d}$, and its dual space is denoted by $\boldsymbol{H}^{-1}(\Omega)$. We introduce the usual Sobolev spaces for vector-valued fields [Assous et al., 2018]

$$\begin{aligned}
\boldsymbol{H}(\mathrm{div}\,;\Omega)&=\big\{\boldsymbol{v}\in\boldsymbol{L}^2(\Omega),\ \mathrm{div}\,\boldsymbol{v}\in\mathrm{L}^2(\Omega)\big\},\\
\boldsymbol{H}_0(\mathrm{div}\,;\Omega)&=\big\{\boldsymbol{v}\in\boldsymbol{H}(\mathrm{div}\,;\Omega),\ \boldsymbol{v}\cdot\boldsymbol{n}=0\text{ on }\partial\Omega\big\},\\
\boldsymbol{H}(\mathrm{div}\,0;\Omega)&=\big\{\boldsymbol{v}\in\boldsymbol{H}(\mathrm{div}\,;\Omega),\ \mathrm{div}\,\boldsymbol{v}=0\big\},\\
\boldsymbol{H}(\mathbf{curl}\,;\Omega)&=\big\{\boldsymbol{v}\in\boldsymbol{L}^2(\Omega),\ \mathbf{curl}\,\boldsymbol{v}\in\boldsymbol{L}^2(\Omega)\big\},\qquad\text{for } d=3,\\
\boldsymbol{H}_0(\mathbf{curl}\,;\Omega)&=\big\{\boldsymbol{v}\in\boldsymbol{H}(\mathbf{curl}\,;\Omega),\ \boldsymbol{v}\times\boldsymbol{n}=0\text{ on }\partial\Omega\big\},\qquad\text{for } d=3.
\end{aligned}$$

Unless otherwise specified, $\boldsymbol{H}(\mathrm{div}\,;\Omega)$ is endowed with the norm $\boldsymbol{v}\mapsto\big(\|\boldsymbol{v}\|^2+\|\mathrm{div}\,\boldsymbol{v}\|^2\big)^{1/2}$ and $\boldsymbol{H}(\mathbf{curl}\,;\Omega)$ with the norm $\boldsymbol{v}\mapsto\big(\|\boldsymbol{v}\|^2+\|\mathbf{curl}\,\boldsymbol{v}\|^2\big)^{1/2}$.

The outline is as follows. In Section 2.1, we introduce the T-coercivity approach, and explain how it can be applied to solve the Stokes problem theoretically. Then, in Section 2.2, we develop the

abstract framework underlying the approach for mixed problems, including saddle-point, augmented and perturbed ones. In Sections 2.3, 2.4 and 2.5, we propose some applications, respectively to electromagnetism, nearly-incompressible elasticity, and diffusion. Then, in Section 2.6, we propose the *natural* extension of the T-coercivity approach for the conforming approximation of mixed problems. As before, we begin by the Stokes problem, then we consider the numerical analysis for mixed problems in general, before describing how the approach can be applied to electromagnetism, nearly-incompressible elasticity, and diffusion. We conclude by a list of further extensions and recent applications of the T-coercivity approach.

2.1 T-coercivity for the Stokes problem

The starting point of our study is to propose a T-coercivity approach to solve Stokes problem. Let $\Omega \subset \mathbb{R}^d$ be a domain. We consider the Stokes problem with homogeneous Dirichlet boundary conditions: given a prescribed body force $\boldsymbol{f} \in \boldsymbol{H}^{-1}(\Omega)$, find the velocity $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega)$ and the pressure $p \in \mathrm{L}^2_0(\Omega)$ such that

$$\begin{aligned} -\nu\Delta\boldsymbol{u} + \nabla p &= \boldsymbol{f}, \quad && \text{in } \Omega, \\ \operatorname{div}\boldsymbol{u} &= 0, && \text{in } \Omega, \\ \boldsymbol{u} &= 0, && \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $\nu > 0$ denotes the fluid's viscosity.

The standard method to solve Problem (2.1) – see [Girault and Raviart, 1986] – consists in a *one-plus-one* approach. The problem is split into a coercive part

$$a(\boldsymbol{u},\boldsymbol{v}) = \nu \int_\Omega \nabla\boldsymbol{u} : \nabla\boldsymbol{v}\,\mathrm{d}x$$

and divergence constraint terms of the form

$$b(\boldsymbol{v},q) = -\int_\Omega q \operatorname{div}\boldsymbol{v}\,\mathrm{d}x,$$

so that the weak formulation of Problem (2.1) reads: find $(\boldsymbol{u},p) \in \boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)$ such that

$$\begin{aligned} a(\boldsymbol{u},\boldsymbol{v}) + b(\boldsymbol{v},p) &= \langle \boldsymbol{f},\boldsymbol{v}\rangle, \quad && \forall \boldsymbol{v} \in \boldsymbol{H}^1_0(\Omega), \\ b(\boldsymbol{u},q) &= 0, && \forall q \in \mathrm{L}^2_0(\Omega), \end{aligned} \tag{2.2}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality product in $\boldsymbol{H}^1_0(\Omega)$. The well-posedness of Problem (2.2) then follows from Ladyzhenskaya–Babuška–Brezzi's theory [Ladyzhenskaya, 1969; Babuška, 1973; Brezzi, 1974] since the bilinear form a is coercive on $\boldsymbol{H}^1_0(\Omega)$ and the bilinear form b satisfies the *inf-sup condition*

$$\inf_{q\in\mathrm{L}^2_0(\Omega)\setminus\{0\}} \sup_{\boldsymbol{v}\in\boldsymbol{H}^1_0(\Omega)\setminus\{0\}} \frac{b(\boldsymbol{v},q)}{\|\nabla\boldsymbol{v}\|\|q\|} \geq \underline{\beta} \tag{2.3}$$

for some constant $\underline{\beta} > 0$.

Here, we are going to give an alternative proof that Problem (2.1) is well-posed by analysing the *all-in-one* bilinear form defined on $\boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)$

$$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v},q)\big) = \nu \int_\Omega \nabla\boldsymbol{u} : \nabla\boldsymbol{v}\,\mathrm{d}x - \int_\Omega p \operatorname{div}\boldsymbol{v}\,\mathrm{d}x - \int_\Omega q \operatorname{div}\boldsymbol{u}\,\mathrm{d}x$$

instead of splitting it into two bilinear forms a and b as in (2.2). This bilinear form is not coercive since

$$\mathcal{A}\big((0,p),(0,p)\big) = 0, \quad \forall p \in \mathrm{L}^2_0(\Omega).$$

For this reason, we use the notion of T-coercivity [Ciarlet Jr, 2012; Chesnel and Ciarlet, 2013], which can be seen as a reformulation of Banach-Nečas-Babuška's theory [Ern and Guermond, 2021a, Theorem 25.9]. For an Hilbert space W, let $\mathcal{L}(W)$ denote the set of bounded operators on W. The definition and the main property of T-coercivity are recalled below.

Definition 2.1. [Chesnel and Ciarlet, 2013, Definition 3] Let W be a Hilbert space and let $\mathcal{A}(\cdot,\cdot)$ be a continuous bilinear form over $W \times W$. We say that $\mathcal{A}$ is T-*coercive* if there exists a bijective operator $\mathtt{T} \in \mathcal{L}(W)$ and $\underline{\alpha} > 0$ such that

$$|\mathcal{A}(u, \mathtt{T}u)| \geq \underline{\alpha}\|u\|_W^2, \quad \forall u \in W.$$

Proposition 2.2. *[Chesnel and Ciarlet, 2013, Theorem 1] Let W be a Hilbert space. Let $\ell(\cdot)$ be a continuous linear form over W and $\mathcal{A}(\cdot,\cdot)$ be a continuous bilinear form over $W \times W$. The problem*

$$\begin{cases} \text{Find } u \in W \quad \text{such that} \\ \forall v \in W, \quad \mathcal{A}(u,v) = \ell(v) \end{cases}$$

is well-posed if and only if $\mathcal{A}$ is T*-coercive. If so, it holds that*

$$\|u\|_W \leq \frac{C_\ell}{\underline{\alpha}} |||\mathtt{T}|||, \tag{2.4}$$

with C_ℓ the continuity constant of the linear form ℓ and

$$|||\mathtt{T}||| = \sup_{v \in W \setminus \{0\}} \frac{\|\mathtt{T}v\|_W}{\|v\|_W}$$

When the bilinear form $\mathcal{A}(\cdot,\cdot)$ is in addition symmetric, the requirement that the operator T *is bijective can be dropped.*

2.1.1 Proving well-posedness with T-coercivity

With the T-coercivity tool in mind, we are now ready to establish the main result of this section. To that aim, we use the result below, see for instance [Girault and Raviart, 1986, Corollary I.2.4]. Let $q \in \mathrm{L}^2_0(\Omega)$. Then, there exists $\boldsymbol{v}_q \in \boldsymbol{H}^1_0(\Omega)$ satisfying

$$-\operatorname{div} \boldsymbol{v}_q = q. \tag{2.5}$$

In addition, there exists a constant $C_{\mathrm{div}} > 0$ independent of q such that

$$\|\nabla \boldsymbol{v}_q\| \leq C_{\mathrm{div}} \|q\|. \tag{2.6}$$

Theorem 2.3. *The problem*

$$\begin{cases} \text{Find } (\boldsymbol{u}, p) \in \boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega) \quad \text{such that} \\ \forall (\boldsymbol{v}, q) \in \boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega), \quad \mathcal{A}\big((\boldsymbol{u}, p), (\boldsymbol{v}, q)\big) = \langle \boldsymbol{f}, \boldsymbol{v} \rangle \end{cases} \tag{2.7}$$

is well-posed and

$$\|(\boldsymbol{u}, p)\|_{\boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)} \leq \frac{2 \max\Big(\sqrt{2}\nu C_{\mathrm{div}}^2, C_{\mathrm{div}}\big(2 + \nu^2 C_{\mathrm{div}}^2\big)^{1/2}\Big)}{\min(\nu^2 C_{\mathrm{div}}^2, 1)} \|\boldsymbol{f}\|_{\boldsymbol{H}^{-1}(\Omega)}. \tag{2.8}$$

Proof. The linear form defined by

$$\ell\big((\boldsymbol{v},q)\big) = \langle \boldsymbol{f}, \boldsymbol{v}\rangle, \quad \forall (\boldsymbol{v},q) \in \boldsymbol{H}^1_0(\Omega), \times \mathrm{L}^2_0(\Omega)$$

is continuous over $\boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)$ in view of the inequality

$$\ell\big((\boldsymbol{v},q)\big) \leq \|\boldsymbol{f}\|_{\boldsymbol{H}^{-1}(\Omega)} \|(\boldsymbol{v},q)\|_{\boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega)}. \tag{2.9}$$

The bilinear form $\mathcal{A}$ is continuous over $\big(\boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)\big)^2$ and we observe that it is also symmetric.

Then, from Proposition 2.2, it is sufficient to show that the bilinear form $\mathcal{A}$ is T-coercive. For a given $(\boldsymbol{u},p) \in \boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)$, we look for an element $(\boldsymbol{v}^*,q^*)$ of $\boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)$ depending continously on $(\boldsymbol{u},p)$ and such that

$$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) \geq \underline{\alpha} \|(\boldsymbol{u},p)\|^2_{\boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega)}$$

for some constant $\underline{\alpha} > 0$. In order to get an intuitive idea of the construction of $(\boldsymbol{v}^*,q^*)$, let us start with specific elements $(\boldsymbol{u},p)$.

- If $p = 0$, then $\|(\boldsymbol{u},p)\|^2_{\boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega)} = \|\nabla \boldsymbol{u}\|^2$ and

  $$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) = \nu \int_\Omega \nabla \boldsymbol{u} : \nabla \boldsymbol{v}^* \,\mathrm{d}x - \int_\Omega \operatorname{div} \boldsymbol{u}\, q^* \,\mathrm{d}x,$$

  so that we can take $\boldsymbol{v}^* = \boldsymbol{u}$ and $q^* = p = 0$.

- If $\boldsymbol{u} = 0$, then $\|(\boldsymbol{u},p)\|^2_{\boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega)} = \|p\|^2$ and

  $$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) = -\int_\Omega p \operatorname{div} \boldsymbol{v}^* \,\mathrm{d}x.$$

  In order to recover the expected term $\|p\|^2$ in the above expression, we have to choose $\boldsymbol{v}^*$, the divergence of which is "as close as possible" to $-p$. The idea is now to choose $\boldsymbol{v}^* = \boldsymbol{v}_p$, where $\boldsymbol{v}_p$ is as in (2.5)-(2.6). Hence, taking $q^* = 0$, we find

  $$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) = \|p\|^2,$$

  and (2.6) ensures that the pair $(\boldsymbol{v}_p, 0)$ depends continously on $(0,p)$ in $\boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)$.

- If $\operatorname{div} \boldsymbol{u} = 0$, then

  $$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) = \nu \int_\Omega \nabla \boldsymbol{u} : \nabla \boldsymbol{v}^* \,\mathrm{d}x - \int_\Omega p \operatorname{div} \boldsymbol{v}^* \,\mathrm{d}x.$$

  Since we need to get a term of the form $\|\nabla \boldsymbol{u}\|^2$ but also of the form $\|p\|^2$, we combine the previous two cases by setting $\boldsymbol{v}^* = \lambda \boldsymbol{u} + \boldsymbol{v}_p$, where λ is a positive coefficient to be adjusted and $\boldsymbol{v}_p$ is the divergence lifting from (2.5) – (2.6). Now, we compute

  $$\begin{aligned}\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) &= \nu\lambda \int_\Omega \nabla \boldsymbol{u} : \nabla \boldsymbol{u} \,\mathrm{d}x + \nu \int_\Omega \nabla \boldsymbol{u} : \nabla \boldsymbol{v}_p \,\mathrm{d}x - \lambda \int_\Omega p \operatorname{div} \boldsymbol{u} \,\mathrm{d}x - \int_\Omega p \operatorname{div} \boldsymbol{v}_p \,\mathrm{d}x \\ &= \nu\lambda \|\nabla \boldsymbol{u}\|^2 + \nu \int_\Omega \nabla \boldsymbol{u} : \nabla \boldsymbol{v}_p \,\mathrm{d}x + \|p\|^2\end{aligned}$$

  since $\operatorname{div} \boldsymbol{u} = 0$ and $-\operatorname{div} \boldsymbol{v}_p = p$. For all $\eta > 0$, Young's inequality implies that

  $$\begin{aligned}\int_\Omega \nabla \boldsymbol{u} : \nabla \boldsymbol{v}_p \,\mathrm{d}x &\geq -\frac{\eta}{2} \|\nabla \boldsymbol{u}\|^2 - \frac{1}{2\eta} \|\nabla \boldsymbol{v}_p\|^2 \\ &\geq -\frac{\eta}{2} \|\nabla \boldsymbol{u}\|^2 - \frac{C^2_{\mathrm{div}}}{2\eta} \|p\|^2 \quad \text{in virtue of (2.6)},\end{aligned}$$

and thus

$$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) \geq \nu\left(\lambda-\frac{\eta}{2}\right)\|\nabla\boldsymbol{u}\|^2+\left(1-\frac{\nu C_{\rm div}^2}{2\eta}\right)\|p\|^2.$$

Hence, by setting $\eta=\lambda=\nu C_{\rm div}^2$, we obtain

$$\mathcal{A}\big((\boldsymbol{u},p),(\boldsymbol{v}^*,q^*)\big) \geq \frac{\nu^2 C_{\rm div}^2}{2}\|\nabla\boldsymbol{u}\|^2+\frac{1}{2}\|p\|^2.$$

In the general case, we choose $\boldsymbol{v}^*=\lambda\boldsymbol{u}+\boldsymbol{v}_p$ with $\lambda=\nu C_{\rm div}^2$ and $q^*=-\lambda p$ so that, even if $\operatorname{div}\boldsymbol{u}\neq 0$, the term $-\lambda\int_\Omega p\operatorname{div}\boldsymbol{u}\,\mathrm{d}x$ cancels with the term $-\int_\Omega \operatorname{div}\boldsymbol{u}\,q^*\,\mathrm{d}x$ and we get the same results as in the case $\operatorname{div}\boldsymbol{u}=0$. Namely, the bilinear form $\mathcal{A}$ is T-coercive for the mapping

$$\begin{aligned}\mathtt{T}:\boldsymbol{H}_0^1(\Omega)\times \mathrm{L}_0^2(\Omega) &\longrightarrow \boldsymbol{H}_0^1(\Omega)\times \mathrm{L}_0^2(\Omega)\\ (\boldsymbol{u},p)&\longmapsto \big(\nu C_{\rm div}^2\boldsymbol{u}+\boldsymbol{v}_p,-\nu C_{\rm div}^2 p\big),\end{aligned}$$

where $\boldsymbol{v}_p$ is defined by (2.5) with estimate (2.6), and it holds that

$$\mathcal{A}\big((\boldsymbol{u},p),\mathtt{T}(\boldsymbol{u},p)\big) \geq \frac{\nu^2 C_{\rm div}^2}{2}\|\nabla\boldsymbol{u}\|^2+\frac{1}{2}\|p\|^2 \geq \frac{1}{2}\min(\nu^2 C_{\rm div}^2,1)\|(\boldsymbol{u},p)\|^2_{\boldsymbol{H}_0^1(\Omega)\times \mathrm{L}_0^2(\Omega)}. \tag{2.10}$$

Thanks to (2.5)-(2.6), T belongs to $\mathcal{L}(\boldsymbol{H}_0^1(\Omega)\times \mathrm{L}_0^2(\Omega))$. More precisely, we have

$$\begin{aligned}\|\mathtt{T}(\boldsymbol{u},p)\|^2_{\boldsymbol{H}_0^1(\Omega)\times \mathrm{L}_0^2(\Omega)} &= \|\nu C_{\rm div}^2\boldsymbol{u}+\boldsymbol{v}_p\|^2_{\boldsymbol{H}_0^1(\Omega)}+\|\nu C_{\rm div}^2 p\|^2\\ &\leq 2(\nu C_{\rm div}^2)^2\|\nabla\boldsymbol{u}\|^2+2\|\nabla\boldsymbol{v}_p\|^2+(\nu C_{\rm div}^2)^2\|p\|^2\\ &\leq 2(\nu C_{\rm div}^2)^2\|\nabla\boldsymbol{u}\|^2+\big(2C_{\rm div}^2+(\nu C_{\rm div}^2)^2\big)\|p\|^2\end{aligned}$$

and thus

$$|||\mathtt{T}||| \leq \max\Big(\sqrt{2}\nu C_{\rm div}^2, C_{\rm div}\big(2+\nu^2 C_{\rm div}^2\big)^{1/2}\Big). \tag{2.11}$$

Using (2.9), (2.10) and (2.11) in the stability estimate (2.4), we finally obtain (2.8). □

Remark 2.4. The previous result readily extends to the case of a non-null divergence constraint

$$\begin{aligned}-\nu\Delta\boldsymbol{u}+\nabla p&=\boldsymbol{f}, &&\text{in }\Omega,\\ \operatorname{div}\boldsymbol{u}&=g, &&\text{in }\Omega,\\ \boldsymbol{u}&=0, &&\text{on }\partial\Omega,\end{aligned}$$

with $g\in \mathrm{L}_0^2(\Omega)$, leading to the stability estimate

$$\|(\boldsymbol{u},p)\|_{\boldsymbol{H}_0^1(\Omega)\times \mathrm{L}_0^2(\Omega)} \leq \frac{2\max\Big(\sqrt{2}\nu C_{\rm div}^2, C_{\rm div}\big(2+\nu^2 C_{\rm div}^2\big)^{1/2}\Big)}{\min(\nu^2 C_{\rm div}^2,1)}\|(\boldsymbol{f},g)\|_{\boldsymbol{H}^{-1}(\Omega)\times \mathrm{L}_0^2(\Omega)}. \tag{2.12}$$

2.1.2 Comments

The stability estimates (2.8) and (2.12) are valid for all $C_{\rm div}$ that fulfills (2.6). On the other hand, one has

$$\lim_{C_{\rm div}\to\infty}\frac{2\max\Big(\sqrt{2}\nu C_{\rm div}^2, C_{\rm div}\big(2+\nu^2 C_{\rm div}^2\big)^{1/2}\Big)}{\min(\nu^2 C_{\rm div}^2,1)}=+\infty,$$

i.e. the stability estimates become meaningless for large $C_{\rm div}$.

Going through the proof of Theorem 2.3, we observe that the constant obtained in (2.8) and (2.12) is just one of the many bounds one can achieve with T-coercivity for the Stokes problem.

Indeed, the operator $\mathtt{T}$ is in general not unique. In particular, one can choose any positive value of λ, so that there exists a family of admissible operators $\mathtt{T}$ in the sense of Definition 2.1, which shows the flexibility of the approach.

Let us provide an illustration. For small viscosity ν (the domain Ω being fixed), it is well-known that the stability constant appearing in the estimate

$$\|(\boldsymbol{u},p)\|_{\boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega)} \leq C(\nu)\,\|(\boldsymbol{f},g)\|_{\boldsymbol{H}^{-1}(\Omega)\times \mathrm{L}^2_0(\Omega)}$$

behaves likes $O(\nu^{-1})$. For instance, for the velocity $\boldsymbol{u}$, the result is elementarily obtained by taking the test field $(\boldsymbol{v},q)=(\boldsymbol{u},p)$ in (2.2). On the other hand, in (2.8) and (2.12), we find a behavior in $O(\nu^{-2})$. But, if one is interested in obtaining a less severe blowup, one can simply choose

$$\eta = \frac{\nu C_{\mathrm{div}}^2}{2(1-\frac{\nu}{2})} \quad \text{and} \quad \lambda = \frac{1}{2}(1+\eta)$$

in the above proof, for all $0<\nu\leq 1$. Then, one finds that

$$\underline{\alpha} = \frac{\nu}{2} \quad \text{and} \quad |||\mathtt{T}||| \leq \max\Big(\frac{1}{\sqrt{2}}(1+C_{\mathrm{div}}^2), \big(2C_{\mathrm{div}}^2+\frac{1}{4}(1+C_{\mathrm{div}}^2)^2\big)^{1/2}\Big),$$

so that (2.4) actually yields a stability constant in $O(\nu^{-1})$.

Remark 2.5. Note that the dependence of the previous estimates on viscosity can be removed by using the scaled norm $\|\cdot\|_\nu$ defined by $\|\boldsymbol{u}\|_\nu = \nu\int_\Omega \nabla\boldsymbol{u}:\nabla\boldsymbol{v}\,\mathrm{d}x$ for all $\boldsymbol{u}\in\boldsymbol{H}^1_0(\Omega)$.

Theorem 2.3 provides a fully constructive proof for the well-posedness of Stokes problem, which is an emblematic example of mixed problem. In the next section, we show that the T-coercivity approach employed here is in fact very general and can be extended to a large class of saddle-point problems.

2.2 Abstract framework

We start with the classical statements regarding the definition of saddle-point problems, and the equivalent conditions to ensure an inf-sup condition on the constraint. Then, we proceed with the design of abstract operators $\mathtt{T}$ to ensure well-posedness for saddle-point problems, and for augmented saddle-point problems.

2.2.1 Saddle-point problems in Hilbert spaces

Let V and Q be two Hilbert spaces. In the Hilbert space Q, we introduce the canonical isomorphism $\mathbb{1}_{Q\to Q'}: Q\to Q'$ defined by

$$\langle \mathbb{1}_{Q\to Q'}p, q\rangle_{Q',Q} = (p,q)_Q, \quad \forall p\in Q,\ \forall q\in Q,$$

which is a bijective isometry according to Riesz Theorem. As a matter of fact, its inverse $\mathbb{1}_{Q'\to Q}$ is also a bijective isometry, and

$$(\mathbb{1}_{Q'\to Q}g, q)_Q = \langle g,q\rangle_{Q',Q}, \quad \forall g\in Q',\ \forall q\in Q.$$

We then introduce two bilinear forms $a(\cdot,\cdot)$ on $V\times V$ and $b(\cdot,\cdot)$ on $V\times Q$ that are assumed to be continuous, *i.e.* there exist $C_a>0$ and $C_b>0$ such that

$$a(u,v)\leq C_a\|u\|_V\|v\|_V, \quad \forall u\in V, \forall v\in V, \tag{2.13}$$
$$b(v,q)\leq C_b\|v\|_V\|q\|_Q, \quad \forall v\in V, \forall q\in Q. \tag{2.14}$$

We denote by A and B the linear continuous operators associated with a and b, defined by

$$\begin{aligned} A \in \mathcal{L}(V, V'), \quad \langle Au, v\rangle_{V',V} = a(u,v), \quad \forall u \in V, \forall v \in V, \\ B \in \mathcal{L}(V, Q'), \quad \langle Bv, q\rangle_{Q',Q} = b(v,q), \quad \forall v \in V, \forall q \in Q. \end{aligned}$$

The adjoint operator of B is given by

$$B^* \in \mathcal{L}(Q, V'), \quad \langle B^* q, v\rangle_{V',V} = \langle Bv, q\rangle_{Q',Q} = b(v,q), \quad \forall v \in V, \forall q \in Q.$$

Given $f \in V'$ and $g \in Q'$, we consider the saddle-point problem: find $(u,p) \in V \times Q$ such that

$$\begin{aligned} Au + B^* p &= f, \quad \text{in } V', \\ Bu &= g, \quad \text{in } Q'. \end{aligned} \tag{2.15}$$

Or, equivalently, in variational form:

$$\begin{cases} \text{Find } (u,p) \in V \times Q \quad \text{such that} \\ \forall v \in V, \quad a(u,v) + b(v,p) = \langle f, v\rangle_{V',V}, \\ \forall q \in Q, \quad b(u,q) = \langle g, q\rangle_{Q',Q}. \end{cases} \tag{2.16}$$

As for the Stokes problem, we write Problem (2.15) as an *all-in-one* variational formulation

$$\begin{cases} \text{Find } (u,p) \in V \times Q \quad \text{such that} \\ \forall (v,q) \in V \times Q, \quad \mathcal{A}\big((u,p),(v,q)\big) = \langle f, v\rangle_{V',V} + \langle g, q\rangle_{Q',Q}, \end{cases} \tag{2.17}$$

where

$$\mathcal{A}\big((u,p),(v,q)\big) = a(u,v) + b(v,p) + b(u,q).$$

In what follows, we show that Problem (2.17) is well-posed using the notion of T-coercivity, with slightly different techniques depending on the assumptions made on the bilinear form a.

Regarding the form $b(\cdot,\cdot)$ and the operator B, one has the well-known result below, see for instance [Girault and Raviart, 1986, Lemma I.4.1][1] or [Ern and Guermond, 2021a, Lemma C.44], which can be viewed as a reformulation of Banach's Closed Range Theorem.

Theorem 2.6. *The following three statements are equivalent:*

(i) There exists $\underline{\beta} > 0$ *such that*

$$\inf_{q \in Q\setminus\{0\}} \sup_{v \in V\setminus\{0\}} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \underline{\beta}. \tag{2.18}$$

(ii) $B : (\operatorname{Ker} B)^\perp \to Q'$ *is an* isomorphism, *and*

$$\|Bv\|_{Q'} \geq \underline{\beta}\, \|v\|_V, \quad \forall v \in (\operatorname{Ker} B)^\perp.$$

(iii) There exists an isomorphic operator $L_B : Q' \to (\operatorname{Ker} B)^\perp$ *such that*

$$B(L_B g) = g \quad \text{and} \quad \|g\|_{Q'} \geq \underline{\beta}\, \|L_B g\|_V, \quad \forall g \in Q'.$$

[1] Item (iii) below is a rephrasing of the original statement, because it is better suited for our purposes. For details, see the proof of Lemma I.4.1. of [Girault and Raviart, 1986] p. 59, item 2. The operator L_B is a *right-inverse* of the operator B.

Since our aim is to build operators $\mathtt{T}$ from $V \times Q$ to itself, we first introduce the operator

$$\mathtt{B} = \mathbb{1}_{Q' \to Q} \circ B : V \to Q.$$

For all $v \in V$, $\|Bv\|_{Q'} = \|\mathbb{1}_{Q' \to Q}(Bv)\|_Q = \|\mathtt{B}v\|_Q$ and, for all $(v, q) \in V \times Q$,

$$b(v, q) = \langle Bv, q\rangle_{Q',Q} = \langle \mathbb{1}_{Q \to Q'}(\mathtt{B}v), q\rangle_{Q',Q} = (\mathtt{B}v, q)_Q. \tag{2.19}$$

Whenever applicable, we also introduce its *right-inverse*

$$\mathtt{L}_B = L_B \circ \mathbb{1}_{Q \to Q'} : Q \to (\operatorname{Ker} \mathtt{B})^\perp.$$

Observe that

$$b(\mathtt{L}_B p, q) = \langle B\mathtt{L}_B p, q\rangle_{Q',Q} = \langle \mathbb{1}_{Q \to Q'} p, q\rangle_{Q',Q} = (p, q)_Q, \quad \forall p \in Q,\ \forall q \in Q. \tag{2.20}$$

Under these notation, items (ii)-(iii) of Theorem 2.6 now write
(ii) $\mathtt{B} : (\operatorname{Ker} \mathtt{B})^\perp \to Q$ is an *isomorphism*, and

$$\|\mathtt{B}v\|_Q \geq \underline{\beta}\|v\|_V, \quad \forall v \in (\operatorname{Ker} \mathtt{B})^\perp. \tag{2.21}$$

(iii) There exists an isomorphic operator $\mathtt{L}_B : Q \to (\operatorname{Ker} \mathtt{B})^\perp$ such that

$$\mathtt{B}(\mathtt{L}_B q) = q \quad \text{and} \quad \|q\|_Q \geq \underline{\beta}\|\mathtt{L}_B q\|_V, \quad \forall q \in Q. \tag{2.22}$$

For convenience, we often use $\beta = \underline{\beta}^{-1}$, so that

$$\|\mathtt{L}_B q\|_V \leq \beta\|q\|_Q, \quad \forall q \in Q.$$

2.2.2 How to achieve T-coercivity for saddle-point problems?

If a is coercive on the whole space V, we can extend the proof of Theorem 2.3 in the following way.

Theorem 2.7. *Assume that (2.18) holds true and that the form a is symmetric and positive. If there exists a constant $\alpha > 0$ such that*

$$a(u, u) \geq \alpha\|u\|_V^2, \quad \forall u \in V, \tag{2.23}$$

then there exists a unique solution to Problem (2.17) and

$$\|(u, p)\|_{V \times Q} \leq \frac{2\max\Big(\sqrt{2}C_a\beta^2, \beta\big(2 + C_a^2\beta^2\big)^{1/2}\Big)}{\min(\alpha C_a\beta^2, 1)}\|(f, g)\|_{V' \times Q'}. \tag{2.24}$$

Proof. First, we note that the symmetry of the bilinear form a implies that $\mathcal{A}$ is also symmetric. Then, we follow the same ideas as in the proof of Theorem 2.3, replacing $\boldsymbol{v}_p$ by $\mathtt{L}_B p$. We introduce the mapping

$$\begin{aligned} \mathtt{T} : V \times Q &\longrightarrow V \times Q \\ (u, p) &\longmapsto \big(\lambda u + \mathtt{L}_B p, -\lambda p\big) \end{aligned}$$

and we compute

$$\begin{aligned} \mathcal{A}\big((u, p), \mathtt{T}(u, p)\big) &= a(u, \lambda u) + a(u, \mathtt{L}_B p) + b(\lambda u, p) + b(\mathtt{L}_B p, p) - b(u, \lambda p) \\ &= \lambda a(u, u) + a(u, \mathtt{L}_B p) + \|p\|_Q^2, \end{aligned}$$

in view of (2.20).

Because the form a is symmetric and positive, we can apply Young's inequality: for any $\eta > 0$,

$$a(u, \mathtt{L}_B p) \geq -\frac{\eta}{2} a(u,u) - \frac{1}{2\eta} a(\mathtt{L}_B p, \mathtt{L}_B p).$$

Taking into account (2.13) and (2.22), the latter being equivalent to (2.18), we get

$$a(\mathtt{L}_B p, \mathtt{L}_B p) \leq C_a \|\mathtt{L}_B p\|_V^2 \leq C_a \beta^2 \|p\|_Q^2$$

and thus

$$a(u, \mathtt{L}_B p) \geq -\frac{\eta}{2} a(u,u) - \frac{C_a \beta^2}{2\eta} \|p\|_Q^2.$$

Hence, recalling (2.23), if $\lambda - \frac{\eta}{2} > 0$ it follows that

$$\mathcal{A}\big((u,p), \mathtt{T}(u,p)\big) \geq \alpha \left(\lambda - \frac{\eta}{2}\right) \|u\|_V^2 + \left(1 - \frac{C_a \beta^2}{2\eta}\right) \|p\|_Q^2.$$

Setting in particular $\eta = \lambda = C_a \beta^2$, we infer that

$$\mathcal{A}\big((u,p), \mathtt{T}(u,p)\big) \geq \alpha \frac{C_a \beta^2}{2} \|u\|_V^2 + \frac{1}{2} \|p\|_Q^2 \geq \frac{1}{2} \min(\alpha C_a \beta^2, 1) \|(u,p)\|_{V \times Q}^2, \tag{2.25}$$

which proves that $\mathcal{A}$ is $\mathtt{T}$-coercive.

Since $\mathtt{T}(u,p) = (C_a \beta^2 u + \mathtt{L}_B p, -C_a \beta^2 p)$, it holds that

$$\begin{aligned} \|\mathtt{T}(u,p)\|_{V \times Q}^2 &= \|C_a \beta^2 u + \mathtt{L}_B p\|_V^2 + \|C_a \beta^2 p\|_Q^2 \\ &\leq 2 (C_a \beta^2)^2 \|u\|_V^2 + 2 \|\mathtt{L}_B p\|_V^2 + (C_a \beta^2)^2 \|p\|_Q^2 \\ &\leq 2 (C_a \beta^2)^2 \|u\|_V^2 + \big(2\beta^2 + (C_a \beta^2)^2\big) \|p\|_Q^2, \end{aligned}$$

which yields

$$|||\mathtt{T}||| \leq \max\Big(\sqrt{2} C_a \beta^2, \beta \big(2 + C_a^2 \beta^2\big)^{1/2}\Big). \tag{2.26}$$

Lastly, we observe that

$$\langle f, v \rangle_{V',V} + \langle g, q \rangle_{Q',Q} \leq \|(f,g)\|_{V' \times Q'} \|(v,q)\|_{V \times Q}. \tag{2.27}$$

Combining (2.25), (2.26) and (2.27), the stability estimate (2.4) furnishes exactly (2.24). □

Remark 2.8. By applying Theorem 2.7 to Stokes problem, we recover stability estimates (2.8) and (2.12) from the correspondence $\alpha = \nu$, $C_a = \nu$ and $\beta = C_{\mathrm{div}}$.

Remark 2.9. The all-in-one bilinear form $\mathcal{A}$ can also be studied using Banach-Nečas-Babuška's theory inf-sup conditions, see [Ern and Guermond, 2021a, Theorem 49.15 and Lemma 53.12].

In Ladyzhenskaya–Babuška–Brezzi's theory and in many applications, the bilinear form a is not coercive on the whole space V but only on the kernel of the operator $\mathtt{B}$. This is for instance the case in electromagnetism, which will be detailed in Section 2.3. The next result shows how to address this situation in the $\mathtt{T}$-coercivity framework (provided that the form a is symmetric and positive), thus establishing the equivalence between the two theories.

Theorem 2.10. *Assume that the form a is symmetric and positive.*
1. If (2.18) holds true, and if there exists a constant $\alpha_0 > 0$ such that

$$a(u_0, u_0) \geq \alpha_0 \|u_0\|_V^2, \quad \forall u_0 \in \operatorname{Ker} \mathtt{B}, \tag{2.28}$$

then the form $\mathcal{A}$ is T*-coercive. In other words, Problem* (2.17) *is well-posed and*

$$\|(u,p)\|_{V\times Q} \leq C\|(f,g)\|_{V'\times Q'}, \tag{2.29}$$

with C a constant depending only on α_0, β, C_a and C_b.
2. Conversely, if Problem (2.17) *is well-posed, that is, if the form $\mathcal{A}$ is* T*-coercive, then (2.18) and* (2.28) *both hold.*

Proof. 1. We consider the mapping

$$\begin{aligned} \mathtt{T} : V\times Q &\longrightarrow V\times Q \\ (u,p) &\longmapsto \big(\lambda u + \mathtt{L}_B p, -\lambda p + \lambda\mu\,\mathtt{B}u\big). \end{aligned}$$

This is almost the same mapping as the one used in the proof of Theorem 2.7. The only difference is the term $\lambda\mu\,\mathtt{B}u$, which is going to help us handling the extra terms that do not belong to the kernel of B by adjusting the value of the constant μ. We get

$$\begin{aligned} \mathcal{A}\big((u,p),\mathtt{T}(u,p)\big) &= a(u,\lambda u) + a(u,\mathtt{L}_B p) + b(\lambda u,p) + b(\mathtt{L}_B p,p) - b(u,\lambda p) + b(u,\lambda\mu\,\mathtt{B}u) \\ &= \lambda a(u,u) + a(u,\mathtt{L}_B p) + \|p\|_Q^2 + \lambda\mu\|\mathtt{B}u\|_Q^2 \end{aligned}$$

because $b(\mathtt{L}_B p,p) = \|p\|_Q^2$ as previously, and

$$b(u,\mathtt{B}u) = \big\langle Bu,\mathtt{B}u\big\rangle_{Q',Q} = \big(\mathbb{1}_{Q'\to Q}(Bu),\mathtt{B}u\big)_Q = \|\mathtt{B}u\|_Q^2.$$

Since the form a is symmetric and positive, one may use Young's inequality. By proceeding as in the proof of Theorem 2.7 and after setting $\lambda = C_a\beta^2$, we know that

$$\lambda a(u,u) + a(u,\mathtt{L}_B p) + \|p\|_Q^2 \geq \frac{C_a\beta^2}{2}a(u,u) + \frac{1}{2}\|p\|_Q^2,$$

from which we deduce

$$\mathcal{A}\big((u,p),\mathtt{T}(u,p)\big) \geq \frac{C_a\beta^2}{2}\big(a(u,u) + 2\mu\|\mathtt{B}u\|_Q^2\big) + \frac{1}{2}\|p\|_Q^2.$$

To compensate the lack of coercivity of a outside $\operatorname{Ker}\mathtt{B}$, we use the decomposition $u = u_0 + \bar{u}$ with $u_0 \in \operatorname{Ker}\mathtt{B}$ and $\bar{u} \in (\operatorname{Ker}\mathtt{B})^\perp$. Following [Boffi et al., 2013, p. 254], Young's inequality yields

$$\begin{aligned} a(u,u) &= a(u_0,u_0) + 2a(u_0,\bar{u}) + a(\bar{u},\bar{u}) \\ &\geq (1-\theta)a(u_0,u_0) + \left(1-\frac{1}{\theta}\right)a(\bar{u},\bar{u}) \\ &\geq (1-\theta)a(u_0,u_0) + \left(C_a - \frac{C_a}{\theta}\right)\|\bar{u}\|_V^2 \end{aligned}$$

for all $0<\theta<1$. Since $u_0 \in \operatorname{Ker}\mathtt{B}$, we have $\|\mathtt{B}u\|_Q^2 = \|\mathtt{B}\bar{u}\|_Q^2$. Moreover, using (2.21) yields $\|\mathtt{B}\bar{u}\|_Q^2 \geq \beta^{-2}\|\bar{u}\|_V^2$. Thus

$$a(u,u) + 2\mu\|\mathtt{B}u\|_Q^2 \geq (1-\theta)a(u_0,u_0) + \left(C_a - \frac{C_a}{\theta} + \frac{2\mu}{\beta^2}\right)\|\bar{u}\|_V^2. \tag{2.30}$$

Choosing $\theta = \frac{1}{2}$ and $\mu = \frac{3}{4}C_a\beta^2$, it holds that

$$a(u,u) + 2\mu\|\mathtt{B}u\|_Q^2 \geq \frac{1}{2}a(u_0,u_0) + \frac{C_a}{2}\|\bar{u}\|_V^2.$$

Hence, recalling (2.28) and using the inequality $C_a \geq \alpha_0$, we obtain

$$a(u,u) + 2\mu\|\mathtt{B}u\|_Q^2 \geq \frac{\alpha_0}{2}\|u_0\|_V^2 + \frac{\alpha_0}{2}\|\bar{u}\|_V^2 = \frac{\alpha_0}{2}\|u\|_V^2$$

and we conclude that

$$\mathcal{A}\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha_0\frac{C_a\beta^2}{4}\|u\|_V^2 + \frac{1}{2}\|p\|_Q^2. \tag{2.31}$$

From the above, we have

$$\mathtt{T}(u,p) = \big(C_a\beta^2 u + \mathtt{L}_B p, -C_a\beta^2 p + \frac{3}{4}(C_a\beta^2)^2\mathtt{B}u\big).$$

Finally, $\mathtt{T}$ belongs to $\mathcal{L}(V\times Q)$ since $\|\mathtt{L}_B p\|_V \leq \beta\|p\|_Q$ (see (2.22)) and[2]

$$\|\mathtt{B}u\|_Q \leq C_b\|u\|_V. \tag{2.32}$$

The stability estimate (2.29) is then given by (2.4).

2. Conversely, suppose that there exist $\alpha_V > 0$, $\alpha_Q > 0$ and $\mathtt{T} \in \mathcal{L}(V\times Q)$ such that

$$\mathcal{A}\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha_V\|u\|_V^2 + \alpha_Q\|p\|_Q^2, \quad \forall (u,p)\in V\times Q. \tag{2.33}$$

Noting $\mathtt{T} : (u,p) \mapsto \big(\mathtt{T}_V(u,p), \mathtt{T}_Q(u,p)\big)$, we have

$$\mathcal{A}\big((u,p),\mathtt{T}(u,p)\big) = a\big(u,\mathtt{T}_V(u,p)\big) + b\big(\mathtt{T}_V(u,p),p\big) + b\big(u,\mathtt{T}_Q(u,p)\big)$$

and, since $\mathtt{T}$ is bounded,

$$\|\mathtt{T}_V(u,p)\|_V^2 + \|\mathtt{T}_Q(u,p)\|_Q^2 \leq |||\mathtt{T}|||^2\big(\|u\|_V^2 + \|p\|_Q^2\big). \tag{2.34}$$

Now, choosing $u = 0$ in (2.33) and (2.34) yields

$$b\big(\mathtt{T}_V(0,p),p\big) \geq \alpha_Q\|p\|_Q^2 \quad \text{and} \quad \|\mathtt{T}_V(0,p)\|_V \leq |||\mathtt{T}||| \cdot \|p\|_Q, \quad \forall p\in Q.$$

Thus, for $p\in Q\backslash\{0\}$, $\mathtt{T}_V(0,p)\neq 0$, otherwise $b\big(\mathtt{T}_V(0,p),p\big) = 0$, which contradicts $b\big(\mathtt{T}_V(0,p),p\big) > 0$. Then it follows that

$$\sup_{v\in V\backslash\{0\}}\frac{b(v,p)}{\|v\|_V} \geq \frac{b\big(\mathtt{T}_V(0,p),p\big)}{\|\mathtt{T}_V(0,p)\|_V} \geq \frac{\alpha_Q}{|||\mathtt{T}|||}\|p\|_Q, \quad \forall p\in Q\setminus\{0\},$$

which shows that the inf-sup condition (2.18) is fulfilled. Likewise, taking $p = 0$ and $u\in\operatorname{Ker}\mathtt{B}$ in (2.33) and (2.34), we get

$$a\big(u,\mathtt{T}_V(u,0)\big) \geq \alpha_V\|u\|_V^2 \quad \text{and} \quad \|\mathtt{T}_V(u,0)\|_V \leq |||\mathtt{T}|||\,\|u\|_V, \quad \forall u\in\operatorname{Ker}\mathtt{B}.$$

By symmetry and positivity of a, it holds that

$$a\big(u,\mathtt{T}_V(u,0)\big) \leq \big(a(u,u)\big)^{1/2}a\big(\mathtt{T}_V(u,0),\mathtt{T}_V(u,0)\big)^{1/2}.$$

Thus

$$\alpha_V\|u\|_V^2 \leq a\big(u,\mathtt{T}_V(u,0)\big) \leq \big(a(u,u)\big)^{1/2}\big(C_a|||\mathtt{T}|||^2\|u\|_V^2\big)^{1/2}$$

and hence $a(u,u) \geq \frac{\alpha_V^2}{C_a|||\mathtt{T}|||^2}\|u\|_V^2$ for all $u\in\operatorname{Ker}\mathtt{B}$, which proves (2.28). □

Remark 2.11. The $\mathtt{T}$-coercivity estimate (2.31) is very close to the case where a is coercive on the whole space V. As a matter of fact, the only difference compared to (2.25) is that the constant before the term $\|u\|_V^2$ is twice as small, with $\alpha_0 = \alpha$.

[2] Classically,

$$\begin{aligned}\|\mathtt{B}u\|_Q^2 &= (\mathtt{B}u,\mathtt{B}u)_Q = \big\langle \mathbb{1}_{Q\to Q'}(\mathtt{B}u),\mathtt{B}u\big\rangle_{Q',Q} \quad \text{by definition of } \mathbb{1}_{Q\to Q'},\\ &= \big\langle \mathbb{1}_{Q\to Q'}\circ\mathbb{1}_{Q'\to Q}(Bu),\mathtt{B}u\big\rangle_{Q',Q} = \big\langle Bu,\mathtt{B}u\big\rangle_{Q',Q} \quad \text{since } \mathbb{1}_{Q\to Q'}\circ\mathbb{1}_{Q'\to Q} = \mathrm{Id}_{Q'},\\ &= b(u,\mathtt{B}u) \leq C_b\|u\|_V\|\mathtt{B}u\|_Q \quad \text{by definition and continuity of } b \text{ (2.14)}.\end{aligned}$$

2.2.3 Augmented saddle-point problems

Let $c(\cdot,\cdot)$ be a positive and continuous bilinear form defined on $Q \times Q$, namely

$$c(p,p) \geq 0, \quad \forall p \in Q \quad \text{and} \quad \exists C_c > 0, \quad c(p,q) \leq C_c \|p\|_Q \|q\|_Q, \quad \forall p \in Q, \forall q \in Q. \tag{2.35}$$

In some cases, the assumption on the positivity of the form c can be relaxed, see Remark 2.13. We denote by C the linear operator associated with the form c, defined by

$$C \in \mathcal{L}(Q, Q'), \quad \langle Cp, q\rangle_{Q',Q} = c(p,q), \quad \forall p \in Q, \forall q \in Q.$$

The *all-in-one* approach developed previously also enables us to deal with augmented saddle-point problems: given $f \in V'$ and $g \in Q'$, find $(u,p) \in V \times Q$ such that

$$\begin{aligned} Au + B^* p &= f, \quad \text{in } V', \\ Bu - Cp &= g, \quad \text{in } Q', \end{aligned} \tag{2.36}$$

where the operator C possibly acts as a *small perturbation* of the original saddle-point problem (2.15). The weak formulation of (2.36) reads:

$$\begin{cases} \text{Find } (u,p) \in V \times Q \quad \text{such that} \\ \forall (v,q) \in V \times Q, \quad \mathcal{A}_c\big((u,p),(v,q)\big) = \langle f, v\rangle_{V',V} + \langle g, q\rangle_{Q',Q}, \end{cases} \tag{2.37}$$

with

$$\mathcal{A}_c\big((u,p),(v,q)\big) = a(u,v) + b(v,p) + b(u,q) - c(p,q).$$

As before, the bilinear form a is supposed to be symmetric and positive.

2.2.4 How to achieve T-coercivity for augmented saddle-point problems?

Once again, we distinguish the case where the form a is coercive on V or only on $\operatorname{Ker} \mathtt{B}$. If the form a is coercive on V, the results from the un-augmented case allow straightforwardly to handle the augmented one.

Theorem 2.12. *Assume that (2.18) holds true, that the form c fulfills (2.35) and that the form a is symmetric and positive. If there exists a constant $\alpha > 0$ such that*

$$a(u,u) \geq \alpha \|u\|_V^2, \quad \forall u \in V,$$

then there exists a unique solution to Problem (2.37).

Proof. With the same operator $\mathtt{T}$ as for the un-augmented problem, namely

$$\begin{aligned} \mathtt{T} : V \times Q &\longrightarrow V \times Q \\ (u,p) &\longmapsto \big(C_a \beta^2 u + \mathtt{L}_B p, -C_a \beta^2 p\big), \end{aligned}$$

it holds that

$$\mathcal{A}_c\big((u,p), \mathtt{T}(u,p)\big) = C_a \beta^2 a(u,u) + a(u, \mathtt{L}_B p) + \|p\|_Q^2 + C_a \beta^2 c(p,p).$$

Therefore, a similar argument as in Theorem 2.7 furnishes

$$\mathcal{A}_c\big((u,p), \mathtt{T}(u,p)\big) \geq \alpha \frac{C_a \beta^2}{2} \|u\|_V^2 + \frac{1}{2} \|p\|_Q^2 + C_a \beta^2 c(p,p),$$

which shows that $\mathcal{A}_c$ is $\mathtt{T}$-coercive since c is positive. □

Remark 2.13. A particular case that appears in many applications – see Section 2.4 for the example of nearly-incompressible elasticity – is when c has the form

$$c(p,q) = \varepsilon(p,q)_Q, \quad \varepsilon \geq 0.$$

In this case, we obtain the estimate

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha\frac{C_a\beta^2}{2}\|u\|_V^2 + \left(\frac{1}{2} + \varepsilon C_a\beta^2\right)\|p\|_Q^2, \tag{2.38}$$

so that the augmentation c improves the constant before the term $\|p\|_Q^2$ and thus stabilizes the bilinear form $\mathcal{A}_c$. Moreover, the above estimate is robust for small values of ε. Besides, it even allows to take negative values of ε. Indeed, if $\varepsilon < 0$, we have

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha\frac{C_a\beta^2}{2}\|u\|_V^2 + \left(\frac{1}{2} - |\varepsilon|\, C_a\beta^2\right)\|p\|_Q^2.$$

Hence, the bilinear form $\mathcal{A}_c$ remains $\mathtt{T}$-coercive whenever $|\varepsilon| < \frac{1}{2C_a\beta^2}$.

Let us now suppose that a is not coercive on the whole space V but only on the kernel of $\mathtt{B}$. Then, two different situations occur. Either the form c can be viewed as a *small perturbation*, and we shall look for a solution of (2.36) that is *close* to the solution of the original problem (2.15). Or this is not the case, and the form c is viewed as a "fixed" augmentation, and there is no obvious connection *a priori* between the solutions of the augmented and un-augmented problems.

2.2.5 Additional results for small perturbations

We say that c is a small perturbation if it can be written as

$$c(p,q) = \varepsilon c_0(p,q), \quad \varepsilon > 0, \tag{2.39}$$

with ε a small parameter and c_0 a symmetric, positive and continuous form on Q. We start with the simple case

$$c(p,q) = \varepsilon(p,q)_Q, \quad \varepsilon > 0, \tag{2.40}$$

for which the $\mathtt{T}$-coercivity approach yields a shorter proof than the corresponding result stated in Ladyzhenskaya–Babuška–Brezzi's framework, see [Boffi et al., 2013, pages 247-252].

Theorem 2.14. *Assume that (2.18) holds true, that the form a is symmetric and positive, and that c takes the simple form of (2.40). If there exists a constant $\alpha_0 > 0$ such that*

$$a(u_0,u_0) \geq \alpha_0\|u_0\|_V^2, \quad \forall u_0 \in \operatorname{Ker}\mathtt{B}, \tag{2.41}$$

and if ε is small enough*, namely*

$$\varepsilon \leq \frac{1}{2C_a\beta^4C_b^2}\left(2 - \frac{\alpha_0}{C_a}\right), \tag{2.42}$$

then Problem (2.37) is well-posed and

$$\|(u,p)\|_{V\times Q} \leq C\|(f,g)\|_{V'\times Q'}, \tag{2.43}$$

with C a constant depending only on α_0, β, C_a and C_b.

Proof. Here again, we consider the mapping

$$\begin{aligned}\mathtt{T} : V \times Q &\longrightarrow V \times Q\\ (u,p) &\longmapsto \big(\lambda u + \mathtt{L}_B p, -\lambda p + \lambda\mu\,\mathtt{B}u\big).\end{aligned}$$

The beginning of the proof is the same as in Theorem 2.10. Taking into account the extra terms coming from the perturbation, we get

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) = \lambda a(u,u) + a(u,\mathtt{L}_B p) + \|p\|_Q^2 + \lambda\mu\|Bu\|_{Q'}^2 + \lambda c(p,p) - \lambda\mu c(p,\mathtt{B}u).$$

Using Young's inequality and setting $\lambda = C_a\beta^2$, it follows that

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \frac{C_a\beta^2}{2}\big(a(u,u) + 2\mu\|Bu\|_{Q'}^2\big) + \frac{1}{2}\|p\|_Q^2 + C_a\beta^2 c(p,p) - C_a\beta^2\mu c(p,\mathtt{B}u). \quad (2.44)$$

Now, as in (2.30), it holds that

$$a(u,u) + 2\mu\|Bu\|_{Q'}^2 \geq (1-\theta)a(u_0,u_0) + \left(C_a - \frac{C_a}{\theta} + \frac{2\mu}{\beta^2}\right)\|\bar{u}\|_V^2 \quad (2.45)$$

for all $0 < \theta < 1$, where $u = u_0 + \bar{u}$ with $u_0 \in \operatorname{Ker}\mathtt{B}$ and $\bar{u} \in (\operatorname{Ker}\mathtt{B})^\perp$.

Knowing that $c(p,q) = \varepsilon(p,q)_Q$ for all p and q in Q, Young's inequality implies that, for all $\delta > 0$,

$$\begin{aligned}-c(p,\mathtt{B}u) = -c(p,\mathtt{B}\bar{u}) = -\varepsilon(p,\mathtt{B}\bar{u})_Q &\geq -\varepsilon\frac{\delta}{2}\|p\|_Q^2 - \frac{\varepsilon}{2\delta}\|\mathtt{B}\bar{u}\|_Q^2\\ &\geq -\varepsilon\frac{\delta}{2}\|p\|_Q^2 - \varepsilon\frac{C_b^2}{2\delta}\|\bar{u}\|_V^2 \quad \text{in view of (2.32).}\end{aligned}$$

Putting (2.44), (2.45) and the above inequality together, we find that

$$\begin{aligned}\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \frac{C_a\beta^2}{2}&\left((1-\theta)a(u_0,u_0) + \left(C_a - \frac{C_a}{\theta} + \frac{2\mu}{\beta^2} - \mu\varepsilon\frac{C_b^2}{\delta}\right)\|\bar{u}\|_V^2\right)\\ &+ \frac{1}{2}\|p\|_Q^2 + \varepsilon C_a\beta^2\left(1 - \mu\frac{\delta}{2}\right)\|p\|_Q^2.\end{aligned}$$

Hence, choosing $\theta = \frac{1}{2}$, $\mu = C_a\beta^2$ and recalling (2.41), it holds that

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \frac{C_a\beta^2}{2}\left(\frac{\alpha_0}{2}\|u_0\|_V^2 + C_a\left(1 - \varepsilon\frac{\beta^2 C_b^2}{\delta}\right)\|\bar{u}\|_V^2\right) + \frac{1}{2}\|p\|_Q^2 + \varepsilon C_a\beta^2\left(1 - C_a\beta^2\frac{\delta}{2}\right)\|p\|_Q^2. \quad (2.46)$$

Finally, we set $\delta = \frac{1}{C_a\beta^2}$ so that

$$1 - C_a\beta^2\frac{\delta}{2} = \frac{1}{2} \qquad \text{and} \qquad 1 - \varepsilon\frac{\beta^2 C_b^2}{\delta} = 1 - \varepsilon C_a\beta^4 C_b^2 \geq \frac{1}{2}\cdot\frac{\alpha_0}{C_a}$$

in virtue of (2.42). Thus

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha_0\frac{C_a\beta^2}{4}\|u\|_V^2 + \left(\frac{1}{2} + \varepsilon\frac{C_a\beta^2}{2}\right)\|p\|_Q^2, \quad (2.47)$$

where we used that $\|u\|_V^2 = \|u_0\|_V^2 + \|\bar{u}\|_V^2$. All in all, we have chosen

$$\mathtt{T}(u,p) = \big(C_a\beta^2 u + \mathtt{L}_B p, -C_a\beta^2 p + (C_a\beta^2)^2\mathtt{B}u\big).$$

Then, estimate (2.43) follows from (2.4) with a stability constant independent of ε since (2.47) is robust for vanishing ε and since $|||\mathtt{T}|||$ does not depend on ε either. □

Remark 2.15. The final estimate (2.47) is very close to (2.38). The only difference between these two estimates is a factor of 2 between the constants multiplying the norms of u and p, with $\alpha_0 = \alpha$.

Remark 2.16. In Ladyzhenskaya–Babuška–Brezzi's framework, it is commonly assumed that $\varepsilon \leq 1$. On the other hand, in (2.42), we find a smallness condition that depends *explicitly* on the various constants of the problem.

Remark 2.17. The inf-sup condition (2.18) and the continuity of b imply that $\underline{\beta} \leq C_b$, *i.e.* $C_b\beta \geq 1$. Therefore, (2.42) yields in particular

$$\varepsilon \leq \frac{1}{C_a\beta^2},$$

which corresponds to the condition found in Remark 2.13 for negative values of ε. As a matter of fact, the non-coercivity of a on the whole space V calls for the introduction of a term $\mathtt{B}u$ in the mapping $\mathtt{T}$. This term induces an additional term of the form $c(p, \mathtt{B}u)$ in the expression of $\mathcal{A}_c\big((u,p), \mathtt{T}(u,p)\big)$, that can be interpreted as a "negative perturbation" of the bilinear form $\mathcal{A}$.

Now, we move to the case where c is given by (2.39). Let us denote by C_{c_0} the continuity constant of the bilinear form c_0. The next theorem establishes the well-posedness of the perturbed problem for a very general form c_0.

Theorem 2.18. *Assume that (2.18) holds true, and that the bilinear forms a and c_0 are both symmetric and positive. Suppose in addition that there exist $\alpha_0 > 0$ and $\gamma_0 > 0$ such that*

$$a(u_0, u_0) \geq \alpha_0 \|u_0\|_V^2, \quad \forall u_0 \in \operatorname{Ker} \mathtt{B},$$

and

$$c_0(p_0, p_0) \geq \gamma_0 \|p_0\|_Q^2, \quad \forall p_0 \in \operatorname{Ker} \mathtt{B}^*. \tag{2.48}$$

If ε is small enough*, namely*

$$\varepsilon \leq \frac{1}{2C_{c_0}C_a\beta^4C_b^2}, \tag{2.49}$$

then Problem (2.37) is well-posed and

$$\|(u,p)\|_{V\times Q} \leq C\|(f,g)\|_{V'\times Q'},$$

with C a constant depending only on α_0, β, γ_0, C_a and C_b.

Proof. First, we adapt the beginning of the proof of Theorem 2.14 to take into consideration the bilinear form c_0. Since c_0 is symmetric and positive, we can use Young's inequality to obtain

$$\begin{aligned} -c(p, \mathtt{B}u) = -\varepsilon c_0(p, \mathtt{B}\bar{u})_Q &\geq -\varepsilon\frac{\delta}{2}c_0(p,p) - \frac{\varepsilon}{2\delta}c_0(\mathtt{B}\bar{u}, \mathtt{B}\bar{u}) \\ &\geq -\varepsilon\frac{\delta}{2}c_0(p,p) - \varepsilon\frac{C_{c_0}C_b^2}{2\delta}\|\bar{u}\|_V^2 \quad \text{since } \|\mathtt{B}\bar{u}\|_Q^2 \leq C_b^2\|\bar{u}\|_V^2, \end{aligned}$$

and thus (2.46) becomes

$$\begin{aligned} \mathcal{A}_c\big((u,p), \mathtt{T}(u,p)\big) \geq \frac{C_a\beta^2}{2}\left(\frac{\alpha_0}{2}\|u_0\|_V^2 + C_a\left(1 - \varepsilon\frac{C_{c_0}\beta^2C_b^2}{\delta}\right)\|\bar{u}\|_V^2\right) \\ + \frac{1}{2}\|p\|_Q^2 + \varepsilon C_a\beta^2\left(1 - C_a\beta^2\frac{\delta}{2}\right)c_0(p,p), \end{aligned}$$

where $\mathtt{T}$ is the mapping

$$\begin{aligned} \mathtt{T} : V \times Q &\longrightarrow V \times Q \\ (u,p) &\longmapsto \big(C_a\beta^2u + \mathtt{L}_Bp, -C_a\beta^2p + (C_a\beta^2)^2\mathtt{B}u\big). \end{aligned}$$

Setting $\delta = \frac{1}{C_a\beta^2}$ as before, we get the estimate

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha_0\frac{C_a\beta^2}{4}\|u\|_V^2 + \frac{1}{2}\|p\|_Q^2 + \varepsilon\frac{C_a\beta^2}{2}c_0(p,p),$$

as long as $\varepsilon \leq \frac{1}{2C_{c_0}C_a\beta^4C_b^2}(2-\frac{\alpha_0}{C_a})$, which is the case under the assumption (2.49) since $\alpha_0 \leq C_a$.

Then, as the bilinear form c_0 is not necessarily coercive on the whole space Q, we use the decomposition $p = p_0 + \bar{p}$ with $p_0 \in \operatorname{Ker}\mathtt{B}^*$ and $\bar{p} \in (\operatorname{Ker}\mathtt{B}^*)^\perp$. From Young's inequality, we have

$$c_0(p,p) \geq (1-\theta)c_0(p_0,p_0) + \left(C_{c_0} - \frac{C_{c_0}}{\theta}\right)\|\bar{p}\|_Q^2$$

for all $0 < \theta < 1$. Setting $\theta = \frac{1}{2}$ and using (2.48), it follows that

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha_0\frac{C_a\beta^2}{4}\|u\|_V^2 + \frac{1}{2}\|p\|_Q^2 + \varepsilon\frac{C_a\beta^2}{2}\left(\frac{\gamma_0}{2}\|p_0\|_Q^2 - C_{c_0}\|\bar{p}\|_Q^2\right).$$

Now, we notice that

$$\frac{1}{2}\|p\|_Q^2 \geq \frac{1}{8}\|p\|_Q^2 + \frac{3}{8}\|\bar{p}\|_Q^2$$

and that, thanks to (2.49),

$$\frac{3}{8}\|\bar{p}\|_Q^2 = \frac{1}{2}\cdot\frac{3}{4}\|\bar{p}\|_Q^2 \geq \varepsilon C_{c_0}C_a\beta^4C_b^2\cdot\frac{3}{4}\|\bar{p}\|_Q^2 \geq \varepsilon\frac{C_a\beta^2}{2}\cdot\frac{3}{2}C_{c_0}\|\bar{p}\|_Q^2$$

because $C_b\beta \geq 1$. Hence

$$\begin{aligned}\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) &\geq \alpha_0\frac{C_a\beta^2}{4}\|u\|_V^2 + \frac{1}{8}\|p\|_Q^2 + \varepsilon\frac{C_a\beta^2}{2}\left(\frac{\gamma_0}{2}\|p_0\|_Q^2 + \left(\frac{3}{2}C_{c_0} - C_{c_0}\right)\|\bar{p}\|_Q^2\right)\\ &\geq \alpha_0\frac{C_a\beta^2}{4}\|u\|_V^2 + \left(\frac{1}{8} + \varepsilon\gamma_0\frac{C_a\beta^2}{4}\right)\|p\|_Q^2 \quad \text{since } C_{c_0} \geq \gamma_0,\end{aligned}$$

which shows that $\mathcal{A}_c$ is T-coercive. □

Lastly, we mention that an important consequence of the previous result is to estimate the distance between the solution $(u_\varepsilon, p_\varepsilon)$ of the perturbed problem

$$\begin{aligned}Au_\varepsilon + B^*p_\varepsilon &= f, \quad \text{in } V',\\ Bu_\varepsilon - \varepsilon C_0 p_\varepsilon &= g, \quad \text{in } Q',\end{aligned} \tag{2.50}$$

and the solution (u,p) of the original saddle-point problem (2.15) as a function of the penalty parameter ε.

Corollary 2.19. *Assume that (2.18) holds true, that the form a is symmetric and positive, and that c takes the form of (2.39). If there exist $\alpha_0 > 0$ and $\gamma_0 > 0$ such that*

$$a(u_0,u_0) \geq \alpha_0\|u_0\|_V^2, \quad \forall u_0 \in \operatorname{Ker}\mathtt{B}, \qquad c_0(p_0,p_0) \geq \gamma_0\|p_0\|_Q^2, \quad \forall p_0 \in \operatorname{Ker}\mathtt{B}^*,$$

and if

$$\varepsilon \leq \frac{1}{2C_{c_0}C_a\beta^4C_b^2},$$

then we have

$$\|u-u_\varepsilon\|_V + \|p-p_\varepsilon\|_Q \leq C\varepsilon, \tag{2.51}$$

with C a constant depending only on α_0, β, γ_0, C_a, C_b and C_{c_0}.

Proof. Subtracting (2.50) from (2.15), we find that $(u-u_\varepsilon, p-p_\varepsilon)$ solves the system

$$\begin{aligned} A(u-u_\varepsilon)+B^*(p-p_\varepsilon) &= 0, && \text{in } V',\\ B(u-u_\varepsilon)-\varepsilon C_0(p-p_\varepsilon) &= -\varepsilon C_0 p, && \text{in } Q'. \end{aligned}$$

From Theorem 2.18, we infer that

$$\|(u-u_\varepsilon, p-p_\varepsilon)\|_{V\times Q} \le C\|(0,-\varepsilon C_0 p)\|_{V'\times Q'}$$

with C depending only on α_0, β, γ_0, C_a and C_b. Thus

$$\|(u-u_\varepsilon, p-p_\varepsilon)\|_{V\times Q} \le CC_{c_0}\varepsilon\|p\|_Q,$$

which proves (2.51). □

2.2.6 Case of a "fixed" augmentation

If the bilinear form c is not given by (2.39), the extra terms of the form $c(p, \mathtt{B}u)$ arising from the previously considered $\mathtt{T}$-coercivity operator cannot be controlled as before, because there is no factor ε to adjust. Below, we assume that c is coercive on Q, namely that there exists $\gamma > 0$ such that

$$c(p,p) \ge \gamma\|p\|_Q^2, \quad \forall p \in Q. \tag{2.52}$$

So, to control these extra terms, we introduce an operator C^{-1} in the expression of $\mathtt{T}$, where $C^{-1} \in \mathcal{L}(Q', Q)$ is defined by

$$c(C^{-1}g, q) = \langle g, q\rangle_{Q',Q}, \quad \forall g \in Q', \forall q \in Q.$$

One can easily check that the operator C^{-1} satisfies

$$(C_c)^{-1}\|g\|_{Q'} \le \|C^{-1}g\|_Q \le \gamma^{-1}\|g\|_{Q'}, \quad \forall g \in Q',$$

and

$$\langle g, C^{-1}g\rangle_{Q',Q} \ge \frac{\gamma}{C_c^2}\|g\|_{Q'}^2, \quad \forall g \in Q'. \tag{2.53}$$

Theorem 2.20. *Assume that (2.52) holds true and that the bilinear forms a and c are both symmetric and positive. Suppose in addition that there exists a constant $\alpha_B > 0$ such that*

$$a(u,u) + \frac{\gamma}{2C_c^2}\|Bu\|_{Q'}^2 \ge \alpha_B\|u\|_V^2, \quad \forall u \in V, \tag{2.54}$$

then Problem (2.37) is well-posed and

$$\|(u,p)\|_{V\times Q} \le C\|(f,g)\|_{V'\times Q'},$$

with C a constant depending only on α_B, γ and C_b.

Proof. For $\eta, \mu > 0$, we consider the mapping

$$\begin{aligned} \mathtt{T} : V\times Q &\longrightarrow V\times Q\\ (u,p) &\longmapsto \big(u, -\eta p + \mu C^{-1}(Bu)\big). \end{aligned}$$

Then, using the definitions of C^{-1} and B, we compute

$$\begin{aligned} \mathcal{A}_c\big((u,p), \mathtt{T}(u,p)\big) &= a(u,u) + b(u,p) - \eta b(u,p) + \mu b(u, C^{-1}Bu) + \eta c(p,p) - \mu c(p, C^{-1}Bu)\\ &= a(u,u) + (1-\eta)b(u,p) + \mu\langle Bu, C^{-1}Bu\rangle_{Q',Q} + \eta c(p,p) - \mu\langle Bu, p\rangle_{Q',Q}\\ &= a(u,u) + (1-\eta-\mu)b(u,p) + \mu\langle Bu, C^{-1}Bu\rangle_{Q',Q} + \eta c(p,p). \end{aligned}$$

Let us choose $\eta, \mu > 0$ such that $\eta + \mu = 1$ to cancel the second term above. To fix ideas, let $\eta = \mu = 1/2$, so that

$$\mathtt{T}(u,p) = \Big(u, -\frac{1}{2}p + \frac{1}{2}C^{-1}(Bu)\Big) \tag{2.55}$$

and

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) = a(u,u) + \frac{1}{2}\langle Bu, C^{-1}Bu\rangle_{Q',Q} + \frac{1}{2}c(p,p).$$

Owing to (2.53) and (2.52), we deduce that

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq a(u,u) + \frac{\gamma}{2C_c^2}\|Bu\|_{Q'}^2 + \frac{\gamma}{2}\|p\|_Q^2,$$

and the result follows. □

Remark 2.21. The T-coercivity estimate reads

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \alpha_B\|u\|_V^2 + \frac{\gamma}{2}\|p\|_Q^2, \tag{2.56}$$

so that it depends on γ, whereas it was independent of ε in the small perturbation case. Moreover, because of the term $C^{-1}(Bu)$ in (2.55), $|||\mathtt{T}|||$ behaves as γ^{-1}. Nevertheless, the final stability estimate is robust because the value of the constant γ is fixed.

Remark 2.22. Note that Theorem 2.20 does not require the inf-sup condition (2.18) to be true. However, if (2.18) holds, then (2.54) is automatically satisfied. As a matter of fact, for any $u \in V$, using the decomposition $u = u_0 + \bar{u}$ with $u_0 \in \operatorname{Ker}\mathtt{B}$ and $\bar{u} \in (\operatorname{Ker}\mathtt{B})^\perp$, we have seen in the proof of Theorem 2.10 that, for all $0 < \theta < 1$, it holds

$$a(u,u) \geq (1-\theta)a(u_0,u_0) + \left(C_a - \frac{C_a}{\theta}\right)\|\bar{u}\|_V^2 \quad \text{and} \quad \|Bu\|_{Q'}^2 = \|\mathtt{B}\bar{u}\|_Q^2 \geq \beta^{-2}\|\bar{u}\|_V^2.$$

Hence,

$$a(u,u) + \frac{\gamma}{2C_c^2}\|Bu\|_{Q'}^2 \geq (1-\theta)a(u_0,u_0) + \left(C_a - \frac{C_a}{\theta} + \frac{\gamma}{2C_c^2}\beta^{-2}\right)\|\bar{u}\|_V^2.$$

We then observe that

$$\left(C_a - \frac{C_a}{\theta} + \frac{\gamma}{2C_c^2}\beta^{-2}\right) > 0, \quad \forall\theta \in \left(\left(1 + \frac{\gamma}{2C_c^2C_a}\beta^{-2}\right)^{-1}, 1\right),$$

so (2.54) is obtained by choosing some $\theta = \theta(C_a, \beta, C_c, \gamma)$ in the above interval.

Remark 2.23. We will see in Section 2.5 that Theorem 2.20 is sufficient to handle the case of neutron diffusion. Nevertheless, note that assumption (2.54) is not optimal since it depends on the arbitrary choice $\eta = \mu = 1/2$ made in the proof. Looking through the proof, we see that the result of Theorem 2.20 still holds true as long as there exist $\tilde{\alpha}_B > 0$ and $0 < \tilde{\mu} < 1$ such that

$$a(u,u) + \tilde{\mu}\langle Bu, C^{-1}Bu\rangle_{Q',Q} \geq \tilde{\alpha}_B\|u\|_V^2, \quad \forall u \in V.$$

The final T-coercivity estimate then reads

$$\mathcal{A}_c\big((u,p),\mathtt{T}(u,p)\big) \geq \tilde{\alpha}_B\|u\|_V^2 + (1-\tilde{\mu})\gamma\|p\|_Q^2.$$

However, this estimate is possibly less sharp than (2.56) if $\tilde{\mu} > \frac{1}{2}$.

In addition to the Stokes problem, let us see next how other typical examples of mixed formulations fall within the T-coercivity framework.

2.3 Application to electromagnetism

Our goal is to solve the so-called quasi-static magnetic problem set in a homogeneous or an anisotropic medium, surrounded by a perfect conductor (see [Assous et al., 2018, Section 6.4]). The medium is characterized by its dielectric permittivity $\underline{\varepsilon}$ and its magnetic permeability $\underline{\mu}$.

Let Ω be the domain of $\mathbb{R}^3$ in which the problem is set. For simplicity, we assume that Ω is simply connected, with a connected boundary. Moreover, we assume that $\xi \in \{\underline{\varepsilon}, \underline{\mu}\}$ satisfy the following assumption:

$$\begin{cases} \xi \text{ is a real-valued, symmetric, measurable tensor field on } \Omega, \\ \exists \xi_-, \xi_+ > 0,\ \forall \boldsymbol{z} \in \mathbb{R}^3,\ \xi_- |\boldsymbol{z}|^2 \leq \xi \boldsymbol{z} \cdot \boldsymbol{z} \leq \xi_+ |\boldsymbol{z}|^2 \text{ a.e. in } \Omega. \end{cases} \tag{2.57}$$

Because one is dealing with symmetric tensors, if ξ fulfills (2.57), so does ξ^{-1}, with $(\xi^{-1})_+ = (\xi_-)^{-1}$ and $(\xi^{-1})_- = (\xi_+)^{-1}$.

Given $\boldsymbol{H}^\star \in \boldsymbol{L}^2(\Omega)$, such that $\underline{\mu}\boldsymbol{H}^\star \in \boldsymbol{H}_0(\mathrm{div}\,;\Omega) \cap \boldsymbol{H}(\mathrm{div}\,0;\Omega)$ and $\rho \in H^{-1}(\Omega)$, the quasi-static magnetic problem amounts to finding $\boldsymbol{E} \in \boldsymbol{L}^2(\Omega)$ such that

$$\begin{aligned} \underline{\mu}^{-1}\mathbf{curl}\,\boldsymbol{E} &= \boldsymbol{H}^\star, && \text{in } \Omega, \\ \mathrm{div}\,(\underline{\varepsilon}\boldsymbol{E}) &= \rho, && \text{in } \Omega, \\ \boldsymbol{E} \times \boldsymbol{n} &= 0, && \text{on } \partial\Omega. \end{aligned} \tag{2.58}$$

Under the assumptions on $\underline{\varepsilon}$ and $\underline{\mu}$, on the one hand we note that $\boldsymbol{E} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega)$. On the other hand, it is known that the problem (2.58) is well-posed, see for instance [Assous et al., 2018, Theorem 6.1.4]. Below, we propose to recover well-posedness using the T-coercivity approach.

2.3.1 Proving well-posedness with T-coercivity

2.3.1.1 In a homogeneous medium

Let us first assume that $\underline{\varepsilon} = \underline{\mu} = \mathbb{I}_3$ in Ω. To build an all-in-one equivalent variational formulation, we follow *e.g.* [Ciarlet Jr, 2021]. In this case, the electromagnetic energy can be expressed in terms of the electric field as $(\boldsymbol{E}, \boldsymbol{E})_{\boldsymbol{L}^2(\Omega)} + (\mathbf{curl}\,\boldsymbol{E}, \mathbf{curl}\,\boldsymbol{E})_{\boldsymbol{L}^2(\Omega)}$. In other words, it is equal to $\|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathbf{curl}\,;\Omega)}$, where $\|\cdot\|_{\boldsymbol{H}(\mathbf{curl}\,;\Omega)}$ denotes the "natural" norm in $\boldsymbol{H}(\mathbf{curl}\,;\Omega)$. We endow $H^1_0(\Omega)$ with $\|\nabla\cdot\|$ and the corresponding inner product $(\nabla\cdot, \nabla\cdot)_{\boldsymbol{L}^2(\Omega)}$. Bearing in mind that $\mathbf{curl}\,(\nabla p) = 0$, it follows that

$$\|\nabla q\| = \|\nabla q\|_{\boldsymbol{H}(\mathbf{curl}\,;\Omega)}, \quad \forall q \in H^1_0(\Omega).$$

First, for $\boldsymbol{H}^\star \in \boldsymbol{H}_0(\mathrm{div}\,;\Omega) \cap \boldsymbol{H}(\mathrm{div}\,0;\Omega)$ and $\rho \in H^{-1}(\Omega)$, one can prove that the equivalent weak formulation of Problem (2.58) reads: find $\boldsymbol{E} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega)$ such that

$$\begin{aligned} (\mathbf{curl}\,\boldsymbol{E}, \mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} &= (\boldsymbol{H}^\star, \mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}, && \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \\ (\boldsymbol{E}, \nabla q)_{\boldsymbol{L}^2(\Omega)} &= -\langle \rho, q \rangle_{H^{-1}(\Omega), H^1_0(\Omega)}, && \forall q \in H^1_0(\Omega). \end{aligned}$$

Second, in order to fit (2.58) into the abstract framework (2.15), we introduce an artifical pressure unknown $\tilde{p}$ by adding a term $(\boldsymbol{v}, \nabla\tilde{p})_{\boldsymbol{L}^2(\Omega)}$ in the first equation. The previous formulation becomes: find $(\boldsymbol{E}, \tilde{p}) \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \times H^1_0(\Omega)$ such that

$$\begin{aligned} (\mathbf{curl}\,\boldsymbol{E}, \mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} + (\boldsymbol{v}, \nabla\tilde{p})_{\boldsymbol{L}^2(\Omega)} &= (\boldsymbol{H}^\star, \mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}, && \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \\ (\boldsymbol{E}, \nabla q)_{\boldsymbol{L}^2(\Omega)} &= -\langle \rho, q \rangle_{H^{-1}(\Omega), H^1_0(\Omega)}, && \forall q \in H^1_0(\Omega). \end{aligned} \tag{2.59}$$

Indeed, one can easily check that $(\boldsymbol{E}, \tilde{p})$ is solution of (2.59) if and only if $\tilde{p} = 0$ and $\boldsymbol{E}$ is solution of (2.58). So, defining the bilinear forms

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) &= (\mathbf{curl}\,\boldsymbol{u}, \mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}, && \forall \boldsymbol{u} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \\ b(\boldsymbol{v}, q) &= (\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(\Omega)}, && \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \forall q \in H^1_0(\Omega), \end{aligned}$$

the *all-in-one* bilinear and linear forms of Maxwell problem are respectively given by

$$\mathcal{A}\big((\boldsymbol{E},\tilde{p}),(\boldsymbol{v},q)\big) = a(\boldsymbol{E},\boldsymbol{v}) + b(\boldsymbol{v},\tilde{p}) + b(\boldsymbol{E},q), \tag{2.60}$$

$$\ell((\boldsymbol{v},q)) = (\boldsymbol{H}^\star, \mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} - \langle \rho, q\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}. \tag{2.61}$$

The continuity constants are such that $C_a = 1$, $C_b = 1$, and $C_\ell \leq \big(\|\boldsymbol{H}^\star\|^2 + \|\rho\|^2_{H^{-1}(\Omega)}\big)^{1/2}$.

Let us give an explicit expression of the abstract operators

$$\mathtt{B} \in \mathcal{L}\big(\boldsymbol{H}_0(\mathbf{curl}\,;\Omega), H_0^1(\Omega)\big), \qquad \mathtt{L}_B \in \mathcal{L}\big(H_0^1(\Omega), (\operatorname{Ker}\mathtt{B})^\perp\big)$$

corresponding to this problem. According to (2.19), for $\boldsymbol{u} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega)$,

$$\mathtt{B}\boldsymbol{u} = 0 \Longleftrightarrow (\boldsymbol{u}, \nabla q)_{\boldsymbol{L}^2(\Omega)} = 0,\ \forall q \in H_0^1(\Omega) \Longleftrightarrow \operatorname{div}\boldsymbol{u} = 0.$$

Hence,

$$\operatorname{Ker}\mathtt{B} = \boldsymbol{K}_N(\Omega), \quad \text{where} \quad \boldsymbol{K}_N(\Omega) = \boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \cap \boldsymbol{H}(\operatorname{div}0;\Omega). \tag{2.62}$$

In addition, one easily checks that

$$(\operatorname{Ker}\mathtt{B})^\perp = \big\{\boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega),\ \exists q \in H_0^1(\Omega),\ \boldsymbol{v} = \nabla q\big\}. \tag{2.63}$$

With those results, we can characterize $\mathtt{L}_B$. On the one hand, by definition of b, we observe that

$$b(\mathtt{L}_B p, q) = \big(\mathtt{L}_B p, \nabla q\big)_{\boldsymbol{L}^2(\Omega)}, \quad \forall p, q \in H_0^1(\Omega). \tag{2.64}$$

On the other hand, according to (2.20), one has

$$b(\mathtt{L}_B p, q) = (\nabla p, \nabla q)_{\boldsymbol{L}^2(\Omega)}, \quad \forall p, q \in H_0^1(\Omega). \tag{2.65}$$

Putting (2.63), (2.64) and (2.65) together, we deduce that

$$\mathtt{L}_B p = \nabla p.$$

Moreover, for all $p \in H_0^1(\Omega)$, one has

$$\|\mathtt{L}_B p\|_{\boldsymbol{H}(\mathbf{curl}\,;\Omega)} = \|\nabla p\|_{\boldsymbol{H}(\mathbf{curl}\,;\Omega)} = \|\nabla p\|,$$

hence $\mathtt{L}_B$ is an isometry, so $\mathtt{L}_B$ satisfies (2.22) with $\underline{\beta} = 1$. The inf-sup condition (2.18) holds.

Going back to $\operatorname{Ker}\mathtt{B}$ (cf. (2.62)), we recall Weber inequality [Weber, 1980]: there exists $C_K > 1$ such that

$$\|\boldsymbol{k}\|_{\boldsymbol{H}(\mathbf{curl}\,;\Omega)} \leq C_K \|\mathbf{curl}\,\boldsymbol{k}\|, \quad \forall \boldsymbol{k} \in \boldsymbol{K}_N(\Omega).$$

The fact that not only $C_K > 0$, but even $C_K > 1$, stems from the definition of the "natural" norms involved. Hence, Weber inequality says that the form a is coercive on $\operatorname{Ker}\mathtt{B}$, so that all the conditions of Theorem 2.10 are fulfilled, with $\alpha_0 = (C_K)^{-2} < 1$. Precisely, Theorem 2.10 states that the bilinear form $\mathcal{A}$ is T-coercive for the mapping

$$\begin{aligned} \mathtt{T} : \boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \times H_0^1(\Omega) &\longrightarrow \boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \times H_0^1(\Omega) \\ (\boldsymbol{E},\tilde{p}) &\longmapsto \left(\boldsymbol{E} + \nabla\tilde{p}, -\tilde{p} + \frac{3}{4}\phi_{\boldsymbol{E}}\right), \end{aligned} \tag{2.66}$$

where $\phi_{\boldsymbol{E}} = \mathtt{B}\boldsymbol{E} \in H_0^1(\Omega)$. By definition of the operator B (cf. (2.19)), for any $\boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega)$, we have

$$(\nabla(\mathtt{B}\boldsymbol{v}), \nabla q)_{\boldsymbol{L}^2(\Omega)} = b(\boldsymbol{v}, q) = (\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(\Omega)}, \quad \forall q \in H_0^1(\Omega).$$

Therefore, $\mathtt{B}\boldsymbol{v}$ is the unique $\phi_{\boldsymbol{v}} \in H^1_0(\Omega)$ satisfying

$$(\nabla\phi_{\boldsymbol{v}}, \nabla q)_{\boldsymbol{L}^2(\Omega)} = (\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(\Omega)}, \quad \forall q \in H^1_0(\Omega).$$

Furthermore, following (2.31), it holds that

$$\mathcal{A}\big((\boldsymbol{E},\tilde{p}), \mathtt{T}(\boldsymbol{E},\tilde{p})\big) \geq \frac{(C_K)^{-2}}{4}\|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} + \frac{1}{2}\|\nabla\tilde{p}\|^2_{\boldsymbol{L}^2(\Omega)} \geq \underline{\alpha}\big(\|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} + \|\nabla\tilde{p}\|^2_{\boldsymbol{L}^2(\Omega)}\big),$$

with $\underline{\alpha} = \frac{(C_K)^{-2}}{4}$.

To get the stability constant, we need to compute $|||\mathtt{T}|||$, that is, bound $\|\mathtt{T}(\boldsymbol{E},\tilde{p})\|_{\boldsymbol{H}_0(\mathbf{curl};\Omega)\times H^1_0(\Omega)}$ for $(\boldsymbol{E},\tilde{p}) \in \boldsymbol{H}_0(\mathbf{curl};\Omega) \times H^1_0(\Omega)$. We find that

$$\begin{aligned}
\|\mathtt{T}(\boldsymbol{E},\tilde{p})\|^2_{\boldsymbol{H}_0(\mathbf{curl};\Omega)\times H^1_0(\Omega)} &= \|\boldsymbol{E} + \nabla\tilde{p}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} + \left\|-\nabla\tilde{p} + \frac{3}{4}\nabla\phi_{\boldsymbol{E}}\right\|^2_{\boldsymbol{L}^2(\Omega)} \\
&\leq 2\|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} + 2\|\nabla\tilde{p}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} + 2\|\nabla\tilde{p}\|^2_{\boldsymbol{L}^2(\Omega)} + 2\cdot\left(\frac{3}{4}\right)^2\|\nabla\phi_{\boldsymbol{E}}\|^2_{\boldsymbol{L}^2(\Omega)} \\
&\leq 2\|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} + 4\|\nabla\tilde{p}\|^2_{\boldsymbol{L}^2(\Omega)} + 2\cdot\left(\frac{3}{4}\right)^2\|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} \\
&\leq 4\big(\|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega)} + \|\nabla\tilde{p}\|^2_{\boldsymbol{L}^2(\Omega)}\big),
\end{aligned}$$

where we used that $\|\nabla\phi_{\boldsymbol{E}}\|_{\boldsymbol{L}^2(\Omega)} = \|\nabla(\mathtt{B}\boldsymbol{E})\|_{\boldsymbol{L}^2(\Omega)} \leq \|\boldsymbol{E}\|_{\boldsymbol{H}(\mathbf{curl};\Omega)}$ thanks to (2.32). Therefore, $|||\mathtt{T}||| \leq 2$.

Applying (2.4), we conclude that

$$\|\boldsymbol{E}\|_{\boldsymbol{H}(\mathbf{curl};\Omega)} \leq 8C_K^2\big(\|\boldsymbol{H}^\star\|^2 + \|\rho\|^2_{H^{-1}(\Omega)}\big)^{1/2}. \tag{2.67}$$

2.3.1.2 In an anisotropic medium

In an anisotropic medium, let us follow for instance [Ciarlet Jr, 2020] to build an all-in-one equivalent variational formulation. In this case, the electromagnetic energy can be expressed as $(\underline{\varepsilon}\boldsymbol{E}, \boldsymbol{E})_{\boldsymbol{L}^2(\Omega)} + (\underline{\mu}^{-1}\mathbf{curl}\,\boldsymbol{E}, \mathbf{curl}\,\boldsymbol{E})_{\boldsymbol{L}^2(\Omega)}$. Under the assumption (2.57) made on $\underline{\varepsilon}$ and $\underline{\mu}$, we note that we can endow $\boldsymbol{H}_0(\mathbf{curl};\Omega)$ with the inner product $(\cdot,\cdot)_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} : (\boldsymbol{u},\boldsymbol{v}) \mapsto (\underline{\varepsilon}\boldsymbol{u},\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} + (\underline{\mu}^{-1}\mathbf{curl}\,\boldsymbol{u}, \mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}$. The associated scaled norm

$$\|\boldsymbol{u}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} = \big((\underline{\varepsilon}\boldsymbol{u},\boldsymbol{u})_{\boldsymbol{L}^2(\Omega)} + (\underline{\mu}^{-1}\mathbf{curl}\,\boldsymbol{u}, \mathbf{curl}\,\boldsymbol{u})_{\boldsymbol{L}^2(\Omega)}\big)^{1/2}$$

is equivalent to the "natural" norm. Then, we endow $H^1_0(\Omega)$ with the inner product $(\cdot,\cdot)_{1,\underline{\varepsilon}} : (p,q) \mapsto (\underline{\varepsilon}\nabla p, \nabla q)_{\boldsymbol{L}^2(\Omega)}$, and the associated scaled norm

$$\|q\|_{1,\underline{\varepsilon}} = \big((\underline{\varepsilon}\nabla q, \nabla q)_{\boldsymbol{L}^2(\Omega)}\big)^{1/2}$$

is equivalent to $\|\cdot\|_{H^1(\Omega)}$ according to Poincaré inequality. With this choice of norms, for $q \in H^1_0(\Omega)$, one has $\|q\|_{1,\underline{\varepsilon}} = \|\nabla q\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}$. Also, $\mathbb{1}_{H^1_0(\Omega)\to H^{-1}(\Omega)}$ is the isomorphism defined by

$$\langle \mathbb{1}_{H^1_0(\Omega)\to H^{-1}(\Omega)} p, q\rangle_{H^{-1}(\Omega),H^1_0(\Omega)} = (p,q)_{1,\underline{\varepsilon}} = (\underline{\varepsilon}\nabla p, \nabla q)_{\boldsymbol{L}^2(\Omega)}, \quad \forall p,q \in H^1_0(\Omega),$$

while the norm in $H^{-1}(\Omega)$ is

$$\|g\|_{-1,\underline{\varepsilon}^{-1}} = \sup_{q\in H^1_0(\Omega)\setminus\{0\}} \frac{\langle g, q\rangle_{H^{-1}(\Omega),H^1_0(\Omega)}}{\|q\|_{1,\underline{\varepsilon}}}, \quad \forall g \in H^{-1}(\Omega).$$

Finally, for $\xi \in \{\underline{\varepsilon}, \underline{\varepsilon}^{-1}, \underline{\mu}, \underline{\mu}^{-1}\}$, we use the inner product $(\cdot,\cdot)_\xi : (\boldsymbol{u},\boldsymbol{v}) \mapsto (\xi\boldsymbol{u},\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}$, and the associated scaled norm $\|\cdot\|_\xi$ in $\boldsymbol{L}^2(\Omega)$. As we shall see below, these scaled norms and inner products, which are introduced to account for the anisotropic medium, lead to computations that are very similar to those that have been carried out for a homogeneous medium.
As before, in order to fit (2.58) into the abstract framework (2.15), we introduce a vanishing artificial pressure $\tilde{p}$. The resulting formulation is: find $(\boldsymbol{E},\tilde{p}) \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \times H_0^1(\Omega)$ such that

$$\begin{aligned} (\underline{\mu}^{-1}\mathbf{curl}\,\boldsymbol{E},\mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} + (\underline{\varepsilon}\boldsymbol{v},\nabla\tilde{p})_{\boldsymbol{L}^2(\Omega)} &= (\boldsymbol{H}^\star,\mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}, \quad && \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \\ (\underline{\varepsilon}\boldsymbol{E},\nabla q)_{\boldsymbol{L}^2(\Omega)} &= -\langle \rho, q\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}, \quad && \forall q \in H_0^1(\Omega). \end{aligned} \tag{2.68}$$

So, defining the bilinear forms

$$\begin{aligned} a_{\underline{\mu}^{-1}}(\boldsymbol{u},\boldsymbol{v}) &= (\underline{\mu}^{-1}\mathbf{curl}\,\boldsymbol{u},\mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)}, \quad && \forall \boldsymbol{u} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \\ b_{\underline{\varepsilon}}(\boldsymbol{v},q) &= (\underline{\varepsilon}\boldsymbol{v},\nabla q)_{\boldsymbol{L}^2(\Omega)}, && \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega), \forall q \in H_0^1(\Omega), \end{aligned}$$

the *all-in-one* bilinear form of Maxwell problem is now given by

$$\mathcal{A}_{\underline{\varepsilon},\underline{\mu}}\big((\boldsymbol{E},\tilde{p}),(\boldsymbol{v},q)\big) = a_{\underline{\mu}^{-1}}(\boldsymbol{E},\boldsymbol{v}) + b_{\underline{\varepsilon}}(\boldsymbol{v},\tilde{p}) + b_{\underline{\varepsilon}}(\boldsymbol{E},q), \tag{2.69}$$

while the linear form remains defined by (2.61). Thanks to the introduction of scaled norms, we find that the bilinear form $a_{\underline{\mu}^{-1}}$ is continuous on $\boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \times \boldsymbol{H}_0(\mathbf{curl}\,;\Omega)$ with a continuity constant $C_a = 1$, while the bilinear form $b_{\underline{\varepsilon}}$ is continuous on $\boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \times H_0^1(\Omega)$ with a continuity constant $C_b = 1$. With respect to the scaled norms, we have

$$\big|(\boldsymbol{H}^\star,\mathbf{curl}\,\boldsymbol{v})_{\boldsymbol{L}^2(\Omega)} - \langle \rho, q\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}\big| \le \|\boldsymbol{H}^\star\|_{\underline{\mu}}\|\mathbf{curl}\,\boldsymbol{v}\|_{\underline{\mu}^{-1}} + \|\rho\|_{-1,\underline{\varepsilon}^{-1}}\|q\|_{1,\underline{\varepsilon}},$$

so that $C_\ell \le \big(\|\boldsymbol{H}^\star\|_{\underline{\mu}}^2 + \|\rho\|_{-1,\underline{\varepsilon}^{-1}}^2\big)^{1/2}$. Let us give an explicit expression of the abstract operators

$$\mathtt{B}_{\underline{\varepsilon}} \in \mathcal{L}\big(\boldsymbol{H}_0(\mathbf{curl}\,;\Omega), H_0^1(\Omega)\big), \qquad \mathtt{L}_{B_{\underline{\varepsilon}}} \in \mathcal{L}\big(H_0^1(\Omega), (\operatorname{Ker}\mathtt{B}_{\underline{\varepsilon}})^\perp\big).$$

Given $\boldsymbol{u} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega)$, we observe that, by definition of operator $\mathtt{B}_{\underline{\varepsilon}}$ (cf. (2.19))

$$\mathtt{B}_{\underline{\varepsilon}}\boldsymbol{u} = 0 \iff (\underline{\varepsilon}\boldsymbol{u},\nabla q)_{\boldsymbol{L}^2(\Omega)} = 0,\ \forall q \in H_0^1(\Omega) \iff \operatorname{div}(\underline{\varepsilon}\boldsymbol{u}) = 0.$$

Hence,

$$\operatorname{Ker}\mathtt{B}_{\underline{\varepsilon}} = \boldsymbol{K}_N(\Omega;\underline{\varepsilon}), \quad \text{where} \quad \boldsymbol{K}_N(\Omega;\underline{\varepsilon}) = \big\{\boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega),\ \operatorname{div}(\underline{\varepsilon}\boldsymbol{v}) = 0\big\}. \tag{2.70}$$

In addition (see *e.g.* (6.16) in [Assous et al., 2018])

$$(\operatorname{Ker}\mathtt{B}_{\underline{\varepsilon}})^\perp = \big\{\boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega),\ \exists q \in H_0^1(\Omega),\ \boldsymbol{v} = \nabla q\big\}, \tag{2.71}$$

where orthogonality is understood with respect to the inner product $(\cdot,\cdot)_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}\,\cdot}$.
With those results, we can characterize $\mathtt{L}_{B_{\underline{\varepsilon}}}$. By definition of $b_{\underline{\varepsilon}}$, we observe that

$$b_{\underline{\varepsilon}}(\mathtt{L}_{B_{\underline{\varepsilon}}}p,q) = \big(\underline{\varepsilon}(\mathtt{L}_{B_{\underline{\varepsilon}}}p),\nabla q\big)_{\boldsymbol{L}^2(\Omega)}, \quad \forall p,q \in H_0^1(\Omega). \tag{2.72}$$

While, according to (2.20), one has

$$b_{\underline{\varepsilon}}(\mathtt{L}_{B_{\underline{\varepsilon}}}p,q) = (p,q)_{1,\underline{\varepsilon}} = (\underline{\varepsilon}\nabla p,\nabla q)_{\boldsymbol{L}^2(\Omega)}, \quad \forall p,q \in H_0^1(\Omega). \tag{2.73}$$

Putting (2.71), (2.72) and (2.73) together, we deduce that $\mathtt{L}_{B_{\underline{\varepsilon}}}p = \nabla p$. So, for all $p \in H_0^1(\Omega)$, it follows that $\|\mathtt{L}_{B_{\underline{\varepsilon}}}p\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} = \|p\|_{1,\underline{\varepsilon}}$. In other words, $\mathtt{L}_{B_{\underline{\varepsilon}}}$ is an isometry with respect to the scaled

norms: $\mathtt{L}_{B_{\underline{\varepsilon}}}$ satisfies (2.22) with $\underline{\beta} = 1$, and the inf-sup condition (2.18) holds.
Going back to $\operatorname{Ker}\mathtt{B}_{\underline{\varepsilon}}$ (cf. (2.70)), we recall the generalized Weber inequality [Weber, 1980] (or [Assous et al., 2018, Theorem 6.1.4]): there exists $C_K > 1$ such that

$$\|\boldsymbol{k}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} \leq C_K \|\mathbf{curl}\,\boldsymbol{k}\|_{\underline{\mu}^{-1}}, \quad \forall \boldsymbol{k} \in \boldsymbol{K}_N(\Omega;\underline{\varepsilon}).$$

While the bound $C_K > 1$ remains as a consequence of the definition of the scaled norms, the value of the constant C_K now possibly depends on $\underline{\varepsilon}$ or $\underline{\mu}$.
The generalized Weber inequality implies that the form $a_{\underline{\mu}^{-1}}$ is coercive on $\operatorname{Ker}\mathtt{B}_{\underline{\varepsilon}}$: all the conditions of Theorem 2.10 are fulfilled, with $\alpha_0 = (C_K)^{-2} < 1$. Interestingly, Theorem 2.10 states that the bilinear form $\mathcal{A}_{\underline{\varepsilon},\underline{\mu}}$ is $\mathtt{T}$-coercive for the mapping $\mathtt{T}$ that is again given by (2.66), but with the $\underline{\varepsilon}$-dependent $\phi_{\boldsymbol{E}} = \mathtt{B}_{\underline{\varepsilon}}\boldsymbol{E} \in H^1_0(\Omega)$. Based on this observation, the final computations are very close to those of Section 2.3.1.1, replacing the "natural" norms and inner products by their scaled counterparts.

First, using (2.19), we find that $\mathtt{B}_{\underline{\varepsilon}}\boldsymbol{v}$ is the unique $\phi_{\boldsymbol{v}} \in H^1_0(\Omega)$ satisfying

$$(\underline{\varepsilon}\nabla\phi_{\boldsymbol{v}}, \nabla q)_{\boldsymbol{L}^2(\Omega)} = (\underline{\varepsilon}\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(\Omega)}, \quad \forall q \in H^1_0(\Omega). \tag{2.74}$$

Second, following (2.31) and introducing $\underline{\alpha} = \frac{(C_K)^{-2}}{4}$ (which depends on $\underline{\varepsilon}$ or $\underline{\mu}$), it holds that

$$\mathcal{A}_{\underline{\varepsilon},\underline{\mu}}\big((\boldsymbol{E},\tilde{p}),\mathtt{T}(\boldsymbol{E},\tilde{p})\big) \geq \underline{\alpha}\big(\|\boldsymbol{E}\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} + \|\tilde{p}\|^2_{1,\underline{\varepsilon}}\big).$$

Finally, thanks to (2.32), which yields $\|\phi_{\boldsymbol{E}}\|_{1,\underline{\varepsilon}} \leq \|\boldsymbol{E}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}$, we now find that

$$\begin{aligned}\|\mathtt{T}(\boldsymbol{E},\tilde{p})\|^2_{\boldsymbol{H}_0(\mathbf{curl};\Omega)\times H^1_0(\Omega)} &= \|\boldsymbol{E} + \nabla\tilde{p}\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} + \left\|-\tilde{p} + \frac{3}{4}\phi_{\boldsymbol{E}}\right\|^2_{1,\underline{\varepsilon}} \\ &\leq 4\big(\|\boldsymbol{E}\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} + \|\tilde{p}\|^2_{1,\underline{\varepsilon}}\big).\end{aligned}$$

Using the scaled norms, we have again that $|||\mathtt{T}||| \leq 2$, and we conclude with (2.4) that

$$\|\boldsymbol{E}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} \leq 8C_K^2\big(\|\boldsymbol{H}^\star\|^2_{\underline{\mu}} + \|\rho\|^2_{-1,\underline{\varepsilon}^{-1}}\big)^{1/2}. \tag{2.75}$$

2.3.2 Optimized bounds in an anisotropic medium

To achieve $\mathtt{T}$-coercivity, the abstract theory does not take into account the so-called double orthogonality property (or Helmholtz decomposition), which states that for all $\boldsymbol{k} \in \boldsymbol{K}_N(\Omega;\underline{\varepsilon})$ and all $q \in H^1_0(\Omega)$, one has $(\underline{\varepsilon}\boldsymbol{k}, \nabla q)_{\boldsymbol{L}^2(\Omega)} = (\underline{\mu}^{-1}\mathbf{curl}\,\boldsymbol{k}, \mathbf{curl}\,(\nabla q))_{\boldsymbol{L}^2(\Omega)} = 0$, so that

$$\|\boldsymbol{k} + \nabla q\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} = \|\boldsymbol{k}\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} + \|q\|^2_{1,\underline{\varepsilon}}.$$

Indeed, given $\boldsymbol{E} \in \boldsymbol{H}_0(\mathbf{curl};\Omega)$, we note that, with the help of $\phi_{\boldsymbol{E}} \in H^1_0(\Omega)$ solving (2.74), one has the (orthogonal) Helmholtz decomposition $\boldsymbol{E} = \boldsymbol{k}_{\boldsymbol{E}} + \nabla\phi_{\boldsymbol{E}}$, with $\boldsymbol{k}_{\boldsymbol{E}} \in \boldsymbol{K}_N(\Omega;\underline{\varepsilon})$.

We sketch below how one can improve the estimates, see [Ciarlet Jr, 2021] for further details. Let us choose

$$\begin{aligned}\mathtt{T}_{opt} : \boldsymbol{H}_0(\mathbf{curl};\Omega) \times H^1_0(\Omega) &\longrightarrow \boldsymbol{H}_0(\mathbf{curl};\Omega) \times H^1_0(\Omega) \\ (\boldsymbol{E},\tilde{p}) &\longmapsto \big(\boldsymbol{k}_{\boldsymbol{E}} + \nabla\tilde{p}, \phi_{\boldsymbol{E}}\big).\end{aligned}$$

Thanks to the double orthogonality property, one finds easily that $\mathtt{T}_{opt}$ is an isometry and that

$$\begin{aligned}\mathcal{A}_{\underline{\varepsilon},\underline{\mu}}\big((\boldsymbol{E},\tilde{p}),\mathtt{T}_{opt}(\boldsymbol{E},\tilde{p})\big) &= \|\mathbf{curl}\,\boldsymbol{k}_{\boldsymbol{E}}\|^2_{\underline{\mu}^{-1}} + \|\tilde{p}\|^2_{1,\underline{\varepsilon}} + \|\phi_{\boldsymbol{E}}\|^2_{1,\underline{\varepsilon}} \\ &\geq (C_K)^{-2}\big(\|\boldsymbol{E}\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} + \|\tilde{p}\|^2_{1,\underline{\varepsilon}}\big),\end{aligned}$$

where C_K is the constant that appears in the generalized Weber inequality. Applying (2.4), we have the optimized stability estimate

$$\|\boldsymbol{E}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} \leq C_K^2 \big(\|\boldsymbol{H}^\star\|_{\underline{\mu}}^2 + \|\rho\|_{-1,\underline{\varepsilon}^{-1}}^2\big)^{1/2}. \tag{2.76}$$

We conclude that, for all possible choices of coefficients $\underline{\varepsilon}$ and $\underline{\mu}$, there is only a factor 8 difference between the stability constant obtained via the abstract T-coercivity approach, see (2.75), and the optimized stability constant which relies explicitly on the double orthogonality property, see (2.76). This shows the robustness of the abstract theory.

Remark 2.24. One can obtain similar results in more general geometries, such as a non-simply-connected domain, or a non-connected boundary, see [Ciarlet Jr, 2021].

2.4 Application to nearly-incompressible elasticity

In this section, we apply the T-coercivity framework to the equations of elasticity, assuming homogeneous Dirichlet boundary conditions. Let $\Omega \subset \mathbb{R}^d$ be a domain, where $2 \leq d \leq 3$. For a prescribed body force $\boldsymbol{f} \in \boldsymbol{H}^{-1}(\Omega)$, we look for the displacement $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega)$ such that

$$\begin{aligned} -\operatorname{div}\big(\sigma(\boldsymbol{u})\big) &= \boldsymbol{f}, \quad && \text{in } \Omega, \\ \boldsymbol{u} &= 0, && \text{on } \partial\Omega, \end{aligned} \tag{2.77}$$

where $\sigma(\boldsymbol{u})$ denotes the stress tensor. We assume that it is given by Hooke's law

$$\sigma(\boldsymbol{u}) = 2\mu\,\varepsilon(\boldsymbol{u}) + \lambda(\operatorname{div}\boldsymbol{u})I,$$

where $\lambda, \mu > 0$ are the Lamé coefficients of the material and $\varepsilon(\boldsymbol{u}) = \frac{1}{2}\big(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^T\big)$ is the linearized strain tensor. Thanks to Korn inequality [Duvaut and Lions, 1972], the space $\boldsymbol{H}_0^1(\Omega)$ is here endowed with the inner product

$$(\boldsymbol{u}, \boldsymbol{v}) \longmapsto \int_\Omega \varepsilon(\boldsymbol{u}) : \varepsilon(\boldsymbol{v})\,\mathrm{d}x,$$

whose associated norm $\boldsymbol{u} \mapsto \|\varepsilon(\boldsymbol{u})\|$ is equivalent to the $\boldsymbol{H}^1(\Omega)$-norm in $\boldsymbol{H}_0^1(\Omega)$. Introducing the new unknown $p = \lambda \operatorname{div}\boldsymbol{u}$, the elasticity system (2.77) can be written in mixed form as follows: find $\boldsymbol{u} \in \boldsymbol{H}_0^1(\Omega)$ and $p \in \mathrm{L}_0^2(\Omega)$ such that

$$\begin{aligned} -2\mu \operatorname{div}\big(\varepsilon(\boldsymbol{u})\big) - \nabla p &= \boldsymbol{f}, \quad && \text{in } \Omega, \\ \operatorname{div}\boldsymbol{u} - \frac{1}{\lambda}p &= 0, && \text{in } \Omega. \end{aligned}$$

Or equivalently, in variational form: find $\boldsymbol{u} \in \boldsymbol{H}_0^1(\Omega)$ and $p \in \mathrm{L}_0^2(\Omega)$ such that

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) &= \langle \boldsymbol{f}, \boldsymbol{v} \rangle, \quad && \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega), \\ b(\boldsymbol{u}, q) - \frac{1}{\lambda} c_0(p, q) &= 0, && \forall q \in \mathrm{L}_0^2(\Omega), \end{aligned} \tag{2.78}$$

with

$$a(\boldsymbol{u}, \boldsymbol{v}) = 2\mu \int_\Omega \varepsilon(\boldsymbol{u}) : \varepsilon(\boldsymbol{v})\,\mathrm{d}x, \quad b(\boldsymbol{v}, q) = \int_\Omega q \operatorname{div}\boldsymbol{v}\,\mathrm{d}x \quad \text{and} \quad c_0(p, q) = \int_\Omega pq\,\mathrm{d}x.$$

For nearly-incompressible materials, the first Lamé coefficient λ goes to infinity, so that λ^{-1} goes to zero. Therefore, (2.78) can be seen as a small perturbation of Stokes system.

Since the bilinear form a is coercive on the whole space $\boldsymbol{H}_0^1(\Omega)$, we can directly apply Theorem 2.12 in the special case of Remark 2.13. The bilinear form a is continuous and coercive, with

$C_a = \alpha = 2\mu$. In addition, the bilinear form b is continuous and satisfies the inf-sup condition (2.18) with $\beta = C_{\rm div}$ since b is the same form – except to the sign – as for Stokes problem. Then, Theorem 2.12 furnishes that the *all-in-one* bilinear form $\mathcal{A}_c$ defined by

$$\mathcal{A}_c\big((\boldsymbol{u},p),(\boldsymbol{v},q)\big) = 2\mu\int_\Omega \varepsilon(\boldsymbol{u}) : \varepsilon(\boldsymbol{v})\,\mathrm{d}x + \int_\Omega p\,\mathrm{div}\,\boldsymbol{v}\,\mathrm{d}x + \int_\Omega q\,\mathrm{div}\,\boldsymbol{u}\,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x \tag{2.79}$$

is T-coercive for the mapping

$$\begin{aligned} \mathtt{T} : \boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega) &\longrightarrow \boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega)\\ (\boldsymbol{u},p) &\longmapsto \big(2\mu C^2_{\rm div}\,\boldsymbol{u} + \boldsymbol{v}_{-p}, -2\mu C^2_{\rm div}\,p\big), \end{aligned} \tag{2.80}$$

and (2.38) implies that

$$\mathcal{A}_c\big((\boldsymbol{u},p),\mathtt{T}(\boldsymbol{u},p)\big) \geq 2\mu^2 C^2_{\rm div}\,\|\varepsilon(\boldsymbol{u})\|^2 + \left(\frac{1}{2} + \frac{2\mu}{\lambda}C^2_{\rm div}\right)\|p\|^2.$$

Note that this estimate is robust in the incompressible limit, namely for large values of λ.

Finally, replacing ν by 2μ in (2.11) and using (2.4), we get that the unique solution of (2.78) satisfies

$$\|(\boldsymbol{u},p)\|_{\boldsymbol{H}^1_0(\Omega)\times \mathrm{L}^2_0(\Omega)} \leq \frac{2\max\Big(2\sqrt{2}\mu C^2_{\rm div}, C_{\rm div}\big(2+4\mu^2 C^2_{\rm div}\big)^{1/2}\Big)}{\min(4\mu^2 C^2_{\rm div}, 1+4\mu\lambda^{-1}C^2_{\rm div})}\|\boldsymbol{f}\|_{(\boldsymbol{H}^1_0(\Omega))'},$$

where $\|\cdot\|_{(\boldsymbol{H}^1_0(\Omega))'}$ denotes the dual norm of $\|\varepsilon(\cdot)\|$.

Remark 2.25. Here again, the stability constant obtained above depends on the choice of the norms for $\boldsymbol{u}$ and p. In particular, it is possible to remove the dependence on the first Lamé coefficient by considering the scaled norm defined by $\|\boldsymbol{u}\|_\mu = 2\mu\int_\Omega \varepsilon(\boldsymbol{u}) : \varepsilon(\boldsymbol{u})\,\mathrm{d}x$ for all $\boldsymbol{u}\in\boldsymbol{H}^1_0(\Omega)$.

2.5 Application to neutron diffusion

Let $\Omega\subset\mathbb{R}^d$ be a domain, where $2\leq d\leq 3$. We consider the neutron diffusion equation with zero flux boundary condition: given a prescribed fission source $S_f\in L^2(\Omega)$, find $u\in H^1(\Omega)$ such that

$$\begin{aligned} -\mathrm{div}\,(D\nabla u) + \sigma u &= S_f, \quad && \text{in } \Omega,\\ u &= 0, && \text{on } \partial\Omega, \end{aligned} \tag{2.81}$$

where u, D, and σ denote respectively the neutron flux, the diffusion coefficient and the macroscopic absorption cross section. It is assumed that the diffusion coefficient D fulfills (2.57), and that the macroscopic absorption cross section is such that

$$\begin{cases} \sigma \text{ is a real-valued measurable scalar field on } \Omega,\\ \exists \sigma_-,\sigma_+ > 0,\ \sigma_-\leq\sigma\leq\sigma_+ \text{ a.e. in } \Omega. \end{cases} \tag{2.82}$$

Because $S_f\in L^2(\Omega)$, one has $D\nabla u\in\boldsymbol{H}(\mathrm{div}\,;\Omega)$. This problem can be recast equivalently in mixed form, introducing the auxiliary unknown $\boldsymbol{p} = -D\nabla u$, called the neutron current. It reads: find $(u,\boldsymbol{p})\in H^1_0(\Omega)\times\boldsymbol{H}(\mathrm{div}\,;\Omega)$ such that

$$\begin{aligned} \mathrm{div}\,\boldsymbol{p} + \sigma u &= S_f, \quad && \text{in } \Omega,\\ D^{-1}\boldsymbol{p} + \nabla u &= 0, && \text{in } \Omega. \end{aligned} \tag{2.83}$$

It can be shown that equivalent weak form is: find $(u,\boldsymbol{p})\in L^2(\Omega)\times\boldsymbol{H}(\mathrm{div}\,;\Omega)$ such that

$$\int_\Omega \big(v\,\mathrm{div}\,\boldsymbol{p} + \sigma uv - D^{-1}\boldsymbol{p}\cdot\boldsymbol{q} + u\,\mathrm{div}\,\boldsymbol{q}\big)\,\mathrm{d}x = \int_\Omega S_f v\,\mathrm{d}x \quad \forall (v,\boldsymbol{q})\in L^2(\Omega)\times\boldsymbol{H}(\mathrm{div}\,;\Omega). \tag{2.84}$$

Remark 2.26. Among other things, one can recover that the solution $u \in L^2(\Omega)$ from the weak form (2.84) is such that $u \in H^1(\Omega)$, and that $u = 0$ on $\partial\Omega$.

Defining the bilinear forms

$$\begin{aligned} a_{D^{-1}}(\boldsymbol{p},\boldsymbol{q}) &= (D^{-1}\boldsymbol{p},\boldsymbol{q})_{\boldsymbol{L}^2(\Omega)}, \quad && \forall \boldsymbol{p} \in \boldsymbol{H}(\mathrm{div}\,;\Omega), \forall \boldsymbol{q} \in \boldsymbol{H}(\mathrm{div}\,;\Omega), \\ b(\boldsymbol{q},v) &= -(\mathrm{div}\,\boldsymbol{q},v)_{L^2(\Omega)}, && \forall \boldsymbol{q} \in \boldsymbol{H}(\mathrm{div}\,;\Omega), \forall v \in L^2(\Omega), \\ c_\sigma(u,v) &= (\sigma u,v)_{L^2(\Omega)}, && \forall u \in L^2(\Omega), \forall v \in L^2(\Omega), \end{aligned}$$

the *all-in-one* bilinear form of the diffusion problem is given by

$$\mathcal{A}_c\big((\boldsymbol{p},u),(\boldsymbol{q},v)\big) = a_{D^{-1}}(\boldsymbol{p},\boldsymbol{q}) + b(\boldsymbol{q},u) + b(\boldsymbol{p},v) - c_\sigma(u,v). \tag{2.85}$$

Here, we are in the case of a "fixed" augmentation, as treated in Section 2.2.6.

Let us check below that all the conditions of Theorem 2.20 are fulfilled. First, c_σ is coercive on $L^2(\Omega)$ with $\gamma = \sigma_-$. Then, $a_{D^{-1}}$ fulfills (2.13) with $C_a = (D_-)^{-1}$, whereas b fulfills (2.14) with $C_b = 1$. Finally, we look for the condition (2.54). It is straightforward to check that, for all $\boldsymbol{p} \in \boldsymbol{H}(\mathrm{div}\,;\Omega)$, $B\boldsymbol{p} = \mathtt{B}\boldsymbol{p} = -\mathrm{div}\,\boldsymbol{p}$. Hence

$$\begin{aligned} a_{D^{-1}}(\boldsymbol{p},\boldsymbol{p}) + \frac{\gamma}{2C_c^2}\|B\boldsymbol{p}\|^2 &= (D^{-1}\boldsymbol{p},\boldsymbol{p}) + \frac{\sigma_-}{2\sigma_+^2}\|\mathrm{div}\,\boldsymbol{p}\|^2 \\ &\geq \min\Big((D_+)^{-1}, \frac{\sigma_-}{2\sigma_+^2}\Big)\|\boldsymbol{p}\|^2_{\boldsymbol{H}(\mathrm{div}\,;\Omega)}. \end{aligned}$$

Then, Theorem 2.20 establishes that the bilinear form $\mathcal{A}_c$ is T-coercive for the mapping (2.55)

$$\begin{aligned} \mathtt{T} : \boldsymbol{H}(\mathrm{div}\,;\Omega) \times \mathrm{L}^2(\Omega) &\longrightarrow \boldsymbol{H}(\mathrm{div}\,;\Omega) \times \mathrm{L}^2(\Omega) \\ (\boldsymbol{p},u) &\longmapsto \Big(\boldsymbol{p}, \frac{1}{2}(-u - \sigma^{-1}\mathrm{div}\,\boldsymbol{p})\Big). \end{aligned}$$

Furthermore, using the estimate (2.56), it holds that

$$\mathcal{A}_c\big((\boldsymbol{p},u),\mathtt{T}(\boldsymbol{p},u)\big) \geq \min\Big((D_+)^{-1}, \frac{\sigma_-}{2\sigma_+^2}\Big)\|\boldsymbol{p}\|^2_{\boldsymbol{H}(\mathrm{div}\,;\Omega)} + \frac{\sigma_-}{2}\|u\|^2 \geq \underline{\alpha}\|(\boldsymbol{p},u)\|^2_{\boldsymbol{H}(\mathrm{div}\,;\Omega)\times L^2(\Omega)}, \tag{2.86}$$

with $\underline{\alpha} = \frac{1}{2}\min\big(2(D_+)^{-1}, \sigma_-(\sigma_+)^{-2}, \sigma_-\big)$.

There remains to estimate $|||\mathtt{T}|||$. One has

$$\begin{aligned} \|\mathtt{T}(\boldsymbol{p},u)\|^2_{\boldsymbol{H}(\mathrm{div}\,;\Omega)\times L^2(\Omega)} &= \|\boldsymbol{p}\|^2_{\boldsymbol{H}(\mathrm{div}\,;\Omega)} + \frac{1}{4}\|-u - \sigma^{-1}\mathrm{div}\,\boldsymbol{p}\|^2_{L^2(\Omega)} \\ &\leq \|\boldsymbol{p}\|^2_{\boldsymbol{H}(\mathrm{div}\,;\Omega)} + \frac{1}{4}\left((1+3)\|u\|^2_{L^2(\Omega)} + \Big(1+\frac{1}{3}\Big)(\sigma_-)^{-2}\|\mathrm{div}\,\boldsymbol{p}\|^2_{L^2(\Omega)}\right) \\ &\leq \Big(1 + \frac{1}{3}(\sigma_-)^{-2}\Big)\,\|(\boldsymbol{p},u)\|^2_{\boldsymbol{H}(\mathrm{div}\,;\Omega)\times L^2(\Omega)}, \end{aligned}$$

so that $|||\mathtt{T}||| \leq \big(1 + \frac{1}{3}(\sigma_-)^{-2}\big)^{1/2}$. Applying (2.4), we conclude that

$$\|(\boldsymbol{p},u)\|_{\boldsymbol{H}(\mathrm{div}\,;\Omega)\times L^2(\Omega)} \leq \frac{2\big(1 + \frac{1}{3}(\sigma_-)^{-2}\big)^{1/2}}{\min\big(2(D_+)^{-1}, \sigma_-(\sigma_+)^{-2}, \sigma_-\big)}\|S_f\|_{L^2(\Omega)}.$$

Remark 2.27. Some of those computations can be found in [Jamelot and Ciarlet Jr, 2013; Ciarlet Jr et al., 2017]. Here, we see them as a consequence of the general result stated in Theorem 2.20. Note that in [Jamelot and Ciarlet Jr, 2013; Ciarlet Jr et al., 2017], the T-coercivity estimate (2.86) is obtained with a constant $\underline{\alpha}' = \frac{1}{2}\min\big(2(D_+)^{-1}, (\sigma_+)^{-1}, \sigma_-\big)$, which is very close to $\underline{\alpha}$ since $\sigma_-(\sigma_+)^{-2} = (\sigma_+)^{-1} \cdot \frac{\sigma_-}{\sigma_+}$ and $\frac{\sigma_-}{\sigma_+} \leq 1$.

Remark 2.28. Here again, the operator $\mathtt{T}$ is not unique, see for instance [Ern and Guermond, 2021b, Exercise 56.6] for another possible choice of $\mathtt{T}$-coercive operator with a slightly different weak formulation of (2.81).

Remark 2.29. If one wants to obtain estimates without the bounding factors $\sigma_\pm$ and $D_\pm$, a standard path is to imbed the parameters D and σ into the definition of the norms, like it is done in Section 2.3. Namely, one chooses the norms:

$$\begin{aligned}
&\|v\|_\sigma = \big((\sigma v, v)_{L^2(\Omega)}\big)^{1/2},\\
&\|\boldsymbol{q}\|_{D^{-1},\sigma^{-1}\mathrm{div}} = \big((D^{-1}\boldsymbol{q},\boldsymbol{q})_{\boldsymbol{L}^2(\Omega)} + (\sigma^{-1}\mathrm{div}\,\boldsymbol{q},\mathrm{div}\,\boldsymbol{q})_{L^2(\Omega)}\big)^{1/2},\\
&\|(\boldsymbol{q},v)\|_V = \big(\|v\|_\sigma^2 + \|\boldsymbol{q}\|^2_{D^{-1},\sigma^{-1}\mathrm{div}}\big)^{1/2}.
\end{aligned}$$

On the one hand, all norms are "fixed" once the parameters are given. On the other hand, one can easily check that the stability constant is now independent of the bounding factors, by using the same mapping $\mathtt{T}$ as before.

2.6 T-coercivity at the discrete level

Previously, we demonstrated the robustness and the flexibility of the $\mathtt{T}$-coercivity approach to study mixed problems at the continuous level. In this section, we are going to see how $\mathtt{T}$-coercivity also enables us to provide a stable discretization of such problems with mixed finite elements. Let us recall the simple results below [Ciarlet Jr, 2012; Chesnel and Ciarlet, 2013].

Definition 2.30. [Chesnel and Ciarlet, 2013, Definition 5] Let W be a Hilbert space, $\mathcal{A}(\cdot,\cdot)$ be a continuous bilinear form over $W\times W$ and $(W_h)_h$ be conforming approximations of W. We say that $\mathcal{A}$ is *uniformly $\mathtt{T}_h$-coercive* if

$$\exists\alpha^*,\beta^*>0,\ \forall h>0,\ \exists\mathtt{T}_h\in\mathcal{L}(W_h),\quad |\mathcal{A}(u_h,\mathtt{T}_h u_h)|\geq\alpha^*\|u_h\|_W^2,\quad \forall u_h\in W_h,\quad\text{and}\quad |||\mathtt{T}_h|||\leq\beta^*.$$

Proposition 2.31. *[Chesnel and Ciarlet, 2013, Theorem 2] Let W be a Hilbert space, f be an element of W', $\mathcal{A}(\cdot,\cdot)$ be a continuous bilinear form over $W\times W$ and $(W_h)_h$ be conforming approximations of W. Denote by $\boldsymbol{A}_h\in\mathcal{L}(W_h,W_h')$ the discrete operator associated to $\mathcal{A}_{|W_h}$. The problem*

$$\begin{cases}\text{Find } u_h\in W_h \quad \text{such that}\\ \forall v_h\in W_h,\quad \mathcal{A}(u_h,v_h)=\langle f,v_h\rangle\end{cases}$$

is well-posed and $(\boldsymbol{A}_h^{-1})_h$ is uniformly bounded if and only if $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive. In that case, denoting by $C_{\mathcal{A}}$ the continuity constant of the bilinear form $\mathcal{A}$, it holds that

$$\|u-u_h\|_W\leq C\inf_{v_h\in W_h}\|u-v_h\|_W, \tag{2.87}$$

with $C=1+\frac{C_{\mathcal{A}}\beta^*}{\alpha^*}$ independent of h.

Remark 2.32. Proposition 2.31 can be extended to the case where the discrete forms $\mathcal{A}_h$ and f_h differs from the continuous forms $\mathcal{A}$ and f. In that case, Céa's lemma (2.87) becomes

$$\|u-u_h\|_W\leq C\inf_{v_h\in W_h}\big(\|u-v_h\|_W+\mathrm{Cons}_{f,h}+\mathrm{Cons}_{\mathcal{A},h}(v_h)\big),$$

with

$$\mathrm{Cons}_{f,h}=\sup_{v_h\in W_h\setminus\{0\}}\frac{|\langle f-f_h,v_h\rangle|}{\|v_h\|_W}\quad\text{and}\quad \mathrm{Cons}_{\mathcal{A},h}(v_h)=\sup_{w_h\in W_h\setminus\{0\}}\frac{|(\mathcal{A}-\mathcal{A}_h)(v_h,w_h)|}{\|w_h\|_W},\ \forall v_h\in W_h.$$

As before, we start with the leading example of Stokes problem.

2.6.1 Stokes problem

For a given h, the natural discretization of Problem (2.7) reads:

$$\begin{cases} \text{Find } (\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h \quad \text{such that} \\ \forall (\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h, \quad \mathcal{A}\big((\boldsymbol{u}_h, p_h), (\boldsymbol{v}_h, q_h)\big) = \langle \boldsymbol{f}, \boldsymbol{v}_h \rangle, \end{cases} \tag{2.88}$$

where $\boldsymbol{V}_h \subset \boldsymbol{H}^1_0(\Omega)$ and $Q_h \subset \mathrm{L}^2_0(\Omega)$ are two finite dimensional spaces constituting a *conforming* approximation of $\boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega)$.

From Proposition 2.31, we know that Problem (2.88) is well-posed if and only if $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive. To build a suitable mapping $\mathtt{T}_h \in \mathcal{L}(\boldsymbol{V}_h \times Q_h)$, a natural idea is to reproduce the continuous mapping from the proof of Theorem 2.3

$$\begin{aligned} \mathtt{T} : \boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega) &\longrightarrow \boldsymbol{H}^1_0(\Omega) \times \mathrm{L}^2_0(\Omega) \\ (\boldsymbol{u}, p) &\longmapsto \big(\lambda \boldsymbol{u} + \boldsymbol{v}_p, -\lambda p\big) \end{aligned}$$

at the discrete level. The operator $\mathtt{T}$ above depends on the divergence lifting $\boldsymbol{v}_p \in \boldsymbol{H}^1_0(\Omega)$ of the pressure $p \in \mathrm{L}^2_0(\Omega)$ defined by, see (2.5)-(2.6),

$$-\operatorname{div} \boldsymbol{v}_p = p \quad \text{and} \quad \|\nabla \boldsymbol{v}_p\| \leq C_{\mathrm{div}} \|p\|.$$

To obtain a similar lifting in the discrete setting, we consider the continuous lifting of the discrete pressure $p_h \in Q_h \subset \mathrm{L}^2_0(\Omega)$, namely $\boldsymbol{v}_{p_h} \in \boldsymbol{H}^1_0(\Omega)$ such that

$$-\operatorname{div} \boldsymbol{v}_{p_h} = p_h \quad \text{and} \quad \|\nabla \boldsymbol{v}_{p_h}\| \leq C_{\mathrm{div}} \|p_h\|. \tag{2.89}$$

This lifting $\boldsymbol{v}_{p_h}$ does not necessarily belong to the discrete space $\boldsymbol{V}_h \subset \boldsymbol{H}^1_0(\Omega)$, so we need an operator $\Pi_h : \boldsymbol{H}^1_0(\Omega) \longrightarrow \boldsymbol{V}_h$ to project it on $\boldsymbol{V}_h$. Therefore, we consider a discrete mapping of the form

$$\begin{aligned} \mathtt{T}_h : \boldsymbol{V}_h \times Q_h &\longrightarrow \boldsymbol{V}_h \times Q_h \\ (\boldsymbol{u}_h, p_h) &\longmapsto \big(\lambda \boldsymbol{u}_h + \Pi_h(\boldsymbol{v}_{p_h}), -\lambda p_h\big). \end{aligned} \tag{2.90}$$

Now, let us precise under which conditions the bilinear form $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive by mimicking the proof of Theorem 2.3. We compute

$$\mathcal{A}\big((\boldsymbol{u}_h, p_h), \mathtt{T}_h(\boldsymbol{u}_h, p_h)\big) = \nu\lambda \|\nabla \boldsymbol{u}_h\|^2 + \nu \int_\Omega \nabla \boldsymbol{u}_h : \nabla\big(\Pi_h(\boldsymbol{v}_{p_h})\big)\, \mathrm{d}x - \int_\Omega p_h \operatorname{div}\left(\Pi_h \boldsymbol{v}_{p_h}\right) \mathrm{d}x.$$

In order to get a term of the form $\|p_h\|^2$, we assume that

$$\int_\Omega p_h \operatorname{div}\left(\Pi_h \boldsymbol{v}_{p_h}\right) \mathrm{d}x = \int_\Omega p_h \operatorname{div} \boldsymbol{v}_{p_h}\, \mathrm{d}x, \tag{2.91}$$

so that

$$\mathcal{A}\big((\boldsymbol{u}_h, p_h), \mathtt{T}_h(\boldsymbol{u}_h, p_h)\big) = \nu\lambda \|\nabla \boldsymbol{u}_h\|^2 + \nu \int_\Omega \nabla \boldsymbol{u}_h : \nabla\big(\Pi_h(\boldsymbol{v}_{p_h})\big)\, \mathrm{d}x + \|p_h\|^2$$

in view of (2.89). Then, for any $\eta > 0$, Young inequality yields

$$\begin{aligned} \int_\Omega \nabla \boldsymbol{u}_h : \nabla\big(\Pi_h(\boldsymbol{v}_{p_h})\big)\, \mathrm{d}x &\geq -\frac{\eta}{2} \|\nabla \boldsymbol{u}_h\|^2 - \frac{1}{2\eta} \|\nabla\big(\Pi_h(\boldsymbol{v}_{p_h})\big)\|^2 \\ &\geq -\frac{\eta}{2} \|\nabla \boldsymbol{u}_h\|^2 - \frac{C^2_{\mathrm{div}} C^2_\pi}{2\eta} \|p_h\|^2 \end{aligned}$$

provided that there exists a constant $C_\pi > 0$, independent of h and of p_h, such that

$$\|\nabla\big(\Pi_h(\boldsymbol{v}_{p_h})\big)\| \leq C_\pi \|\nabla \boldsymbol{v}_{p_h}\|. \tag{2.92}$$

Hence, it holds that

$$\mathcal{A}\big((\boldsymbol{u}_h,p_h),\mathtt{T}_h(\boldsymbol{u}_h,p_h)\big)\geq\nu\left(\lambda-\frac{\eta}{2}\right)\|\nabla\boldsymbol{u}_h\|^2+\left(1-\frac{\nu C_{\mathrm{div}}^2C_\pi^2}{2\eta}\right)\|p_h\|^2.$$

Setting $\eta=\lambda=\nu C_{\mathrm{div}}^2C_\pi^2$, we obtain

$$\begin{aligned}\mathcal{A}\big((\boldsymbol{u}_h,p_h),\mathtt{T}_h(\boldsymbol{v}_h,p_h)\big)&\geq\frac{\nu^2C_{\mathrm{div}}^2C_\pi^2}{2}\|\nabla u_h\|^2+\frac{1}{2}\|p_h\|^2\\&\geq\frac{1}{2}\min(\nu^2C_{\mathrm{div}}^2C_\pi^2,1)\|(u_h,p_h)\|^2_{\boldsymbol{H}_0^1(\Omega)\times\mathrm{L}_0^2(\Omega)}.\end{aligned}\tag{2.93}$$

Moreover, taking into account (2.92) and mimicking the continuous case (see (2.11)), we have

$$|||\mathtt{T}_h|||\leq\max\Big(\sqrt{2}\nu C_{\mathrm{div}}^2C_\pi^2,C_{\mathrm{div}}C_\pi\big(2+\nu^2C_{\mathrm{div}}^2C_\pi^2\big)^{1/2}\Big).\tag{2.94}$$

So, with the help of the operator $\Pi_h:\boldsymbol{H}_0^1(\Omega)\longrightarrow\boldsymbol{V}_h$, we have proven the following result.

Theorem 2.33. *If there exist a family of operators $(\Pi_h)_h$ and a constant $C_\pi>0$ such that, for all h,*

$$\int_\Omega q_h\,\mathrm{div}\,(\Pi_h\boldsymbol{v})\,\mathrm{d}x=\int_\Omega q_h\,\mathrm{div}\,\boldsymbol{v}\,\mathrm{d}x,\quad\forall\boldsymbol{v}\in\boldsymbol{H}_0^1(\Omega),\forall q_h\in Q_h,\tag{2.95}$$

$$\|\nabla\big(\Pi_h(\boldsymbol{v})\big)\|\leq C_\pi\|\nabla\boldsymbol{v}\|,\quad\forall\boldsymbol{v}\in\boldsymbol{H}_0^1(\Omega),\tag{2.96}$$

then Problem (2.88) is well-posed for all h and

$$\|(u-u_h,p-p_h)\|_{\boldsymbol{H}_0^1(\Omega)\times\mathrm{L}_0^2(\Omega)}\leq C\inf_{(v_h,q_h)\in\boldsymbol{V}_h\times Q_h}\|(u-v_h,p-q_h)\|_{\boldsymbol{H}_0^1(\Omega)\times\mathrm{L}_0^2(\Omega)},\tag{2.97}$$

with

$$C=1+\frac{2\max(\nu,2)\max\Big(\sqrt{2}\nu C_{\mathrm{div}}^2C_\pi^2,C_{\mathrm{div}}C_\pi\big(2+\nu^2C_{\mathrm{div}}^2C_\pi^2\big)^{1/2}\Big)}{\min(\nu^2C_{\mathrm{div}}^2C_\pi^2,1)}.$$

Proof. The previous reasoning shows that the bilinear form $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive for the mapping

$$\begin{aligned}\mathtt{T}_h:\boldsymbol{V}_h\times Q_h&\longrightarrow\boldsymbol{V}_h\times Q_h\\(\boldsymbol{u}_h,p_h)&\longmapsto\big(\nu C_{\mathrm{div}}^2C_\pi^2\boldsymbol{u}_h+\Pi_h(\boldsymbol{v}_{p_h}),-\nu C_{\mathrm{div}}^2C_\pi^2p_h\big)\end{aligned}$$

as long as the two conditions (2.91) and (2.92) are fulfilled for all $p_h\in Q_h$, which is the case if (2.95) and (2.96) hold true. The stability estimate (2.97) then follows by using (2.93) and (2.94) in (2.87). □

The conditions (2.95) and (2.96) correspond exactly to the assumptions of an abstract result known as Fortin's lemma [Fortin, 1977]. Above, the T-coercivity approach allowed us to recover these two conditions in a somewhat direct way. Moreover, we recall that, since the form b fulfills an inf-sup condition (2.3), those conditions (2.95)-(2.96) are equivalent to the so-called *uniform discrete inf-sup condition on the form b*

$$\exists\underline{\beta}'>0,\quad\forall h,\quad\inf_{q_h\in Q_h\setminus\{0\}}\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h\setminus\{0\}}\frac{\int_\Omega q_h\,\mathrm{div}\,\boldsymbol{v}_h\,\mathrm{d}x}{\|\nabla\boldsymbol{v}_h\|\|q_h\|}\geq\underline{\beta}',$$

see for instance [Girault and Raviart, 1986, Lemma II.1.1].
Finally, we recall that, provided there is a basic approximability property (*i.e.* any element of $\boldsymbol{V}\times Q$ can be approximated by a sequence of elements of $(\boldsymbol{V}_h\times Q_h)_h$), the convergence of the discrete solutions to the exact one is a consequence of (2.97).

2.6.2 Approximation of saddle-point problems

We now derive a *conforming* approximation of the abstract problem (2.15), starting from the variational expressions (2.16) or (2.17), the latter with the form

$$\mathcal{A}\big((u,p),(v,q)\big) = a(u,v) + b(v,p) + b(u,q).$$

So, let $(V_h)_h$, resp. $(Q_h)_h$, be two families of finite dimensional subspaces of V, resp. Q. Starting from (2.16), the discrete variational formulation writes

$$\begin{cases} \text{Find } (u_h,p_h) \in V_h \times Q_h \quad \text{such that} \\ \forall v_h \in V_h, \quad a(u_h,v_h) + b(v_h,p_h) = \langle f, v_h\rangle_{V',V} \\ \forall q_h \in Q_h, \quad b(u_h,q_h) = \langle g, q_h\rangle_{Q',Q}. \end{cases}$$

while, starting from (2.17), the *all-in-one* discrete variational formulation writes

$$\begin{cases} \text{Find } (u_h,p_h) \in V_h \times Q_h \quad \text{such that} \\ \forall (v_h,q_h) \in V_h \times Q_h, \quad \mathcal{A}\big((u_h,p_h),(v_h,q_h)\big) = \langle f, v_h\rangle_{V',V} + \langle g, q_h\rangle_{Q',Q}. \end{cases}$$

In abstract form, the *uniform discrete inf-sup condition on the form b* writes

$$\exists \underline{\beta}' > 0, \quad \forall h, \quad \inf_{q_h \in Q_h\setminus\{0\}} \sup_{v_h \in V_h\setminus\{0\}} \frac{b(v_h,q_h)}{\|v_h\|_V \|q_h\|_Q} \geq \underline{\beta}'. \tag{2.98}$$

We suppose that the discrete version of the operator B is the restriction of B to V_h, namely

$$B(V_h) \subset Q_h'. \tag{2.99}$$

Introducing the discrete operators $\mathtt{B}_h : V_h \to Q_h$ such that for all h,

$$(\mathtt{B}_h v_h, q_h)_Q = b(v_h,q_h), \quad \forall (v_h,q_h) \in V_h \times Q_h,$$

the straightforward discrete counterpart of Theorem 2.6 is

Theorem 2.34. *The following three statements are equivalent:*

(i) There exists $\underline{\beta}' > 0$ such that the form b fulfills the uniform discrete inf-sup condition (2.98).

(ii) For all h, $\mathtt{B}_h : (\operatorname{Ker}\mathtt{B}_h)^\perp \to Q_h$ is an isomorphism*, and*

$$\|\mathtt{B}_h v_h\|_Q \geq \underline{\beta}' \|v_h\|_V, \quad \forall v_h \in (\operatorname{Ker}\mathtt{B}_h)^\perp. \tag{2.100}$$

(iii) For all h, there exists an isomorphic operator $\mathtt{L}_{B,h} : Q_h \to (\operatorname{Ker}\mathtt{B}_h)^\perp$ such that

$$\mathtt{B}_h(\mathtt{L}_{B,h} q_h) = q_h \quad \text{and} \quad \|q_h\|_Q \geq \underline{\beta}' \|\mathtt{L}_{B,h} q_h\|_V, \quad \forall q_h \in Q_h. \tag{2.101}$$

Remark 2.35. Obviously, this result also holds if the value of the constant in the discrete inf-sup condition on the form b depends on h, *i.e.* for each h it holds for some $\underline{\beta}'(h) > 0$, with $\lim_{h\to 0} \underline{\beta}'(h) = 0$. In this case however, getting error estimates can be more intricate.

As mentioned above for the Stokes system, one has the Fortin's Lemma (cf. [Girault and Raviart, 1986, Lemma II.1.1]).

Theorem 2.36. *Assume that the form b fulfills an inf-sup condition (2.18). The uniform discrete inf-sup condition on the form b (2.98) holds if, and only if, there exist a family of operators $(\Pi_h)_h$, with $\Pi_h : V \longrightarrow V_h$, and a constant $C_\pi > 0$ such that, for all h,*

$$\begin{aligned} &b(\Pi_h v, q_h) = b(v,q_h), \quad \forall v \in V, \ \forall q_h \in Q_h, \\ &\sup_h |||\Pi_h||| \leq C_\pi. \end{aligned} \tag{2.102}$$

Operators $(\Pi_h)_h$ that fulfill conditions (2.102) are called Fortin operators.

Let us now proceed with the derivation of conditions to ensure that the form $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive. As a general rule, the proofs of the results follow very closely the proofs that were given in the exact case. The straightforwardness of the procedure when going from the continuous to the discrete level is one of the main features of the T-coercivity approach. We give next the discrete counterparts of Theorems 2.7 and 2.10.

Theorem 2.37. *Assume that the form a is symmetric and positive, that there exists a constant $\alpha' > 0$ such that*

$$a(u_h, u_h) \geq \alpha' \|u_h\|_V^2, \quad \forall u_h \in V_h, \tag{2.103}$$

and that the uniform discrete inf-sup condition (2.98) on the form b holds true.
Then the form $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive.

The property (2.103) is sometimes called the the uniform discrete coercivity property.

Proof. Let h be given. We introduce the mapping

$$\begin{aligned} \mathtt{T}_h : V_h \times Q_h &\longrightarrow V_h \times Q_h \\ (u_h, p_h) &\longmapsto \big(\lambda u_h + \mathtt{L}_{B,h} p_h, -\lambda p_h\big). \end{aligned}$$

We then compute

$$\begin{aligned} \mathcal{A}\big((u_h, p_h), \mathtt{T}_h(u_h, p_h)\big) &= a(u_h, \lambda u_h) + a(u_h, \mathtt{L}_{B,h} p_h) + b(\lambda u_h, p_h) + b(\mathtt{L}_{B,h} p_h, p_h) - b(u_h, \lambda p_h) \\ &= \lambda a(u_h, u_h) + a(u_h, \mathtt{L}_{B,h} p_h) + \|p_h\|_Q^2, \text{ according to (2.101)-left.} \end{aligned}$$

Because the form a is symmetric and positive, we can apply Young's inequality: for any $\eta > 0$,

$$a(u_h, \mathtt{L}_{B,h} p_h) \geq -\frac{\eta}{2} a(u_h, u_h) - \frac{1}{2\eta} a(\mathtt{L}_{B,h} p_h, \mathtt{L}_{B,h} p_h).$$

According now to (2.101)-right, we find

$$a(\mathtt{L}_{B,h} p_h, \mathtt{L}_{B,h} p_h) \leq C_a \|\mathtt{L}_{B,h} p_h\|_V^2 \leq C_a (\underline{\beta}')^{-2} \|p_h\|_Q^2.$$

Using assumption (2.103), if $\lambda - \frac{\eta}{2} > 0$, it follows that

$$\mathcal{A}\big((u_h, p_h), \mathtt{T}_h(u_h, p_h)\big) \geq \alpha' \left(\lambda - \frac{\eta}{2}\right) \|u_h\|_V^2 + \left(1 - \frac{C_a (\underline{\beta}')^{-2}}{2\eta}\right) \|p_h\|_Q^2.$$

Setting $\eta = \lambda = C_a (\underline{\beta}')^{-2}$ as in the exact case, we infer that

$$\mathcal{A}\big((u_h, p_h), \mathtt{T}_h(u_h, p_h)\big) \geq \frac{1}{2} \min(\alpha' C_a (\underline{\beta}')^{-2}, 1) \|(u_h, p_h)\|_{V \times Q}^2$$

which proves that $\mathcal{A}$ is $\mathtt{T}_h$-coercive, with a T-coercivity constant $\frac{1}{2} \min(\alpha' C_a (\underline{\beta}')^{-2}, 1) > 0$ that is independent of h.

Since $\mathtt{T}_h(u_h, p_h) = (C_a (\underline{\beta}')^{-2} u_h + \mathtt{L}_{B,h} p_h, -C_a (\underline{\beta}')^{-2} p_h)$, one finds that

$$\begin{aligned} \|\mathtt{T}_h(u_h, p_h)\|_{V \times Q}^2 &\leq 2 (C_a (\underline{\beta}')^{-2})^2 \|u_h\|_V^2 + 2 \|\mathtt{L}_{B,h} p_h\|_V^2 + (C_a (\underline{\beta}')^{-2})^2 \|p_h\|_Q^2 \\ &\leq 2 (C_a (\underline{\beta}')^{-2})^2 \|u_h\|_V^2 + \big(2 (\underline{\beta}')^{-2} + (C_a (\underline{\beta}')^{-2})^2\big) \|p_h\|_Q^2, \end{aligned}$$

where the last inequality follows from (2.101)-right. The bound is valid for all h, which yields

$$\sup_h |||\mathtt{T}_h||| \leq \max\Big(\sqrt{2} C_a (\underline{\beta}')^{-2}, \beta\big(2 + C_a^2 (\underline{\beta}')^{-2}\big)^{1/2}\Big),$$

so the form $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive. □

Remark 2.38. As for the Stokes problem, the discrete right-inverse $\mathtt{L}_{B,h}$ is connected to the Fortin operator Π_h. As a matter of fact, if there exists a family of discrete projectors $(\Pi_h)_h$ verifying (2.102), the operator defined by $\mathtt{L}_{B,h} = \Pi_h(\mathtt{L}_B)$ satisfies (2.101) with $\underline{\beta}' = (C_\pi\beta)^{-1}$ since for all $q_h \in Q_h$

$$\|\Pi_h(\mathtt{L}_B q_h)\|_V \leq C_\pi \|\mathtt{L}_B q_h\|_V \leq C_\pi\beta\|q_h\|_Q,$$

according to (2.22). As a consequence, to perform stability estimates at the discrete level using $\mathtt{T}_h$-coercivity, one has only to replace β by $C_\pi\beta$ in the computations done at the continuous level.

Theorem 2.39. *Assume that the form a is symmetric and positive, that there exists a constant $\alpha_0' > 0$ such that*

$$a(u_{0,h}, u_{0,h}) \geq \alpha_0' \|u_{0,h}\|_V^2, \quad \forall u_{0,h} \in \operatorname{Ker} \mathtt{B}_h, \tag{2.104}$$

and that the uniform discrete inf-sup condition (2.98) on the form b holds true.
Then the form $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive.

The property (2.104) is sometimes called the uniform discrete coercivity property on the kernels.

Proof. Let h be given. We consider the mapping

$$\begin{aligned} \mathtt{T}_h : V_h \times Q_h &\longrightarrow V_h \times Q_h \\ (u_h, p_h) &\longmapsto \big(\lambda u_h + \mathtt{L}_{B,h} p_h, -\lambda p_h + \lambda\mu\, \mathtt{B}_h u_h\big). \end{aligned}$$

As in the proof of Theorem 2.10, we can compute

$$\mathcal{A}\big((u_h,p_h), \mathtt{T}_h(u_h,p_h)\big) = \lambda a(u_h,u_h) + a(u_h, \mathtt{L}_{B,h}p_h) + \|p_h\|_Q^2 + \lambda\mu\|\mathtt{B}_h u_h\|_Q^2$$

because $b(\mathtt{L}_{B,h}p_h, p_h) = \|p_h\|_Q^2$. Since the form a is symmetric and positive, one may use Young's inequality. By proceeding as in the proof of Theorem 2.37 and after setting $\lambda = C_a(\underline{\beta}')^{-2}$, we find that

$$\lambda a(u_h,u_h) + a(u_h, \mathtt{L}_{B,h}, p_h) + \|p_h\|_Q^2 \geq \frac{1}{2} C_a(\underline{\beta}')^{-2} a(u_h,u_h) + \frac{1}{2}\|p_h\|_Q^2,$$

and

$$\mathcal{A}\big((u_h,p_h), \mathtt{T}_h(u_h,p_h)\big) \geq \frac{1}{2} C_a(\underline{\beta}')^{-2}\big(a(u_h,u_h) + 2\mu\|\mathtt{B}_h u_h\|_Q^2\big) + \frac{1}{2}\|p_h\|_Q^2.$$

Then, we use the decomposition $u_h = u_{0,h} + \bar{u}_h$ with $u_{0,h} \in \operatorname{Ker}\mathtt{B}_h$ and $\bar{u}_h \in (\operatorname{Ker}\mathtt{B}_h)^\perp$. As before, Young's inequality yields

$$a(u_h,u_h) \geq (1-\theta)a(u_{0,h},u_{0,h}) + \left(C_a - \frac{C_a}{\theta}\right)\|\bar{u}_h\|_V^2$$

for all $0 < \theta < 1$. Moreover, $\|\mathtt{B}_h u_h\|_Q^2 = \|\mathtt{B}_h\bar{u}_h\|_Q^2 \geq (\underline{\beta}')^2\|\bar{u}_h\|_V^2$ according to (2.100), so that

$$a(u_h,u_h) + 2\mu\|\mathtt{B}_h u_h\|_Q^2 \geq (1-\theta)a(u_{0,h},u_{0,h}) + \left(C_a - \frac{C_a}{\theta} + 2\mu(\underline{\beta}')^2\right)\|\bar{u}_h\|_V^2.$$

Choosing $\theta = \frac{1}{2}$ and $\mu = \frac{3}{4}C_a(\underline{\beta}')^{-2}$, it holds that

$$\begin{aligned} a(u_h,u_h) + 2\mu\|\mathtt{B}_h u_h\|_Q^2 &\geq \frac{1}{2}a(u_{0,h},u_{0,h}) + \frac{C_a}{2}\|\bar{u}_h\|_V^2 \\ &\geq \frac{\alpha_0'}{2}\|u_{0,h}\|_V^2 + \frac{\alpha_0'}{2}\|\bar{u}_h\|_V^2 = \frac{\alpha_0'}{2}\|u_h\|_V^2, \end{aligned}$$

where we used assumption (2.104) and $C_a \geq \alpha_0'$ on the second line.

Finally, we conclude that

$$\mathcal{A}\big((u_h,p_h),\mathtt{T}_h(u_h,p_h)\big) \geq \frac{1}{4}\alpha_0' C_a(\underline{\beta}')^{-2}\|u_h\|_V^2 + \frac{1}{2}\|p_h\|_Q^2,$$

which yields that $\mathcal{A}$ is $\mathtt{T}_h$-coercive, with a T-coercivity constant $\min(\frac{1}{4}\alpha_0' C_a(\underline{\beta}')^{-2}, \frac{1}{2}) > 0$ that is independent of h.

From the above, we have $\mathtt{T}_h(u_h,p_h) = \big(C_a(\underline{\beta}')^{-2}u_h + \mathtt{L}_{B,h}p_h, -C_a(\underline{\beta}')^{-2}p_h + \frac{3}{4}(C_a(\underline{\beta}')^{-2})^2\mathtt{B}_h u_h\big)$, and, noting that $\|\mathtt{B}_h u_h\|_Q \leq C_b\|u_h\|_V$, one concludes that

$$\sup_h |||\mathtt{T}_h||| \leq \infty,$$

so the form $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive. □

Remark 2.40. The reciprocal of Theorem 2.39 is also true. To see this, one simply needs to mimick the proof of Theorem 2.10 - item 2. at the discrete level.

Remark 2.41. Note that replacing $\mathtt{B}u$ by $\mathtt{B}_h u_h$ when going from the continuous operator $\mathtt{T}$ to the discrete operator $\mathtt{T}_h$ is possible because we assumed that B_h is the restriction of B to V_h, see (2.99). If (2.99) does not hold, one introduces $\Phi_h : Q \longrightarrow Q_h$ defined by

$$b(v_h,\Phi_h q) = b(v_h,q), \quad \forall q \in Q, \forall v_h \in V_h,$$

like in [Boffi et al., 2013, Proposition 5.1.2]. Then, one has to replace $\mathtt{B}_h u_h$ by $\Phi_h(\mathtt{B}_h u_h)$ in the previous proof, *i.e.*

$$\begin{aligned}\mathtt{T}_h : V_h \times Q_h &\longrightarrow V_h \times Q_h\\ (u_h,p_h) &\longmapsto \Big(C_a(\underline{\beta}')^{-2}u_h + \mathtt{L}_{B,h}p_h, -C_a(\underline{\beta}')^{-2}p_h + \frac{3}{4}(C_a(\underline{\beta}')^{-2})^2\Phi_h(\mathtt{B}u_h)\Big).\end{aligned}$$

Again, provided a basic approximability property holds, that is, any element of $V \times Q$ can be approximated by a sequence of elements of $(V_h \times Q_h)_h$, convergence will follow under the assumptions of Theorem 2.37 or Theorem 2.39.

2.6.3 Approximation of augmented saddle-point problems

We now approximate the abstract problem (2.36), starting from the variational expression (2.37), with the form

$$\mathcal{A}_c\big((u,p),(v,q)\big) = a(u,v) + b(v,p) + b(u,q) - c(p,q),$$

where $c(\cdot,\cdot)$ is a form defined on $Q \times Q$ that fulfills (2.35); in particular, $c(\cdot,\cdot)$ is *positive*. So, let again $(V_h)_h$, resp. $(Q_h)_h$, be two families of finite dimensional subspaces of V, resp. Q. The discrete variational formulation writes

$$\begin{cases}\text{Find } (u_h,p_h) \in V_h \times Q_h \quad \text{such that}\\ \forall (v_h,q_h) \in V_h \times Q_h, \quad \mathcal{A}_c\big((u_h,p_h),(v_h,q_h)\big) = \langle f, v_h\rangle_{V',V} + \langle g, q_h\rangle_{Q',Q},\end{cases}$$

To ensure that the form $\mathcal{A}_c$ is uniformly $\mathtt{T}_h$-coercive, the proofs once more follow very closely those that were given in the exact case. We give next the discrete counterparts of Theorems 2.12, 2.14 and 2.20.

Theorem 2.42. *Assume that the form a is symmetric, positive, fulfills the uniform discrete coercivity property (2.103), and that the uniform discrete inf-sup condition (2.98) on the form b holds true. Then the form $\mathcal{A}_c$ is uniformly $\mathtt{T}_h$-coercive.*

Theorem 2.43. *Assume that the form a is symmetric and positive, fulfills the uniform discrete coercivity property on the kernels (2.104), and that the uniform discrete inf-sup condition (2.98) on the form b holds true. If moreover the form c is like in (2.40), where ε is* small enough, *namely*

$$\varepsilon \leq \frac{1}{2C_a(C_\pi\beta)^4C_b^2}\Big(2-\frac{\alpha_0}{C_a}\Big),$$

then the form $\mathcal{A}_c$ is uniformly $\mathtt{T}_h$-coercive.

Theorem 2.44. *Assume that (2.52) holds true and that the bilinear forms a and c are both symmetric and positive. If there exists a constant $\alpha'_B > 0$ such that*

$$a(u_h,u_h)+\frac{\gamma}{2C_c^2}\|\mathtt{B}_h u_h\|_Q^2 \geq \alpha'_B\|u_h\|_V^2, \quad \forall u_h \in V_h, \tag{2.105}$$

then the form $\mathcal{A}_c$ is uniformly $\mathtt{T}_h$-coercive.

As before, provided a basic approximability property holds, that is, any element of $V \times Q$ can be approximated by a sequence of elements of $(V_h \times Q_h)_h$, convergence will follow under the assumptions of Theorem 2.42, Theorem 2.43 or Theorem 2.44.

2.6.4 Applications

Let us briefly see how the T-coercivity approach can be used to discretize the mixed problems, that is for Stokes, electromagnetism, nearly-incompressible elasticity and finally neutron diffusion. For each problem, we propose one or several possibilities. Note that, since there is a vast litterature on this topic, there is no need to devise new approximation techniques. On the contrary, the simple framework of the T-coercivity approach provides elementary guidelines to help us choose among existing techniques. In most cases, we emphasize that this leads to explicit discrete operators $\mathtt{T}_h$. And, as we shall see next, the degree of explicitness depends on the problem that is studied.

In each case, the first step is to choose a *conforming* finite element discretization adapted to the space V under consideration. We assume for simplicity that Ω is a polyhedron for $d = 3$, or a polygon for $d = 2$, so one can use meshes made of simplices for the discretization by finite elements. For $k \geq 1$, $\mathcal{P}_k$ stands for the Lagrange finite elements of order k. For Stokes and elasticity, we note that the space $\boldsymbol{H}_0^1(\Omega)$ may be approximated using $(\mathcal{P}_k)^d$ finite elements with $k \geq 2$. For electromagnetism, we have to deal with the space $\boldsymbol{H}_0(\mathbf{curl}\,;\Omega)$, which can be discretized using the (first-kind) Nédélec finite elements of order $k \geq 1$, denoted by $\mathcal{N}_k$. Lastly, for neutron diffusion, we have to deal with the space $\boldsymbol{H}(\mathrm{div}\,;\Omega)$, discretized with the help of the Raviart-Thomas elements of order $k \geq 0$, denoted by $\mathcal{RT}_k$. We refer to [Boffi et al., 2013] for details.

The next step is to choose the *conforming* finite element discretization in the space Q in such a way that convergence of the discrete solutions to the exact one is guaranteed. This occurs as soon as one achieves uniform $\mathtt{T}_h$-coercivity for the all-in-one form $\mathcal{A}$. To that aim, one simply has to build discrete operators $\mathtt{T}_h$ similarly as in the continuous case but using (when applicable) the Fortin operators defined in Theorem 2.36 to project the lifting $\mathtt{L}_\mathtt{B}$ on the discrete space. Note that according to classical theory [Girault and Raviart, 1986; Boffi et al., 2013], the existence of such operators is equivalent to the uniform discrete inf-sup condition (2.98) on the form b. Doing so, the discrete stability estimates then follow from the continuous case by changing the constant β to take into account the influence of the Fortin operators, see Remark 2.38.

First, for Stokes and elasticity, and for $k = 2$, setting $Q_h = \mathcal{P}_1$ leads to Fortin operators $\Pi_h^{\mathcal{P}} : \boldsymbol{H}_0^1(\Omega) \longrightarrow (\mathcal{P}_k)^d$ satisfying (2.95)-(2.96), or the abstract counterpart (2.102): the pair $((\mathcal{P}_2)^d, \mathcal{P}_1)$

is called the Taylor-Hood finite element. In this case, building Fortin operators is a very technical issue. We do not go into the details, and refer instead to Section 8.8 in [Boffi et al., 2013]: we note that the resulting expression of the operator $\Pi_h^{\mathcal{P}}$ is quite involved. This leads to a uniform discrete inf-sup condition on the form b. Regarding uniform $\mathtt{T}_h$-coercivity, one uses (2.90) to define the discrete operators $\mathtt{T}_h$ for Stokes. While, for nearly-incompressible elasticity, the bilinear form defined in (2.79) is uniformly $\mathtt{T}_h$-coercive for the mapping

$$\begin{aligned}\mathtt{T}_h : (\mathcal{P}_2)^d \times \mathcal{P}_1 &\longrightarrow (\mathcal{P}_2)^d \times \mathcal{P}_1\\ (\boldsymbol{u}_h, p_h) &\longmapsto \big(2\mu(C_{\pi,\mathcal{P}}C_{\mathrm{div}})^2\boldsymbol{u}_h + \Pi_h^{\mathcal{P}}(\boldsymbol{v}_{-p_h}), -2\mu(C_{\pi,\mathcal{P}}C_{\mathrm{div}})^2 p_h\big),\end{aligned}$$

according to Theorem 2.42 and (2.80).

For electromagnetism in an anisotropic medium, for $k \geq 1$, setting $Q_h = \mathcal{P}_k$ leads to Fortin operators $\Pi_h^{\mathcal{N}} : \boldsymbol{H}_0(\mathbf{curl}\,;\Omega) \longrightarrow \mathcal{N}_k$ satisfying (2.102). Let us explain below how to proceed. In (2.102), the operators must fulfill the compatibility conditions

$$(\underline{\varepsilon}\Pi_h^{\mathcal{N}}\boldsymbol{v}, \nabla q_h)_{\boldsymbol{L}^2(\Omega)} = (\underline{\varepsilon}\boldsymbol{v}, \nabla q_h)_{\boldsymbol{L}^2(\Omega)}, \quad \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega),\ \forall q_h \in \mathcal{P}_k. \tag{2.106}$$

Under the assumptions (2.57) on $\underline{\varepsilon}$ (and $\underline{\mu}$), the Helmholtz decomposition of $\boldsymbol{v}$, orthogonal with respect to the inner product $(\cdot,\cdot)_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}$, writes $\boldsymbol{v} = \boldsymbol{k}_{\boldsymbol{v}} + \nabla\phi_{\boldsymbol{v}}$, where $\boldsymbol{k}_{\boldsymbol{v}} \in \boldsymbol{K}_N(\Omega;\underline{\varepsilon})$ and $\phi_{\boldsymbol{v}} \in H_0^1(\Omega)$ solves (2.74). Since $q_h \in H_0^1(\Omega)$, we note that

$$(\underline{\varepsilon}\boldsymbol{v}, \nabla q_h)_{\boldsymbol{L}^2(\Omega)} = (\underline{\varepsilon}\nabla\phi_{\boldsymbol{v}}, \nabla q_h)_{\boldsymbol{L}^2(\Omega)}, \quad \forall \boldsymbol{v} \in \boldsymbol{H}_0(\mathbf{curl}\,;\Omega),\ \forall q_h \in \mathcal{P}_k.$$

This leads to the "natural" choice

$$\Pi_h^{\mathcal{N}}\boldsymbol{v} = \nabla(P_h^k\phi_{\boldsymbol{v}}),$$

where $P_h^k : H_0^1(\Omega) \longrightarrow \mathcal{P}_k$ is the orthogonal projection on $\mathcal{P}_k$ with respect to the inner product $(\cdot,\cdot)_{1,\underline{\varepsilon}}$, namely

$$(\underline{\varepsilon}\nabla(P_h^k\phi), \nabla q_h)_{\boldsymbol{L}^2(\Omega)} = (\underline{\varepsilon}\nabla\phi, \nabla q_h)_{\boldsymbol{L}^2(\Omega)}, \quad \forall \phi \in H_0^1(\Omega),\ \forall q_h \in \mathcal{P}_k. \tag{2.107}$$

Indeed, the above problem is well-posed thanks to the assumptions (2.57) on $\underline{\varepsilon}$, and $\Pi_h^{\mathcal{N}}\boldsymbol{v}$ automatically belongs to $\mathcal{N}_k$ because, by design, the finite element space $\mathcal{N}_k$ contains $\nabla[\mathcal{P}_k]$.
With this definition of the operator $\Pi_h^{\mathcal{N}}$, the compatibility conditions (2.106) immediately follow. Furthermore, the uniform bound on the norm of the operators in (2.102) is obtained via

$$\begin{aligned}\|\Pi_h^{\mathcal{N}}\boldsymbol{v}\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} &= \|\nabla(P_h^k\phi_{\boldsymbol{v}})\|^2_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}\\ &= (\underline{\varepsilon}\nabla(P_h^k\phi_{\boldsymbol{v}}), \nabla(P_h^k\phi_{\boldsymbol{v}}))_{\boldsymbol{L}^2(\Omega)}\\ \text{(cf. (2.107))} &= (\underline{\varepsilon}\nabla\phi_{\boldsymbol{v}}, \nabla(P_h^k\phi_{\boldsymbol{v}}))_{\boldsymbol{L}^2(\Omega)}\\ &\leq \|\nabla\phi_{\boldsymbol{v}}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}\,\|\nabla(P_h^k\phi_{\boldsymbol{v}})\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}\\ &= \|\nabla\phi_{\boldsymbol{v}}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}\,\|\Pi_h^{\mathcal{N}}\boldsymbol{v}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}},\end{aligned}$$

so that $\|\Pi_h^{\mathcal{N}}\boldsymbol{v}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}} \leq \|\boldsymbol{v}\|_{\underline{\varepsilon},\underline{\mu}^{-1}\mathbf{curl}}$ by orthogonality of the Helmholtz decomposition.
On the other hand, the uniform coercivity condition on the discrete kernels in Theorem 2.39 is obtained for instance in [Ciarlet Jr, 2020, Theorem 3]. It hinges on the crucial property stating that, given a field $\boldsymbol{v} = \nabla q$ with $q \in H_0^1(\Omega)$, the result of the interpolation of $\boldsymbol{v}$ with the Nédélec interpolation operator may be expressed as ∇q_h for some $q_h \in \mathcal{P}_k$.
Hence, for electromagnetism in an anisotropic medium we infer from Theorem 2.39 that the bilinear form $\mathcal{A}_{\underline{\varepsilon},\underline{\mu}}$ defined in (2.69) is uniformly $\mathtt{T}_h$-coercive for the mapping

$$\begin{aligned}\mathtt{T}_h : \mathcal{N}_k \times \mathcal{P}_k &\longrightarrow \mathcal{N}_k \times \mathcal{P}_k\\ (\boldsymbol{E}_h, \tilde{p}_h) &\longmapsto \left((C_{\pi,\mathcal{N}})^2\boldsymbol{E}_h + \Pi_h^{\mathcal{N}}(\nabla\tilde{p}_h), -(C_{\pi,\mathcal{N}})^2\tilde{p}_h + \frac{3}{4}(C_{\pi,\mathcal{N}})^4\phi_{\boldsymbol{E}_h}\right),\end{aligned}$$

where $\phi_{\boldsymbol{E}_h} \in Q_h$ satisfies the discrete counterpart of (2.74), namely

$$(\underline{\varepsilon}\nabla\phi_{\boldsymbol{E}_h}, \nabla q_h)_{\boldsymbol{L}^2(\Omega)} = (\underline{\varepsilon}\boldsymbol{E}_h, \nabla q_h)_{\boldsymbol{L}^2(\Omega)}, \quad \forall q_h \in \mathcal{P}_k.$$

Alternatively, one can follow Section 2.3.2. In this case, adding indices and superscripts h in the definition of $\mathtt{T}_{opt}$, one considers

$$\begin{aligned}(\mathtt{T}_{opt})_h : \mathcal{N}_k \times \mathcal{P}_k &\longrightarrow \mathcal{N}_k \times \mathcal{P}_k \\ (\boldsymbol{E}_h, \tilde{p}_h) &\longmapsto \big(\boldsymbol{k}^h_{\boldsymbol{E}_h} + \nabla\tilde{p}_h, \phi^h_{\boldsymbol{E}_h}\big),\end{aligned}$$

with the decomposition $\boldsymbol{E}_h = \boldsymbol{k}^h_{\boldsymbol{E}_h} + \nabla\phi^h_{\boldsymbol{E}_h}$. The crucial idea is now to use a discrete Helmholtz decomposition to define $\boldsymbol{k}^h_{\boldsymbol{E}_h}$ and $\phi^h_{\boldsymbol{E}_h}$. Namely, for $\boldsymbol{v}_h \in \mathcal{N}_k$, $\phi^h_{\boldsymbol{v}_h} \in \mathcal{P}_k$ is defined by

$$(\underline{\varepsilon}\nabla\phi^h_{\boldsymbol{v}_h}, \nabla q_h)_{\boldsymbol{L}^2(\Omega)} = (\underline{\varepsilon}\boldsymbol{v}_h, \nabla q_h)_{\boldsymbol{L}^2(\Omega)}, \quad \forall q_h \in \mathcal{P}_k,$$

and $\boldsymbol{k}^h_{\boldsymbol{v}_h} = \boldsymbol{v}_h - \nabla\phi^h_{\boldsymbol{v}_h}$. Doing so, one obtains an (orthogonal) discrete Helmholtz decomposition that is uniformly stable, and it can be checked that the form $\mathcal{A}_{\underline{\varepsilon},\underline{\mu}}$ is uniformly $\mathtt{T}_h$-coercive. Details can be found in [Ciarlet Jr, 2020, Proposition 13].

Last, for neutron diffusion and for $k \geq 1$, one only needs to select Q_h in such a way that (2.99) is fulfilled, namely $\operatorname{div}(\mathcal{RT}_k) \subset Q_h'$. To do so, we set $Q_h = \mathcal{P}_k^{pw}$, for $k \geq 0$, where the superscript ^{pw} stands for piecewise Lagrange finite elements of order k. Assuming for simplicity that σ restricted to any simplex is constant, we introduce the discrete mapping [Jamelot and Ciarlet Jr, 2013; Ciarlet Jr et al., 2017]

$$\begin{aligned}\mathtt{T}_h : \mathcal{RT}_k \times \mathcal{P}_k^{pw} &\longrightarrow \mathcal{RT}_k \times \mathcal{P}_k^{pw} \\ (\boldsymbol{p}_h, u_h) &\longmapsto \left(\boldsymbol{p}_h, \frac{1}{2}(-u_h + \sigma^{-1}\operatorname{div}\boldsymbol{p}_h)\right),\end{aligned}$$

and the property $\operatorname{div}(\mathcal{RT}_k) \subset \mathcal{P}_k^{pw}$ guarantees the uniform $\mathtt{T}_h$-coercivity of the bilinear form (2.85) in virtue of Theorem 2.44.

All basic approximability properties are established in [Boffi et al., 2013], which guarantees convergence in all of the above cases. We again refer to [Boffi et al., 2013] for details and possible extensions, such as the generalized Taylor-Hood elements ($k \geq 3$) or the MINI element for Stokes or elasticity.

Conclusion

We have demonstrated the flexibility of the T-coercivity approach, here applied to classical linear mixed problems, both for the theoretical study of the problems and for their numerical approximation by finite elements. Let us mention some possible extensions, such as nonconforming discretization methods for Stokes [Jamelot, 2023], multigroup diffusion [Giret, 2018] or DDM for diffusion [Ciarlet Jr et al., 2017]. It is our belief that numerous applications can be studied with the T-coercivity approach, both theoretically and numerically. Recent works include application in poromechanics [Barré et al., 2023], time-harmonic Maxwell's equations with impedance surfaces [Levadoux, 2022], and the applications listed in [Hong et al., 2023].

CHAPTER 3

Numerical analysis of an incompressible soft material poromechanics model using T-coercivity

This chapter reproduces results published in *Comptes Rendus. Mécanique*, 351(S1), 1-36 (2023) and obtained in collaboration with Céline Grandmont and Philippe Moireau. Moreover, in September 2022, I presented this work at the *GIMC-SIMAI joint Workshop for Young Scientists* in Pavia, Italy.

Contents

Numerical analysis of an incompressible soft material poromechanics model using T-coercivity

Mathieu Barré[1,2], Céline Grandmont[3,4,5] and Philippe Moireau[1,2]

[1] Inria, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
[2] LMS, École Polytechnique, CNRS, Institut Polytechnique de Paris Route de Saclay, 91120 Palaiseau, France
[3] Inria, 2 Rue Simone Iff, 75012 Paris, France
[4] LJLL, Sorbonne Université, CNRS, 4 Place Jussieu, 75005 Paris, France
[5] Département de Mathématique, Université Libre de Bruxelles CP 214, Boulevard du Triomphe, 1050 Bruxelles, Belgium

Published in *Comptes Rendus. Mécanique*
DOI:10.5802/crmeca.194

Abstract

This article is devoted to the numerical analysis of the full discretization of a generalized poromechanical model resulting from the linearization of an initial model fitted to soft tissue perfusion. Our strategy here is based on the use of energy-based estimates and T-coercivity methods, so that the numerical analysis benefits from the essential tools used in the existence analysis of the continuous-time and continuous-space formulation. In particular, our T-coercivity strategy allows us to obtain the necessary inf-sup condition for the global system from the inf-sup condition restricted to a subsystem having the same structure as the Stokes problem. This allows us to prove that any finite element pair adapted to the Stokes problem is also suitable for this global poromechanical model regardless of porosity and permeability, generalizing previous results from the literature studying this model.

Keywords — Poromechanics, mixture theory, incompressible limit, total discretization, inf-sup stability, energy preserving time-scheme.

Introduction

Poromechanical models describe the mechanical response of saturated porous media in which fluid flow interacts with a deformable structure through the definition of a multiphase continuum framework. Such models were originally developed by the geosciences community [Biot, 1941; Terzaghi, 1943; Russell and Wheeler, 1983], but have reached new application areas such as biomechanics to model perfused living tissues [Yang and Taber, 1991; Huyghe et al., 1992; Khaled and Vafai, 2003; Chapelle et al., 2010; Tully and Ventikos, 2011; Michler et al., 2013; Berger et al., 2016; Vardakis et al., 2016; Chou et al., 2016; Sacco et al., 2017; Lourenco et al., 2022]. In these biomedical applications, physical phenomena such as the fluid inertia and solid quasi-incompressibility may not be neglected, as it was the case in soil engineering, leading to more general formulations. In this spirit, [Chapelle and Moireau, 2014] has proposed a rather general formulation, valid for large strains and adapted to soft tissue perfusion. In a recent paper [Barré et al., 2023], we analyze the linearization of this model in the context of small deformations, small velocities and around a given state of perfusion. Our analysis generalizes previous existence results explored in [Burtschell et al., 2019] and

[Barnafi et al., 2021] by extending the existence to the incompressible case and in the absence of solid viscosity where we face a hyperbolic-parabolic problem under a global incompressibility constraint. In particular, the results obtained in [Barré et al., 2023] are based on the use of energy estimates and T-coercivity. This notion was originally introduced for sign-changing coefficients problems [Chesnel and Ciarlet, 2013] but we took advantage of it in our mixed hyperbolic-parabolic setting. In fact, T-coercivity was moreover recently explored for general mixed formulations in [Barré and Ciarlet Jr, 2022]. In this case, the T-coercivity approach is an alternative to the classical inf-sup condition. It allows us to elegantly combine several transformations defined from the inf-sup condition of subsystems into a general inf-sup condition for the globally coupled problem. This provides a powerful tool to integrate in a unique framework (a) the hyperbolic structure of the solid – in the absence of solid viscosity – and (b) the parabolic structure of the fluid, as well as (c) the divergence constraint on the mixture velocity, which combines the velocities of fluid and solid, without being restricted by porosity.

In this work, we propose to use T-coercivity in the context of numerical analysis by proving the convergence of space and time discretization schemes of the linearized version of the model proposed in [Chapelle and Moireau, 2014]. Again, T-coercivity provides a general framework for the study of such coupled and constrained systems and facilitates the numerical analysis. Firstly, it allows us to easily handle the hyperbolic-parabolic coupling at the discrete level when the solid has no viscosity, in which case the model rewritten in first order form is no longer associated with a coercive form. Secondly, it allows us to find a global inf-sup condition for the coupled problem directly from an inf-sup condition applied to a subsystem that is exactly the Stokes problem. Therefore, we can benefit from all the results of the numerical analysis for the Stokes problem [Pironneau and Glowinski, 1979; Glowinski, 2003; Boffi et al., 2013; Ern and Guermond, 2021a] and show that any pair of finite elements adapted to the Stokes problem provides a way to define a set of finite elements fitted to this general poromechanical model, independently of the porosity that originally appears in the divergence condition. This leads to a generalization of the convergence results obtained in [Burtschell et al., 2019] and [Barnafi et al., 2021], in particular without any restriction on the model parameters and in the incompressible limit case that was not considered in these studies. Furthermore, our analysis takes into account an additional fluid mass input entering the porous medium, which was not included in [Burtschell et al., 2019] and was assumed to be small enough in [Barnafi et al., 2021]. In the case where no restriction is imposed on the fluid mass input, we prove the stability and convergence of the proposed schemes under a smallness condition on the time step.

The paper is organized as follows. In the next section we recall the model formulation, the energy estimates, the existence results and the key properties of T-coercivity. In the third section, we present the time schemes under consideration and in Section 4, we proceed to the space-time convergence analysis. The last section is devoted to numerical illustrations.

3.1 Problem setting

3.1.1 Presentation of the model

In this work, we consider a poromechanics model describing the motion of an elastic medium filled by an incompressible viscous fluid. This model arises from the linearization of the nonlinear poromechanics model introduced in [Chapelle and Moireau, 2014] in the context of soft-tissue perfusion. The porous medium is modeled as a mixture of a solid phase and a fluid phase that cohabit and interact at each point of the domain Ω. For all $x \in \Omega \subset \mathbb{R}^d$ $(d = 2, 3)$, a porosity $0 \leq \phi(x) \leq 1$ is given, which corresponds to the fraction of fluid within the porous mixture, whereas $1 - \phi(x)$ represents the fraction of elastic medium. The macroscopic state variables are the solid displacement u_s, the fluid velocity v_f and the interstitial pressure p, namely the fluid pressure in the pores. The governing equations derived in [Burtschell et al., 2019; Barré et al., 2023] by linearizing

the model from [Chapelle and Moireau, 2014] read:

$$\begin{cases} \rho_s(1-\phi)\,\partial_{tt}^2 u_s - \operatorname{div}\big(\sigma_s(u_s)\big) - \operatorname{div}\big(\sigma_s^{\mathrm{vis}}(\partial_t u_s)\big) \\ \qquad\qquad - \phi^2\, k_f^{-1}(v_f - \partial_t u_s) + (b-\phi)\,\nabla p = \rho_s(1-\phi)\, f, & \text{in } \Omega\times(0,T), \text{(3.1a)} \\ \rho_f\phi\,\partial_t v_f - \operatorname{div}\big(\phi\,\sigma_f(v_f)\big) + \phi^2\, k_f^{-1}(v_f - \partial_t u_s) - \theta\, v_f + \phi\,\nabla p = \rho_f\phi\, f, & \text{in } \Omega\times(0,T), \text{(3.1b)} \\ \dfrac{b-\phi}{\kappa}\,\partial_t p + \operatorname{div}\big((b-\phi)\,\partial_t u_s + \phi\, v_f\big) = \rho_f^{-1}\theta, & \text{in } \Omega\times(0,T). \text{(3.1c)} \end{cases}$$

In the above system, the first equation (3.1a) is the solid mass momentum balance, the second one (3.1b) is the fluid mass momentum balance equation, and the third one (3.1c) corresponds to the mass balance equation for the global mixture incorporating both the solid and fluid phases.

The solid and fluid densities are denoted by ρ_s and ρ_f, so that $\rho_s(1-\phi)\partial_{tt}^2 u_s$ and $\rho_f\phi\,\partial_t v_f$ represent respectively the accelerations of solid and fluid particles within the mixture. We assume that the structure stress tensor $\sigma_s(u_s)$ follows Hooke's law

$$\sigma_s(u_s) = \lambda\,\mathrm{Tr}(\varepsilon(u_s))I + 2\mu\,\varepsilon(u_s),$$

where λ and μ are two Lamé constants characterizing the macroscopic behavior of the solid perforated part, and $\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^T)$ is the linearized Green-Lagrange strain tensor. Similarly, we suppose that the fluid stress tensor is given by

$$\sigma_f(v_f) = \lambda_f\,\mathrm{Tr}(\varepsilon(v_f))I + 2\mu_f\,\varepsilon(v_f),$$

and that the solid additional viscosity reads $\sigma_s^{\mathrm{vis}}(\partial_t u_s) = 2\eta\,\varepsilon(\partial_t u_s)$, with μ_f and η denoting the fluid and solid viscosities. The solid and fluid equations are coupled by a term $\phi^2 k_f^{-1}(v_f - \partial_t u_s)$ translating the friction between the two phases. This friction term is proportional to the filtration velocity $\phi(v_f - \partial_t u_s)$ through a coefficient ϕk_f^{-1}, where k_f denotes the hydraulic conductivity tensor, namely the ratio between the intrinsic permeability and the fluid viscosity. Moreover, the solid and fluid dynamics are coupled by the gradient of pressure ∇p, which is splitted into a contribution $(b-\phi)\nabla p$ in (3.1a) and $\phi\nabla p$ in (3.1b), where b is the Biot-Willis coefficient that takes into account the pressure-deformation coupling at the pore scale. The interstitial pressure dynamics is governed by the mass balance equation (3.1c) involving the solid grain bulk modulus κ, or more precisely the storage coefficient $\frac{b-\phi}{\kappa}$. Finally, in addition to the porosity and the Biot-Willis coefficient, the input data are the applied exterior body force f, distributed with a coefficient $\rho_s(1-\phi)$ among the solid and $\rho_f\phi$ among the fluid, and a volumic fluid mass source term described by a scalar function θ which is assumed to depend only on space.

As shown in [Barré et al., 2023, Section 1.1], this model can be seen as a generalization of Darcy, Brinkman and Biot equations. As a matter of fact, (3.1) includes inertial and viscous effects both for the solid and fluid phases, while most of standard poromechanics models – see for instance [Biot, 1941, 1955; Biot and Temple, 1972] – do not consider the fluid velocity as a primary state variable.

In what follows, we will focus on the case where the solid is non-viscous and incompressible, so that we may assume that $\eta = 0$, $b = 1$ and $\kappa = \infty$. The last hypothesis is motivated by the targeted physiological applications, since most of biological tissues are nearly incompressible. Under such assumptions, system (3.1) becomes: find (u_s, v_f, p) such that

$$\begin{cases} \rho_s(1-\phi)\,\partial_{tt}^2 u_s - \operatorname{div}\big(\sigma_s(u_s)\big) \\ \qquad\qquad - \phi^2\, k_f^{-1}(v_f - \partial_t u_s) + (1-\phi)\,\nabla p = \rho_s(1-\phi)\, f, & \text{in } \Omega\times(0,T), \text{(3.2a)} \\ \rho_f\phi\,\partial_t v_f - \operatorname{div}\big(\phi\,\sigma_f(v_f)\big) + \phi^2\, k_f^{-1}(v_f - \partial_t u_s) - \theta\, v_f + \phi\,\nabla p = \rho_f\phi\, f, & \text{in } \Omega\times(0,T), \text{(3.2b)} \\ \operatorname{div}\big((1-\phi)\,\partial_t u_s + \phi\, v_f\big) = \rho_f^{-1}\theta, & \text{in } \Omega\times(0,T). \text{(3.2c)} \end{cases}$$

Note that in this case, the interstitial pressure is no longer a state variable since the term $\frac{b-\phi}{\kappa}\partial_t p$ vanishes, but rather a Lagrange multiplier associated with the incompressibility constraint

$$\operatorname{div}\left((1-\phi)\,\partial_t u_s + \phi\, v_f\right) = \rho_f^{-1}\theta.$$

The model (3.2) has to be complemented with initial and boundary conditions. For the sake of simplicity, we will restrict our study to the case of homogeneous Dirichlet boundary conditions

$$\begin{aligned} u_s &= 0, \quad \text{on } \partial\Omega, \\ v_f &= 0, \quad \text{on } \partial\Omega, \end{aligned} \tag{3.3}$$

where the motion of the porous medium is fixed on the boundary. For other types of boundary conditions such as Neumann or total stress boundary conditions, we refer the reader to [Burtschell et al., 2019]. Furthermore, we assume that an initial condition $(u_{s0}, v_{s0}, v_{f_0})$ is given, so that

$$\begin{cases} u_s(0) = u_{s0}, & \text{in } \Omega, \\ \partial_t u_s(0) = v_{s0}, & \text{in } \Omega, \\ v_f(0) = v_{f_0}, & \text{in } \Omega. \end{cases}$$

Through the rest of the paper, we will suppose that $(u_{s0}, v_{s0}, v_{f_0})$ is sufficiently regular.

3.1.2 Energy balance

One of the specificities of the model (3.1) – and also of the original non-linear model proposed in [Chapelle and Moireau, 2014] – compared to other poromechanics models is that it satisfies a natural energy balance. Before deriving this balance, we observe that we may assume without loss of generality that the right-hand side of the constraint equation is equal to zero. As a matter of fact, if it is not the case, we can build a divergence lifting v_θ such that $\operatorname{div} v_\theta = \rho_f^{-1}\theta$ and perform the change of variable $(u_s, v_f) \mapsto \left(u_s - \int_0^t v_\theta \,\mathrm{d}s, v_f - v_\theta\right)$. The existence of such a lifting requires that θ is regular enough and that

$$\int_\Omega \theta \,\mathrm{d}x = 0,$$

where the last assumption is a compatibility condition coming from the Dirichlet boundary condition. Indeed, if u_s and v_f satisfy (3.2c) and (3.3), then Stokes formula implies that

$$\int_\Omega \theta \,\mathrm{d}x = \rho_f \int_\Omega \operatorname{div}\left((1-\phi)\,\partial_t u_s + \phi\, v_f\right)\mathrm{d}x = \rho_f \int_{\partial\Omega} \left((1-\phi)\,\partial_t u_s + \phi\, v_f\right)\cdot n \,\mathrm{d}s = 0.$$

For all these reasons, we will suppose from now on that the right-hand side of (3.2c) is equal to zero. Note that the fluid mass input term θ also appears in (3.2b) through the term $-\theta v_f$. We will keep this term in (3.2b) since it is not affected by the above lifting, leading to a more general result than in [Burtschell et al., 2019] where it is assumed that $\theta = 0$.

Formally, multiplying (3.2a) by $\partial_t u_s$, (3.2b) by v_f and integrating by parts in space, we then obtain the energy identity

$$\begin{aligned} &\underbrace{\frac{\rho_s}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega (1-\phi)|\partial_t u_s|^2 \,\mathrm{d}x}_{\text{Structure kinetic energy}} + \underbrace{\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x}_{\text{Structure mechanical energy}} + \underbrace{\frac{\rho_f}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi|v_f|^2 \,\mathrm{d}x}_{\text{Fluid kinetic energy}} \\ &\quad + \underbrace{\int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(v_f)\,\mathrm{d}x}_{\text{Viscous dissipation within the fluid}} + \underbrace{\int_\Omega \phi^2 k_f^{-1}(v_f - \partial_t u_s)\cdot(v_f - \partial_t u_s)\,\mathrm{d}x}_{\text{Friction dissipation between solid and fluid phases}} \\ &\quad = \underbrace{\int_\Omega \theta|v_f|^2\,\mathrm{d}x}_{\text{Incoming rate of fluid kinetic energy}} + \underbrace{\int_\Omega \rho_s(1-\phi) f\cdot\partial_t u_s \,\mathrm{d}x + \int_\Omega \rho_f \phi f\cdot v_f \,\mathrm{d}x}_{\text{Power of external forces}}, \end{aligned} \tag{3.4}$$

where the physical meaning of each of the terms is indicated below them.

Guided by the above identity, we make the following hypotheses on the data:

($h1$) The constants ρ_s, ρ_f, μ_f, λ, μ are assumed to be strictly positive;

($h2$) The porosity $\phi \in \mathrm{H}^{d/2+r}(\Omega)$ with $r > 0$, and is such that there exists $(\phi_{\min}, \phi_{\max})$ satisfying

$$0 < \phi_{\min} \leq \phi(x) \leq \phi_{\max} < 1, \quad \forall x \in \Omega;$$

($h3$) The hydraulic conductivity tensor k_f is invertible and there exists $k_{\min}^{-1} > 0$ such that

$$k_f^{-1} v \cdot v \geq k_{\min}^{-1} |v|^2, \quad \forall v \in \mathbb{R}^d;$$

($h4$) $\theta \in L^\infty(\Omega)$ in addition to $\int_\Omega \theta \, \mathrm{d}x = 0$;

($h5$) $f \in L^2(0,T;[\mathrm{L}^2(\Omega)]^d)$.

Remark 3.1. In ($h4$), we assume for the sake of simplicity that the fluid mass input term is independent of time. It simplifies the analysis, but the time-dependent case could be handled by supposing that θ is regular enough, in particular $\theta \in \mathrm{C}^0\big([0,T] \times \Omega\big)$, see [Barré et al., 2023].

Remark 3.2. If the right-hand side of (3.2c) is not assumed to be equal to zero, an extra term $\displaystyle\int_\Omega \frac{p}{\rho_f} \theta \, \mathrm{d}x$ appears in the right-hand side of (3.4), which corresponds to an incoming rate of Gibbs free energy, see [Chapelle and Moireau, 2014].

Under assumptions ($h1$) – ($h5$), the application of Grönwall Lemma to the energy balance (3.4) allows us to control the growth of the total energy defined by

$$\mathcal{E}(t) = \frac{\rho_s}{2} \int_\Omega (1-\phi) |\partial_t u_s(t)|^2 \, \mathrm{d}x + \frac{1}{2} \int_\Omega \sigma_s\big(u_s(t)\big) : \varepsilon\big(u_s(t)\big) \, \mathrm{d}x + \frac{\rho_f}{2} \int_\Omega \phi |v_f(t)|^2 \, \mathrm{d}x,$$

and of the total dissipation defined by

$$\mathcal{D}(t) = \int_\Omega \phi \, \sigma_f\big(v_f(t)\big) : \varepsilon\big(v_f(t)\big) \, \mathrm{d}x + \int_\Omega \phi^2 k_f^{-1} \big(v_f(t) - \partial_t u_s(t)\big) \cdot \big(v_f(t) - \partial_t u_s(t)\big) \, \mathrm{d}x.$$

As a matter of fact, with these notation, (3.4) reads: for each $t \in (0,T)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}(t) + \mathcal{D}(t) = \int_\Omega \theta \, |v_f(t)|^2 \, \mathrm{d}x + F(t),$$

with

$$F(t) = \int_\Omega \rho_s (1-\phi) f(t) \cdot \partial_t u_s(t) \, \mathrm{d}x + \int_\Omega \rho_f \phi f(t) \cdot v_f(t) \, \mathrm{d}x.$$

Then, three different situations occur depending on the fluid mass input term θ: either (a) θ is *negative*, or (b) θ is possibly positive but remains *small* – in a sense specified below, or finally (c) θ is possibly positive and *large*.

(a) If θ is negative, namely if fluid mass is removed from the system, then

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}(t) + \mathcal{D}(t) + \int_\Omega |\theta| \, |v_f(t)|^2 \, \mathrm{d}x &= F(t) \\ &\leq \|f(t)\|_{\rho_s(1-\phi)} \, \|\partial_t u_s(t)\|_{\rho_s(1-\phi)} + \|f(t)\|_{\rho_f \phi} \, \|v_f(t)\|_{\rho_f \phi} \\ &\leq \sqrt{2} \big(\|f(t)\|_{\rho_s(1-\phi)} + \|f(t)\|_{\rho_f \phi} \big) \mathcal{E}(t)^{1/2}, \end{aligned}$$

where we used the notation $\|v\|_\alpha$ for the $[\mathrm{L}^2(\Omega)]^d$ norm scaled by a function $\alpha(x)$, namely $\|v\|_\alpha^2 = \int_\Omega \alpha\,|v|^2\,\mathrm{d}x$. Therefore, Grönwall Lemma yields: for each $t \in (0,T)$,

$$\mathcal{E}(t) + \int_0^t \mathcal{D}(s)\,\mathrm{d}s + \int_0^t \int_\Omega |\theta|\,|v_f(s)|^2\,\mathrm{d}x\,\mathrm{d}s \le \left(\mathcal{E}(0) + \frac{\sqrt{2}}{2}\int_0^t \big(\|f(s)\|_{\rho_s(1-\phi)} + \|f(s)\|_{\rho_f\phi}\big)\,\mathrm{d}s\right)^2.$$

(b) When θ can be positive but small enough, the incoming rate of fluid kinetic energy can be compensated by the fluid viscous dissipation. To do so, let us recall Korn inequality [Ciarlet, 1988], which states that there exists $C > 0$ such that

$$\int_\Omega \varepsilon(v) : \varepsilon(v)\,\mathrm{d}x \ge C\|v\|^2_{[\mathrm{H}^1_0(\Omega)]^d}, \quad \forall v \in [\mathrm{H}^1_0(\Omega)]^d. \tag{3.5}$$

Combining (3.5) with Poincaré inequality, we know that there exists a constant $C_d > 0$ such that

$$\int_\Omega |v|^2\,\mathrm{d}x \le C_d \int_\Omega \varepsilon(v) : \varepsilon(v)\,\mathrm{d}x, \quad \forall v \in [\mathrm{H}^1_0(\Omega)]^d.$$

Hence

$$\int_\Omega \theta\,|v_f(t)|^2\,\mathrm{d}x \le \frac{C_d\,\|\theta\|_{\mathrm{L}^\infty(\Omega)}}{2\mu_f\phi_{\min}} \int_\Omega \phi\,\sigma_f\big(v_f(t)\big) : \varepsilon\big(v_f(t)\big)\,\mathrm{d}x,$$

so that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(t) + \left(1 - \frac{C_d\,\|\theta\|_{\mathrm{L}^\infty(\Omega)}}{2\mu_f\phi_{\min}}\right)\int_\Omega \phi\,\sigma_f\big(v_f(t)\big) : \varepsilon\big(v_f(t)\big)\,\mathrm{d}x$$
$$+ \int_\Omega \phi^2 k_f^{-1}\big(v_f(t) - \partial_t u_s(t)\big)\cdot\big(v_f(t) - \partial_t u_s(t)\big)\,\mathrm{d}x \le \sqrt{2}\big(\|f(t)\|_{\rho_s(1-\phi)} + \|f(t)\|_{\rho_f\phi}\big)\mathcal{E}(t)^{1/2},$$

provided that

$$\frac{C_d\,\|\theta\|_{\mathrm{L}^\infty(\Omega)}}{2\mu_f\phi_{\min}} \le 1. \tag{3.6}$$

As a consequence, if (3.6) is satisfied, then for each $t \in (0,T)$ we have

$$\mathcal{E}(t) + \left(1 - \frac{C_d\,\|\theta\|_{\mathrm{L}^\infty(\Omega)}}{2\mu_f\phi_{\min}}\right)\int_0^t\int_\Omega \phi\,\sigma_f\big(v_f(s)\big) : \varepsilon\big(v_f(s)\big)\,\mathrm{d}x\,\mathrm{d}s$$
$$+ \int_0^t\int_\Omega \phi^2 k_f^{-1}\big(v_f(s) - \partial_t u_s(s)\big)\cdot\big(v_f(s) - \partial_t u_s(s)\big)\,\mathrm{d}x\,\mathrm{d}s$$
$$\le \left(\mathcal{E}(0) + \frac{\sqrt{2}}{2}\int_0^t \big(\|f(s)\|_{\rho_s(1-\phi)} + \|f(s)\|_{\rho_f\phi}\big)\,\mathrm{d}s\right)^2.$$

(c) In the general case where θ can be positive and taking possibly large values, we use Young inequality to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(t) + \mathcal{D}(t) = \int_\Omega \theta\,|v_f(t)|^2\,\mathrm{d}x + F(t) \le \frac{2\,\|\theta\|_{\mathrm{L}^\infty(\Omega)}}{\rho_f\phi_{\min}}\cdot\frac{1}{2}\int_\Omega \rho_f\phi_{\min}\,|v_f(t)|^2\,\mathrm{d}x$$
$$+ \frac{1}{2}\|f(t)\|^2_{\rho_s(1-\phi)} + \frac{1}{2}\|\partial_t u_s(t)\|^2_{\rho_s(1-\phi)} + \frac{1}{2}\|f(t)\|^2_{\rho_f\phi} + \frac{1}{2}\|v_f(t)\|^2_{\rho_f\phi}$$
$$\le \left(1 + \frac{2\|\theta\|_{L^\infty(\Omega)}}{\rho_f\phi_{\min}}\right)\mathcal{E}(t) + \frac{1}{2}\|f(t)\|^2_{\rho_s(1-\phi)} + \frac{1}{2}\|f(t)\|^2_{\rho_f\phi},$$

leading to

$$\mathcal{E}(t)+\int_0^t \mathcal{D}(s)\,\mathrm{d}s \leq \exp\left(\left(1+\frac{2\|\theta\|_{L^\infty(\Omega)}}{\rho_f\phi_{\min}}\right)t\right)\left(\mathcal{E}(0)+\frac{1}{2}\int_0^t \left(\|f(s)\|^2_{\rho_s(1-\phi)}+\|f(s)\|^2_{\rho_f\phi}\right)\mathrm{d}s\right). \quad (3.7)$$

In this article, to remain as general as possible, we will focus on the case c) where θ may take large values and for which the solution possibly shows an exponential growth as in (3.7). Indeed, this general case was not covered in the litterature, in particular [Burtschell et al., 2019; Barnafi et al., 2021]. Note however that in the case where θ satisfies (3.6) as it is assumed in [Barnafi et al., 2021], our analysis also provides error estimates with no exponential growth, see Remarks 3.11, 3.18 and 3.22. The energy estimate (3.7), which has been theoretically proven in [Barré et al., 2023], is a fundamental property of the system and its discrete counterpart will be the cornerstone of the numerical analysis. But before proposing a discretization of Problem (3.2) in the next section, let us briefly recall existence and uniqueness results at the continuous level and introduce a few notation.

3.1.3 Existence results

From the theoretical point of view, Problem (3.2) combines two major difficulties. The first one is that the solid equation (3.2a) is hyperbolic, whereas the fluid equation (3.2b) is parabolic. The second one is the incompressibility constraint (3.2c) coupling the solid and fluid velocities. Therefore, system (3.2) is a strongly coupled problem with a hyperbolic-parabolic coupling that also involves a saddle-point structure associated with a non-standard divergence constraint.

The existence of strong, mild and weak solutions of Problem (3.2) has been studied and justified in detail in [Barré et al., 2023] using a semigroup approach and the notion of T-coercivity [Chesnel and Ciarlet, 2013]. The first step is to formulate our problem as a first-order evolution system. Introducing the solid velocity variable $v_s = \partial_t u_s$, Problem (3.2) can be rewritten as: find (u_s, v_s, v_f, p) such that

$$\begin{cases} \partial_t u_s - v_s = 0, & (3.8a)\\ \rho_s(1-\phi)\,\partial_t v_s - \operatorname{div}\left(\sigma_s(u_s)\right) - \phi^2\, k_f^{-1}(v_f - v_s) + (1-\phi)\,\nabla p = \rho_s(1-\phi)\,f, & (3.8b)\\ \rho_f\phi\,\partial_t v_f - \operatorname{div}\left(\phi\,\sigma_f(v_f)\right) + \phi^2\, k_f^{-1}(v_f - v_s) - \theta\, v_f + \phi\,\nabla p = \rho_f\phi\, f, & (3.8c)\\ \operatorname{div}\left((1-\phi)\,v_s + \phi\, v_f\right) = 0. & (3.8d) \end{cases}$$

Then, denoting by $z = (u_s, v_s, v_f)$ the state variable, we seek for a solution z in the energy space

$$H = [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d,$$

endowed with the scalar product

$$(z,y)_H = \int_\Omega \sigma_s(u_s) : \varepsilon(d_s) + \int_\Omega \rho_s(1-\phi)\,v_s \cdot w_s\,\mathrm{d}x + \int_\Omega \rho_f\phi\, v_f \cdot w_f\,\mathrm{d}x,$$

for any $y = (d_s, w_s, w_f)$ belonging to H, and with the corresponding norm

$$\|z\|^2_H = \int_\Omega \sigma_s(u_s) : \varepsilon(u_s)\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,|v_s|^2\,\mathrm{d}x + \int_\Omega \rho_f\phi\,|v_f|^2\,\mathrm{d}x,$$

associated with the energy balance (3.4). Note that this norm is equivalent to the canonical norm on H thanks to Korn inequality (3.5). Setting

$$V = [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \quad \text{and} \quad Q = \mathrm{L}^2_0(\Omega),$$

we define the *all-in-one* mixed bilinear form *incorporating the constraint*

$$\begin{aligned}\mathcal{A}\big((z,p),(y,q)\big) = &-\int_\Omega \sigma_s(v_s) : \varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s) : \varepsilon(w_s)\,\mathrm{d}x \\ &+ \int_\Omega \phi^2\, k_f^{-1}(v_f - v_s)\cdot(w_f - w_s)\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_f) : \varepsilon(w_f)\,\mathrm{d}x - \int_\Omega \theta v_f \cdot w_f\,\mathrm{d}x \\ &- \int_\Omega p\,\mathrm{div}\left((1-\phi)\,w_s + \phi\,w_f\right)\mathrm{d}x + \int_\Omega \mathrm{div}\left((1-\phi)\,v_s + \phi\,v_f\right) q\,\mathrm{d}x, \end{aligned} \tag{3.9}$$

for all $z = (u_s, v_s, v_f)$, $y = (d_s, w_s, w_f)$ in V and p, q in Q. Within this functional framework, the mixed formulation of Problem (3.8) reads

$$\begin{cases} \text{Find } z = (u_s, v_s, v_f) \in C^1\big([0,T];H\big) \cap C^0\big([0,T];V\big) \text{ and } p \in C^0\big([0,T];Q\big) \quad \text{such that} \\ \big(\dot z(t), y\big)_H + \mathcal{A}\big((z(t),p(t)),(y,q)\big) = (g(t),y)_H, \quad \forall y \in V, \forall q \in Q, \end{cases} \tag{3.10}$$

with $\dot z = \frac{\mathrm{d}}{\mathrm{d}t} z$ and $g(t) = (0, f(t), f(t))$. From [Barré et al., 2023, Theorem 3.14], we know that this formulation is well-posed. The solution of Problem (3.10) satisfies (3.8b), (3.8c) and (3.8d) in $[\mathrm{L}^2(\Omega)]^d$, whereas the identity (3.8a) is fulfilled in the space $[\mathrm{H}^1_0(\Omega)]^d$, endowed with the specific scalar product $(u,v) \mapsto \int_\Omega \sigma_s(u) : \varepsilon(v)$ adapted to the elasticity operator $-\,\mathrm{div}\left(\sigma_s(\cdot)\right)$.

The proof of well-posedness in [Barré et al., 2023] hinges on showing that the evolution operator associated with (3.8) is maximal-accretive. This property can be proven using the notion of T-coercivity, that will also be a central tool for the numerical analysis of the discrete problem and that we present below.

3.1.4 The T-coercivity approach

The T-coercivity approach is a reformulation of Banach-Nečas-Babuška theory for the study of well-posedness and numerical approximation of non-coercive problems. T-coercivity was originally introduced for problems involving an invertible operator perturbed by a compact term [Buffa, 2005; Ciarlet Jr, 2012] and problems with sign-changing coefficients, see for instance [Bonnet-Ben Dhia et al., 2010b; Chesnel and Ciarlet, 2013; Bonnet-Ben Dhia et al., 2014a; Bunoiu et al., 2021; Halla, 2021]. More recently, it was applied to saddle-point problems [Barré et al., 2023; Barré and Ciarlet Jr, 2022]. This approach is particularly appropriate here as it allows us to handle the two difficulties of the problem – the incompressibility constraint and the non-coercivity of the underlying operator coming from the hyperbolic-parabolic coupling – in a *monolithic* way by analyzing the *all-in-one* bilinear form (3.9).

For the sake of completeness, the definition and main property of T-coercivity are recalled below at the continuous level.

Definition 3.3. Let W be a Hilbert space and let $\mathcal{A}(\cdot,\cdot)$ be a continuous bilinear form over $W \times W$. We say that $\mathcal{A}$ is T-*coercive* if there exists a bijective operator $\mathtt{T} \in \mathcal{L}(W)$ and $\underline{\alpha} > 0$ such that

$$|\mathcal{A}(u, \mathtt{T}u)| \geq \underline{\alpha}\,\|u\|_W^2\,, \quad \forall u \in W.$$

Proposition 3.4. *Let W be a Hilbert space. Let $\ell(\cdot)$ be a continuous linear form over W and $\mathcal{A}(\cdot,\cdot)$ be a continuous bilinear form over $W \times W$. The problem*

$$\begin{cases} \textit{Find } u \in W \textit{ such that} \\ \forall v \in W, \quad \mathcal{A}(u,v) = \ell(v) \end{cases}$$

is well-posed if and only if $\mathcal{A}$ is T*-coercive.*

The all-in-one bilinear form $\mathcal{A}$ is not coercive, but it was shown in [Barré et al., 2023, Proposition 3.15] that the bilinear form $\mathcal{A}_{\lambda_0}$ defined by

$$\mathcal{A}_{\lambda_0}\big((z,p),(y,q)\big) = \mathcal{A}\big((z,p),(y,q)\big) + \lambda_0 (z,y)_H, \tag{3.11}$$

is T-coercive provided that the parameter λ_0 is large enough. More precisely, we have the following result.

Proposition 3.5. *If $\lambda_0 > (\rho_f \phi_{\min})^{-1} \|\theta\|_{L^\infty(\Omega)}$, then the bilinear form $\mathcal{A}_{\lambda_0}$ is T-coercive for the mapping*

$$T : (u_s, v_s, v_f, p) \longmapsto \Big(\beta u_s + \gamma v_s, \alpha v_s - v_p, \alpha v_f - v_p, \alpha p\Big), \tag{3.12}$$

where $v_p \in [\mathrm{H}^1_0(\Omega)]^d$ is a divergence lifting defined by

$$\operatorname{div} v_p = p \quad \text{and} \quad \|\nabla v_p\| \leq C_{\mathrm{div}} \|p\|, \tag{3.13}$$

with $C_{\mathrm{div}} > 0$, and α, β and γ are constants depending on λ_0 and the various physical parameters.

As can be seen in Definition 3.3 and Proposition 3.5, the T-coercivity framework relies on the *explicit* building of an operator T such that the bilinear form under study is T-coercive. This explicit realization provides insights on how to design a suitable approximation of the continuous problem when going at the discrete level. Indeed, the following results [Chesnel and Ciarlet, 2013] indicate that if one is able to reproduce the continuous mapping T in the discrete setting, then the convergence of the associated discrete solution is ensured.

Definition 3.6. Let W be a Hilbert space, $\mathcal{A}(\cdot,\cdot)$ be a continuous bilinear form over $W \times W$ and $(W_h)_h$ be conforming approximations of W. We say that $\mathcal{A}$ is *uniformly $\mathtt{T}_h$-coercive* if

$$\exists \alpha^*, \beta^* > 0,\ \forall h > 0,\ \exists \mathtt{T}_h \in \mathcal{L}(W_h), \quad |\mathcal{A}(u_h, \mathtt{T}_h u_h)| \geq \alpha^* \|u_h\|^2_W, \quad \forall u_h \in W_h, \quad \text{and} \quad |||\mathtt{T}_h||| \leq \beta^*.$$

Proposition 3.7. *Assume that the hypotheses of Proposition 3.4 hold and that the bilinear form $\mathcal{A}$ is T-coercive. Let $(W_h)_h$ be conforming approximations of W, and denote by $\boldsymbol{A}_h \in \mathcal{L}(W_h, W_h')$ the discrete operator associated with $\mathcal{A}_{|W_h}$. The problem*

$$\begin{cases} \text{Find } u_h \in W_h \text{ such that} \\ \forall v_h \in W_h, \quad \mathcal{A}(u_h, v_h) = \ell(v_h) \end{cases}$$

is well-posed and $(\boldsymbol{A}_h^{-1})_h$ is uniformly bounded if and only if $\mathcal{A}$ is uniformly $\mathtt{T}_h$-coercive. In that case, denoting by $C_{\mathcal{A}}$ the continuity constant of the bilinear form $\mathcal{A}$, it holds that

$$\|u - u_h\|_W \leq C \inf_{v_h \in W_h} \|u - v_h\|_W, \tag{3.14}$$

with $C = 1 + \frac{C_{\mathcal{A}} \beta^*}{\alpha^*}$ independent of h.

The approximation property (3.14) will enable us to build a discrete projection operator on the finite dimension space considered that is adapted to the specific structure of Problem (3.8).

3.2 Two discretization schemes

3.2.1 Semi-discrete time discretization

We propose two monolithic time schemes to discretize Problem (3.8). The first one is a Crank-Nicolson scheme [Crank and Nicolson, 1947], in which both the solid and fluid quantities are discretized using a midpoint rule. In the second one, the solid part is still discretized with a midpoint rule but the fluid and pressure parts are approximated with an implicit backward Euler method.

This second scheme is motivated by the fact of reproducing in a linearized setting the splitting scheme introduced in [Burtschell et al., 2017] for the non-linear model from [Chapelle and Moireau, 2014], in which the solid and fluid parts are discretized respectively with Newmark and backward Euler schemes following [Hauret and Le Tallec, 2006]. These schemes are close to those studied in [Burtschell et al., 2019] and [Barnafi et al., 2021] but include the additional fluid mass term θ and cover the incompressible regime.

The interval $(0,T)$ is divided into n_T time intervals. Let us denote by $\Delta t = \frac{T}{n_T}$ the time step of the method, and by $t^n = n\Delta t$ the discrete times, with initial time $t^0 = 0$ and final time $t^{n_T} = n_T \Delta t = T$. The continuous solution (u_s, v_s, v_f, p) at time t^n will be approximated by the semi-discrete solution $(u_s^n, v_s^n, v_f^n, p^n)$, which is initialized by

$$(u_s^0, v_s^0, v_f^0, p^0) = (u_{s0}, v_{s0}, v_{f_0}, 0).$$

Moreover, we will denote by $u_s^{n+\frac{1}{2}}$, $v_s^{n+\frac{1}{2}}$, $v_f^{n+\frac{1}{2}}$ and $p^{n+\frac{1}{2}}$ the midpoint quantities

$$u_s^{n+\frac{1}{2}} = \frac{u_s^{n+1} + u_s^n}{2}, \quad v_s^{n+\frac{1}{2}} = \frac{v_s^{n+1} + v_s^n}{2}, \quad v_f^{n+\frac{1}{2}} = \frac{v_f^{n+1} + v_f^n}{2}, \quad p^{n+\frac{1}{2}} = \frac{p^{n+1} + p^n}{2},$$

which correspond to an approximation of the solution at time $t^{n+\frac{1}{2}} = \left(n + \frac{1}{2}\right)\Delta t$.

Under these notation, the proposed semi-discrete Crank-Nicolson scheme reads:

$$\begin{cases} \dfrac{u_s^{n+1} - u_s^n}{\Delta t} - v_s^{n+\frac{1}{2}} = 0, & \text{(3.15a)} \\ \rho_s(1-\phi)\dfrac{v_s^{n+1} - v_s^n}{\Delta t} - \operatorname{div}\left(\sigma_s(u_s^{n+\frac{1}{2}})\right) \\ \qquad - \phi^2 k_f^{-1}(v_f^{n+\frac{1}{2}} - v_s^{n+\frac{1}{2}}) + (1-\phi)\nabla p^{n+\frac{1}{2}} = \rho_s(1-\phi) f^{n+\frac{1}{2}}, & \text{(3.15b)} \\ \rho_f \phi \dfrac{v_f^{n+1} - v_f^n}{\Delta t} - \operatorname{div}\left(\phi\, \sigma_f(v_f^{n+\frac{1}{2}})\right) \\ \qquad + \phi^2 k_f^{-1}(v_f^{n+\frac{1}{2}} - v_s^{n+\frac{1}{2}}) - \theta\, v_f^{n+\frac{1}{2}} + \phi \nabla p^{n+\frac{1}{2}} = \rho_f \phi\, f^{n+\frac{1}{2}}, & \text{(3.15c)} \\ \operatorname{div}\left((1-\phi)\, v_s^{n+\frac{1}{2}} + \phi\, v_f^{n+\frac{1}{2}}\right) = 0, & \text{(3.15d)} \end{cases}$$

where the discrete external body force $f^{n+\frac{1}{2}}$ is defined by

$$f^{n+\frac{1}{2}} = \frac{f(t^{n+1}) + f(t^n)}{2}.$$

The second proposed scheme, which will be referred to as backward Euler scheme, then consists in

$$\begin{cases} \dfrac{u_s^{n+1} - u_s^n}{\Delta t} - v_s^{n+\frac{1}{2}} = 0, & \text{(3.16a)} \\ \rho_s(1-\phi)\dfrac{v_s^{n+1} - v_s^n}{\Delta t} - \operatorname{div}\left(\sigma_s(u_s^{n+\frac{1}{2}})\right) \\ \qquad - \phi^2 k_f^{-1}(v_f^{n+1} - v_s^{n+\frac{1}{2}}) + (1-\phi)\nabla p^{n+1} = \rho_s(1-\phi) f^{n+\frac{1}{2}}, & \text{(3.16b)} \\ \rho_f \phi \dfrac{v_f^{n+1} - v_f^n}{\Delta t} - \operatorname{div}\left(\phi\, \sigma_f(v_f^{n+1})\right) \\ \qquad + \phi^2 k_f^{-1}(v_f^{n+1} - v_s^{n+\frac{1}{2}}) - \theta\, v_f^{n+1} + \phi \nabla p^{n+1} = \rho_f \phi\, f^{n+\frac{1}{2}}, & \text{(3.16c)} \\ \operatorname{div}\left((1-\phi)\, v_s^{n+\frac{1}{2}} + \phi\, v_f^{n+1}\right) = 0. & \text{(3.16d)} \end{cases}$$

Note that a similar scheme was proposed in [Burtschell et al., 2019] to discretize the viscous and compressible system (3.1) with $\theta = 0$. However, the convergence estimates in [Burtschell et al., 2019] depend on κ and hence are not valid for the limit case $\kappa = \infty$. Here, we consider the non-viscous and incompressible case with $\theta \neq 0$, which leads to additional difficulties since we have to deal with a hyperbolic-parabolic coupled system with a constraint on the mixture velocity and with possible unstabilities arising from the fluid additional mass input.

Remark 3.8. The two schemes (3.15) and (3.16) are written as a four-field formulation to benefit from the existence results obtained at the continuous level. However, it is more efficient in practice to eliminate the solid velocity variable thanks to the relations

$$\begin{aligned} v_s^{n+\frac{1}{2}} &= \frac{u_s^{n+1} - u_s^n}{\Delta t}, \qquad v_s^{n+1} = 2v_s^{n+\frac{1}{2}} - v_s^n = 2\frac{u_s^{n+1} - u_s^n}{\Delta t} - v_s^n, \\ \frac{v_s^{n+1} - v_s^n}{\Delta t} &= \frac{2}{\Delta t}\big(v_s^{n+\frac{1}{2}} - v_s^n\big) = \frac{2}{\Delta t^2}\big(u_s^{n+1} - u_s^n - \Delta t v_s^n\big), \end{aligned} \tag{3.17}$$

and solve a three-field formulation.

3.2.2 Fully discrete schemes

For the space discretization, we consider two finite dimensional spaces $X_h \subset [\mathrm{H}_0^1(\Omega)]^d$ and $Q_h \subset \mathrm{L}_0^2(\Omega)$ constituting a *conforming* approximation of $[\mathrm{H}_0^1(\Omega)]^d$ and $\mathrm{L}_0^2(\Omega)$. We seek for the vectorial quantities – both solid and fluid – in the discrete space X_h and for the pressure in the discrete space Q_h. Moreover, in order to take into account the incompressibility constraint (3.8d), we assume that (X_h, Q_h) are selected in order to satisfy the *uniform discrete inf-sup condition*

$$\exists \beta > 0, \forall p_h \in Q_h, \qquad \sup_{v_h \in X_h} \frac{\displaystyle\int_\Omega \operatorname{div} v_h \, p_h \, \mathrm{d}x}{\|v_h\|_{[\mathrm{H}_0^1(\Omega)]^d}} \geq \beta \, \|p_h\| \, . \tag{3.18}$$

Note that this is the inf-sup condition associated with the standard divergence constraint that has been widely studied in the scope of Stokes equation. This condition does not depend on the porosity, as opposed to the hypotheses made in [Burtschell et al., 2019]. Therefore, to choose the pair (X_h, Q_h), we can use the large literature existing on this topic for Stokes equations [Glowinski, 2003; Girault and Raviart, 1986; Boffi et al., 2013]: possible choices include for instance Taylor-Hood elements or the MINI element. All these choices rely on the design of a *Fortin operator* $\Pi_h : [\mathrm{H}_0^1(\Omega)]^d \mapsto X_h$ satisfying, for each $v \in [\mathrm{H}_0^1(\Omega)]^d$,

- For all $q_h \in Q_h$,
$$\int_\Omega \operatorname{div}\big(\Pi_h(v)\big) \, q_h \, \mathrm{d}x = \int_\Omega \operatorname{div} v \, q_h \, \mathrm{d}x, \tag{3.19}$$
- There exists a constant $C_\pi > 0$ independent of h such that
$$\big\|\nabla\big(\Pi_h(v)\big)\big\| \leq C_\pi \, \|\nabla v\| \, . \tag{3.20}$$

The existence of such an operator is ensured by the inf-sup condition (3.18) by virtue of the Closed Range Theorem. Following [Barré and Ciarlet Jr, 2022], we will use this Fortin operator rather than the inf-sup condition (3.18) to build a $\mathtt{T}_h$-coercivity mapping adapted to the mixture's divergence constraint (3.8d) and to the specific structure of Problem (3.8).

After selecting the spaces X_h and Q_h, the fully-discrete versions of the Crank-Nicolson scheme (3.15) and the backward Euler scheme (3.16) respectively amount to finding $u_{s,h}^{n+1}, v_{s,h}^{n+1}, v_{f,h}^{n+1}, p_h^{n+1} \in$

$X_h \times X_h \times X_h \times Q_h$ at each time step such that for all $(d_{s,h}, w_{s,h}, w_{f,h}, q_h) \in X_h \times X_h \times X_h \times Q_h$,

$$\begin{aligned}
&\int_\Omega \sigma_s\left(\frac{u_{s,h}^{n+1}-u_{s,h}^{n}}{\Delta t}\right) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,\frac{v_{s,h}^{n+1}-v_{s,h}^{n}}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,\frac{v_{f,h}^{n+1}-v_{f,h}^{n}}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad - \int_\Omega \sigma_s(v_{s,h}^{n+\frac12}) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \sigma_s(u_{s,h}^{n+\frac12}) : \varepsilon(w_{s,h})\,\mathrm{d}x \\
&+ \int_\Omega \phi\,\sigma_f(v_{f,h}^{n+\frac12}) : \varepsilon(w_{f,h})\,\mathrm{d}x - \int_\Omega \theta v_{f,h}^{n+\frac12}\cdot w_{f,h}\,\mathrm{d}x + \int_\Omega \phi^2\,k_f^{-1}(v_{f,h}^{n+\frac12}-v_{s,h}^{n+\frac12})\cdot(w_{f,h}-w_{s,h})\,\mathrm{d}x \\
&\qquad - \int_\Omega p_h^{n+\frac12}\,\mathrm{div}\left((1-\phi)\,w_{s,h}+\phi\,w_{f,h}\right)\mathrm{d}x + \int_\Omega \mathrm{div}\left((1-\phi)\,v_{s,h}^{n+\frac12}+\phi\,v_{f,h}^{n+\frac12}\right)q_h\,\mathrm{d}x \\
&\qquad\qquad = \int_\Omega \rho_s(1-\phi)\,f^{n+\frac12}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,f^{n+\frac12}\cdot w_{f,h}\,\mathrm{d}x, \quad (3.21)
\end{aligned}$$

or

$$\begin{aligned}
&\int_\Omega \sigma_s\left(\frac{u_{s,h}^{n+1}-u_{s,h}^{n}}{\Delta t}\right) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,\frac{v_{s,h}^{n+1}-v_{s,h}^{n}}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,\frac{v_{f,h}^{n+1}-v_{f,h}^{n}}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad - \int_\Omega \sigma_s(v_{s,h}^{n+\frac12}) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \sigma_s(u_{s,h}^{n+\frac12}) : \varepsilon(w_{s,h})\,\mathrm{d}x \\
&+ \int_\Omega \phi\,\sigma_f(v_{f,h}^{n+1}) : \varepsilon(w_{f,h})\,\mathrm{d}x - \int_\Omega \theta v_{f,h}^{n+1}\cdot w_{f,h}\,\mathrm{d}x + \int_\Omega \phi^2\,k_f^{-1}(v_{f,h}^{n+1}-v_{s,h}^{n+\frac12})\cdot(w_{f,h}-w_{s,h})\,\mathrm{d}x \\
&\qquad - \int_\Omega p_h^{n+1}\,\mathrm{div}\left((1-\phi)\,w_{s,h}+\phi\,w_{f,h}\right)\mathrm{d}x + \int_\Omega \mathrm{div}\left((1-\phi)\,v_{s,h}^{n+\frac12}+\phi\,v_{f,h}^{n+1}\right)q_h\,\mathrm{d}x \\
&\qquad\qquad = \int_\Omega \rho_s(1-\phi)\,f^{n+\frac12}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,f^{n+\frac12}\cdot w_{f,h}\,\mathrm{d}x. \quad (3.22)
\end{aligned}$$

Moreover, both schemes are initialized by

$$\left(u_{s,h}^0, v_{s,h}^0, v_{f,h}^0\right) = I_h\left(u_{s0}, v_{s0}, v_{f0}\right),$$

where I_h is the interpolation operator from $[\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d$ to $X_h \times X_h \times X_h$.

Setting $V_h = X_h \times X_h \times X_h$, introducing the notation

$$z_h^{n+1} = (u_{s,h}^{n+1}, v_{s,h}^{n+1}, v_{f,h}^{n+1}), \quad z_h^{n+\frac12} = (u_{s,h}^{n+\frac12}, v_{s,h}^{n+\frac12}, v_{f,h}^{n+\frac12}), \quad y_h = (d_{s,h}, w_{s,h}, w_{f,h}),$$

and recalling the definition of the bilinear form $\mathcal{A}$, the weak formulations (3.21) and (3.22) can be condensed into

$$\begin{cases} \text{Find } z_h^{n+1} \in V_h \text{ and } p_h^{n+1} \in Q_h \quad \text{such that for all } (y_h, q_h) \in V_h \times Q_h, \\ \left(\frac{z_h^{n+1}-z_h^n}{\Delta t}, y_h\right)_H + \mathcal{A}\big((z_h^{n+\frac12}, p_h^{n+\frac12}), (y_h, q_h)\big) = \big(g^{n+\frac12}, y_h\big)_H, \end{cases} \quad (3.23)$$

for the Crank-Nicolson scheme and

$$\begin{cases} \text{Find } z_h^{n+1} \in V_h \text{ and } p_h^{n+1} \in Q_h \quad \text{such that for all } (y_h, q_h) \in V_h \times Q_h, \\ \left(\frac{z_h^{n+1}-z_h^n}{\Delta t}, y_h\right)_H + \mathcal{A}\big((u_{s,h}^{n+\frac12}, v_{s,h}^{n+\frac12}, v_{f,h}^{n+1}, p_h^{n+1}), (y_h, q_h)\big) = \big(g^{n+\frac12}, y_h\big)_H, \end{cases} \quad (3.24)$$

for the backward Euler scheme, with $g^{n+\frac12} = (0, f^{n+\frac12}, f^{n+\frac12})$.

Remark 3.9. As already noticed in Remark 3.8, the four-field formulations (3.23) and (3.24) are convenient for the theoretical and numerical analysis of the problem but are not optimal when it comes to numerical efficiency. For example, to implement (3.21) in practice, it is preferable to remove the solid velocity variable using (3.17) and to solve the three-field formulation: find $u_{s,h}^{n+1}, v_{f,h}^{n+1}, p_h^{n+1} \in X_h \times X_h \times Q_h$ such that for all $(w_{s,h}, w_{f,h}, q_h) \in X_h \times X_h \times Q_h$,

$$\begin{aligned}
&\int_\Omega \rho_s(1-\phi)\,\frac{2}{\Delta t^2}\big(u_{s,h}^{n+1} - u_{s,h}^n - \Delta t v_{s,h}^n\big)\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,\frac{v_{f,h}^{n+1}-v_{f,h}^n}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad + \int_\Omega \sigma_s(u_{s,h}^{n+\frac12}) : \varepsilon(w_{s,h})\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(v_{f,h}^{n+\frac12}) : \varepsilon(w_{f,h})\,\mathrm{d}x - \int_\Omega \theta v_{f,h}^{n+\frac12}\cdot w_{f,h}\,\mathrm{d}x \\
&+ \int_\Omega \phi^2\,k_f^{-1}\left(v_{f,h}^{n+\frac12} - \frac{u_{s,h}^{n+1}-u_{s,h}^n}{\Delta t}\right)\cdot(w_{f,h}-w_{s,h})\,\mathrm{d}x - \int_\Omega p_h^{n+\frac12}\,\mathrm{div}\,\big((1-\phi)\,w_{s,h} + \phi\,w_{f,h}\big)\,\mathrm{d}x \\
&+ \int_\Omega \mathrm{div}\left((1-\phi)\,\frac{u_{s,h}^{n+1}-u_{s,h}^n}{\Delta t} + \phi\,v_{f,h}^{n+\frac12}\right) q_h\,\mathrm{d}x = \int_\Omega \rho_s(1-\phi)\,f^{n+\frac12}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,f^{n+\frac12}\cdot w_{f,h}\,\mathrm{d}x,
\end{aligned}$$

and then to post-process the solid velocity node by node with the formula

$$v_{s,h}^{n+1} = 2\frac{u_{s,h}^{n+1}-u_{s,h}^n}{\Delta t} - v_{s,h}^n.$$

We are now going to use the T-coercivity approach presented in Section 3.1.4 to see under which conditions the discrete problems (3.23) and (3.24) are well-posed. We will see that the two bilinear forms involved in (3.23) and (3.24) are closely related to the family of bilinear forms

$$\mathcal{A}_{\lambda_0}\big((z,p),(y,q)\big) = \mathcal{A}\big((z,p),(y,q)\big) + \lambda_0 (z,y)_H$$

introduced in (3.11). Therefore, we start with the more general result below.

Lemma 3.10. *Assume that (h1)−(h4) hold and that the discrete inf-sup condition (3.18) is satisfied. Let ℓ be a continuous linear form on $V \times Q$. If $\lambda_0 > (\rho_f\phi_{\min})^{-1}\,\|\theta\|_{\mathrm{L}^\infty(\Omega)}$, then the bilinear form $\mathcal{A}_{\lambda_0}$ is uniformly $\mathtt{T}_h$-coercive. In particular, the problem*

$$\begin{cases} \text{Find } (z_h,p_h) \in V_h \times Q_h \quad \text{such that} \\ \mathcal{A}_{\lambda_0}\big((z_h,p_h),(y_h,q_h)\big) = \ell\big((y_h,q_h)\big), \quad \forall y_h \in V_h, \forall q_h \in Q_h, \end{cases}$$

is well-posed and admits a solution that is uniformly bounded with respect to h. Moreover, there exists a constant $C > 0$ independent of h such that

$$\|(z,p) - (z_h,p_h)\|_{V\times Q} \le C \inf_{(y_h,q_h)\in V_h\times Q_h} \|(z,p) - (y_h,q_h)\|_{V\times Q}\,, \tag{3.25}$$

where (z,p) is the solution of the continuous problem

$$\begin{cases} \text{Find } (z,p) \in V \times Q \quad \text{such that} \\ \mathcal{A}_{\lambda_0}\big((z,p),(y,q)\big) = \ell\big((y,q)\big), \quad \forall y \in V, \forall q \in Q. \end{cases}$$

Proof. Let $\lambda_0 > (\rho_f\phi_{\min})^{-1}\,\|\theta\|_{\mathrm{L}^\infty(\Omega)}$. We are going to reproduce the T-coercive mapping (3.12) used at the continuous level in the discrete setting. To do so, mimicking (3.13), for all $p_h \in Q_h$, we introduce $v_{p_h} \in [\mathrm{H}_0^1(\Omega)]^d$ such that

$$\mathrm{div}\,v_{p_h} = p_h \quad \text{and} \quad \|\nabla v_{p_h}\| \le C_{\mathrm{div}}\,\|p_h\|\,. \tag{3.26}$$

Since v_{p_h} does not necessarily belong to the discrete space X_h, we project it on X_h using the Fortin operator Π_h and consider a mapping of the form

$$\mathtt{T}_h : (u_{s,h}, v_{s,h}, v_{f,h}, p_h) \longmapsto \Big(\beta u_{s,h} + \gamma v_{s,h}, \alpha v_{s,h} - \Pi_h v_{p_h}, \alpha v_{f,h} - \Pi_h v_{p_h}, \alpha p_h\Big),$$

where α, β and γ are some constants to be adjusted.

First, we observe that $|||\mathtt{T}_h|||$ is bounded uniformly with respect to h since

$$\big\|\nabla\big(\Pi_h v_{p_h}\big)\big\| \leq C_\pi \, \|\nabla v_{p_h}\| \leq C_\pi C_{\mathrm{div}} \, \|p_h\| \tag{3.27}$$

by virtue of (3.26)-right and (3.20).

Thanks to the divergence-compatibility property of the operator Π_h, the proof then follows the same lines as in the continuous level, see [Barré et al., 2023, Proposition 3.15]. Indeed, we compute

$$\begin{aligned}
\mathcal{A}_{\lambda_0}\big((z_h, p_h), \mathtt{T}_h(z_h, p_h)\big) &= \lambda_0 \int_\Omega \beta\, \sigma_s(u_{s,h}) : \varepsilon(u_{s,h})\,\mathrm{d}x + \lambda_0 \int_\Omega \gamma\, \sigma_s(u_{s,h}) : \varepsilon(v_{s,h})\,\mathrm{d}x \\
&- \int_\Omega \beta\, \sigma_s(v_{s,h}) : \varepsilon(u_{s,h})\,\mathrm{d}x - \int_\Omega \gamma\, \sigma_s(v_{s,h}) : \varepsilon(v_{s,h})\,\mathrm{d}x - \int_\Omega \sigma_s(u_{s,h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x \\
&+ \lambda_0 \int_\Omega \rho_s(1-\phi)\big(\alpha\,|v_{s,h}|^2 - v_{s,h}\cdot \Pi_h v_{p_h}\big)\,\mathrm{d}x + \int_\Omega \alpha\, \sigma_s(u_{s,h}) : \varepsilon(v_{s,h})\,\mathrm{d}x \\
&+ \int_\Omega \alpha\, \phi^2\, k_f^{-1}(v_{f,h} - v_{s,h})\cdot(v_{f,h} - v_{s,h})\,\mathrm{d}x + \int_\Omega \big(\lambda_0 \rho_f \phi - \theta\big)\big(\alpha\,|v_{f,h}|^2 - v_{f,h}\cdot \Pi_h v_{p_h}\big)\,\mathrm{d}x \\
&+ \int_\Omega \phi\,\big(\alpha\, \sigma_f(v_{f,h}) : \varepsilon(v_{f,h}) - \sigma_f(v_{f,h}) : \varepsilon(\Pi_h v_{p_h})\big)\,\mathrm{d}x - \int_\Omega p_h \,\mathrm{div}\,\big((1-\phi)\,\alpha v_{s,h} + \phi\,\alpha v_{f,h}\big)\,\mathrm{d}x \\
&+ \int_\Omega p_h \,\mathrm{div}\,\big((1-\phi)\,\Pi_h v_{p_h} + \phi\,\Pi_h v_{p_h}\big)\,\mathrm{d}x + \int_\Omega \mathrm{div}\,\big((1-\phi)\,v_{s,h} + \phi\, v_{f,h}\big)\alpha p_h\,\mathrm{d}x.
\end{aligned}$$

Note that the term $-\int_\Omega p_h \,\mathrm{div}\,\big((1-\phi)\,\alpha v_{s,h} + \phi\,\alpha v_{f,h}\big)\,\mathrm{d}x$ and $\int_\Omega \mathrm{div}\,\big((1-\phi)\,v_{s,h} + \phi\, v_{f,h}\big)\alpha p_h\,\mathrm{d}x$ cancel out, and that

$$\int_\Omega p_h \,\mathrm{div}\,\big((1-\phi)\,\Pi_h v_{p_h} + \phi\,\Pi_h v_{p_h}\big)\,\mathrm{d}x = \int_\Omega p_h \,\mathrm{div}\,\big(\Pi_h v_{p_h}\big)\,\mathrm{d}x = \int_\Omega p_h \mathrm{div}\, v_{p_h}\,\mathrm{d}x = \int_\Omega p_h^2\,\mathrm{d}x,$$

thanks to (3.19) and (3.26)-left. Now, we set $\beta = \frac{\alpha}{2}$ and $\gamma = -\frac{\alpha}{2\lambda_0}$ in order to remove the terms of the form $\int_\Omega \sigma_s(u_{s,h}) : \varepsilon(v_{s,h})\,\mathrm{d}x$. Consequently, we have

$$\begin{aligned}
&\mathcal{A}_{\lambda_0}\big((z_h, p_h), \mathtt{T}_h(z_h, p_h)\big) \qquad\qquad (3.28)\\
&\qquad \geq \frac{\lambda_0 \alpha}{2} \int_\Omega \sigma_s(u_{s,h}) : \varepsilon(u_{s,h})\,\mathrm{d}x - \int_\Omega \sigma_s(u_{s,h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x \\
&\qquad + \frac{\alpha}{2\lambda_0} \int_\Omega \sigma_s(v_{s,h}) : \varepsilon(v_{s,h})\,\mathrm{d}x + \lambda_0 \rho_s (1-\phi_{\max}) \int_\Omega \big(\alpha\,|v_{s,h}|^2 - v_{s,h}\cdot \Pi_h v_{p_h}\big)\,\mathrm{d}x \\
&\qquad + \big(\lambda_0 \rho_f \phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)}\big) \int_\Omega \big(\alpha\,|v_{f,h}|^2 - v_{f,h}\cdot \Pi_h v_{p_h}\big)\,\mathrm{d}x \\
&\qquad + \phi_{\min} \int_\Omega \big(\alpha\, \sigma_f(v_{f,h}) : \varepsilon(v_{f,h}) - \sigma_f(v_{f,h}) : \varepsilon(\Pi_h v_{p_h})\big)\,\mathrm{d}x + \int_\Omega p_h^2\,\mathrm{d}x. \qquad (3.29)
\end{aligned}$$

Next, for all $\delta > 0$, Young inequality yields

$$\begin{aligned}
-\int_\Omega \sigma_s(u_{s,h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x &\geq -\frac{\delta}{2}\int_\Omega \sigma_s(u_{s,h}) : \varepsilon(u_{s,h})\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega \sigma_s(\Pi_h v_{p_h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x,\\
-\int_\Omega \sigma_f(v_{f,h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x &\geq -\frac{\delta}{2}\int_\Omega \sigma_f(v_{f,h}) : \varepsilon(v_{f,h})\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega \sigma_f(\Pi_h v_{p_h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x,\\
-\int_\Omega v_{s,h}\cdot\Pi_h v_{p_h}\,\mathrm{d}x &\geq -\frac{\delta}{2}\int_\Omega |v_{s,h}|^2\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega |\Pi_h v_{p_h}|^2\,\mathrm{d}x,\\
-\int_\Omega v_{f,h}\cdot\Pi_h v_{p_h}\,\mathrm{d}x &\geq -\frac{\delta}{2}\int_\Omega |v_{f,h}|^2\,\mathrm{d}x - \frac{1}{2\delta}\int_\Omega |\Pi_h v_{p_h}|^2\,\mathrm{d}x.
\end{aligned} \tag{3.30}$$

To control the new terms appearing in (3.30), we use the inequalities

$$\|\mathrm{div}\, v\|^2 \leq d\,\|\nabla v\|^2 \quad \text{and} \quad \|\varepsilon(v)\| \leq \|\nabla v\|, \quad \forall v \in [\mathrm{H}^1_0(\Omega)]^d,$$

together with (3.27) to retrieve

$$\begin{aligned}
\int_\Omega \sigma_f(\Pi_h v_{p_h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x &= \lambda_f \|\mathrm{div}\,(\Pi_h v_{p_h})\|^2 + 2\mu_f \|\varepsilon(\Pi_h v_{p_h})\|^2 \leq C_\pi^2 C_{\mathrm{div}}^2(\lambda_f d + 2\mu_f)\|p_h\|^2,\\
\int_\Omega \sigma_s(\Pi_h v_{p_h}) : \varepsilon(\Pi_h v_{p_h})\,\mathrm{d}x &= \lambda \|\mathrm{div}\,(\Pi_h v_{p_h})\|^2 + 2\mu \|\varepsilon(\Pi_h v_{p_h})\|^2 \leq C_\pi^2 C_{\mathrm{div}}^2(\lambda d + 2\mu)\|p_h\|^2.
\end{aligned} \tag{3.31}$$

Furthermore, denoting by C_p the constant of Poincaré inequality, it holds that

$$\|\Pi_h v_{p_h}\|^2 \leq C_p \|\nabla(\Pi_h v_{p_h})\|^2 \leq C_p C_\pi^2 C_{\mathrm{div}}^2 \|p_h\|^2. \tag{3.32}$$

Using (3.30), (3.31) and (3.32) to bound from below the right-hand side of (3.29) and rearranging terms, we obtain

$$\begin{aligned}
\mathcal{A}_{\lambda_0}\big((z_h,p_h),\mathtt{T}_h(z_h,p_h)\big) \geq{}& \Big(\frac{\lambda_0\alpha}{2} - \frac{\delta}{2}\Big)\|u_{s,h}\|_{\mathrm{s}}^2 + \frac{\alpha}{2\lambda_0}\|v_{s,h}\|_{\mathrm{s}}^2 + \lambda_0\rho_s(1-\phi_{\max})\Big(\alpha-\frac{\delta}{2}\Big)\|v_{s,h}\|^2\\
&+ (\lambda_0\rho_f\phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)})\Big(\alpha-\frac{\delta}{2}\Big)\|v_{f,h}\|^2 + 2\mu_f\phi_{\min}\Big(\alpha-\frac{\delta}{2}\Big)\|\varepsilon(v_{f,h})\|^2 + \Big(1-\frac{\delta^*}{2\delta}\Big)\|p_h\|^2,
\end{aligned}$$

where

$$\delta^* = C_\pi^2 C_{\mathrm{div}}^2\big(\lambda d + 2\mu + \lambda_0\rho_s(1-\phi_{\max})C_p + (\lambda_0\rho_f\phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)})C_p + \phi_{\min}(\lambda_f d + 2\mu_f)\big).$$

Thanks to the assumption $\lambda_0 > (\rho_f\phi_{\min})^{-1}\|\theta\|_{\mathrm{L}^\infty(\Omega)}$, we have $\delta^* > 0$.

Hence, setting $\delta = \delta^*$ and $\alpha = \alpha^* = \max\big(\delta^*, \frac{2\delta^*}{\lambda_0}\big)$, we get

$$\mathcal{A}_{\lambda_0}\big((z_h,p_h),\mathtt{T}_h(z_h,p_h)\big) \geq \frac{\delta^*}{2}\|u_{s,h}\|_{\mathrm{s}}^2 + \frac{\alpha^*}{2\lambda_0}\|v_{s,h}\|_{\mathrm{s}}^2 + \mu_f\phi_{\min}\delta^*\|\varepsilon(v_{f,h})\|^2 + \frac{1}{2}\|p_h\|^2. \tag{3.33}$$

Finally, we infer that $\mathcal{A}_{\lambda_0}$ is $\mathtt{T}_h$-coercive for the mapping

$$\mathtt{T}_h : (u_{s,h}, v_{s,h}, v_{f,h}, p_h) \longmapsto \Big(\frac{\alpha^*}{2}u_{s,h} - \frac{\alpha^*}{2\lambda_0}v_{s,h}, \alpha^* v_{s,h} - \Pi_h v_{p_h}, \alpha^* v_{f,h} - \Pi_h v_{p_h}, \alpha^* p_h\Big).$$

□

Remark 3.11. If θ is small, namely if it satisfies (3.6), then the condition $\lambda_0 > (\rho_f\phi_{\min})^{-1}\|\theta\|_{\mathrm{L}^\infty(\Omega)}$ can be dropped.

Coming back to the weak formulation of the Crank-Nicolson scheme, we get the following well-posedness result.

Theorem 3.12. *Assume that* $(h1)-(h4)$ *hold and that the discrete inf-sup condition* (3.18) *is satisfied. If we have in addition*

$$\Delta t < \frac{2\rho_f \phi_{\min}}{\|\theta\|_{L^\infty(\Omega)}}, \tag{3.34}$$

then Problem (3.23) *is well-posed.*

Proof. Isolating the unknown z_h^{n+1}, the formulation (3.23) is equivalent to

$$\begin{cases} \text{Find } z_h^{n+1} \in V_h \text{ and } p_h^{n+1} \in Q_h \quad \text{such that for all } (y_h, q_h) \in V_h \times Q_h, \\ 2(\Delta t)^{-1}\big(z_h^{n+1}, y_h\big)_H + \mathcal{A}\big((z_h^{n+1}, p_h^{n+1}), (y_h, q_h)\big) \\ \qquad\qquad = 2(\Delta t)^{-1}\big(z_h^{n}, y_h\big)_H - \mathcal{A}\big((z_h^{n}, p_h^{n}), (y_h, q_h)\big) + 2\big(g^{n+\frac{1}{2}}, y_h\big)_H. \end{cases}$$

Here, we see that the bilinear form involved for solving the discrete problem at time t^{n+1} is a perturbation of the bilinear form $\mathcal{A}$. Moreover, the perturbed form is exactly the same than the one studied at the continuous level in Proposition 3.5. Indeed, recalling the notation (3.11), we get the formulation

$$\begin{cases} \text{Find } z_h^{n+1} \in V_h \text{ and } p_h^{n+1} \in Q_h \quad \text{such that for all } (y_h, q_h) \in V_h \times Q_h, \\ \mathcal{A}_{2(\Delta t)^{-1}}\big((z_h^{n+1}, p_h^{n+1}), (y_h, q_h)\big) = 2(\Delta t)^{-1}\big(z_h^{n}, y_h\big)_H - \mathcal{A}\big((z_h^{n}, p_h^{n}), (y_h, q_h)\big) + 2\big(g^{n+\frac{1}{2}}, y_h\big)_H. \end{cases} \tag{3.35}$$

To ensure well-posedness, we know from Proposition 3.7 that it is sufficient to prove that the bilinear form $\mathcal{A}_{2(\Delta t)^{-1}}$ is $\mathtt{T}_h$-coercive. Applying Lemma 3.10, we find that this problem is well-posed provided that $2(\Delta t)^{-1} > (\rho_f \phi_{\min})^{-1} \|\theta\|_{\mathrm{L}^\infty(\Omega)}$, which corresponds exactly to the time step restriction (3.34). □

Remark 3.13. Here, we propose a stable space discretization that is indepedent of the porosity ϕ. This is motivated by the fact that in the original non-linear model derived in [Chapelle and Moireau, 2014], the porosity is an unknown state variable depending on time, so that any space discretization must be robust with respect to this parameter. If we allow the choice of the pair (X_h, Q_h) to depend on ϕ, one has to study the influence of the porosity on the constant $\underline{\beta}$ involved in the discrete inf-sup condition

$$\exists \underline{\beta} > 0, \forall p_h \in Q_h, \qquad \sup_{(v_{s,h}, v_{f,h}) \in X_h \times X_h} \frac{\displaystyle\int_\Omega \operatorname{div}\left((1-\phi)\, v_{s,h} + \phi\, v_{f,h}\right) p_h \,\mathrm{d}x}{\|(v_{s,h}, v_{f,h})\|_{[\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{H}_0^1(\Omega)]^d}} \geq \underline{\beta}\, \|p_h\|,$$

see [Barnafi et al., 2021] for a discussion on this topic.

Remark 3.14. If we chose different discretization spaces for the solid and the fluid, namely if $V_h = X_{s,h} \times X_{s,h} \times X_{f,h}$ with $X_{s,h} \neq X_{f,h}$, the proof of Lemma 3.10 can be extended provided that there exists a Fortin operator $\Pi_h : [\mathrm{H}_0^1(\Omega)]^d \mapsto X_{s,h} \cap X_{f,h}$ verifying (3.19) and (3.20). Hence, the well-posedness of the discrete problem is guaranteed under the inf-sup condition

$$\exists \beta > 0, \forall p_h \in Q_h, \qquad \sup_{v_h \in X_{s,h} \cap X_{f,h}} \frac{\displaystyle\int_\Omega \operatorname{div} v_h\, p_h \,\mathrm{d}x}{\|v_h\|_{[\mathrm{H}_0^1(\Omega)]^d}} \geq \beta\, \|p_h\|.$$

Remark 3.15. Note that the result of Theorem 3.12 does not require any assumption on the size of the permeability tensor k_f, contrary to the assumptions made in [Barnafi et al., 2021] for the compressible case.

Remark 3.16. If θ depends on time, the bilinear form $\mathcal{A}_{\lambda_0}$ also depends on time. Nevertheless, the result of Lemma 3.10 could be extended as long as we assume that $\theta \in C^0\big([0,T] \times \Omega\big)$ and $\lambda_0 > (\rho_f \phi_{\min})^{-1} \|\theta\|_{C^0\left([0,T]\times\Omega\right)}$.

For the backward Euler scheme, we obtain well-posedness under a time step condition that is slightly more restrictive than (3.34).

Theorem 3.17. *Assume that* $(h1)-(h4)$ *hold and that the discrete inf-sup condition* (3.18) *is satisfied. If we have in addition*

$$\Delta t < \frac{\rho_f \phi_{\min}}{\|\theta\|_{L^\infty(\Omega)}}, \tag{3.36}$$

then Problem (3.24) *is well-posed.*

Proof. Let us rewrite (3.24) by isolating the unknown $(u_{s,h}^{n+\frac{1}{2}}, v_{s,h}^{n+\frac{1}{2}}, v_{f,h}^{n+1}, p_h^{n+1})$. Writing $u_s^{n+1}-u_s^n = 2(u_s^{n+\frac{1}{2}} - u_s^n)$ and $v_s^{n+1} - v_s^n = 2(v_s^{n+\frac{1}{2}} - v_s^n)$, we obtain the following discrete problem: find $(u_{s,h}^{n+\frac{1}{2}}, v_{s,h}^{n+\frac{1}{2}}, v_{f,h}^{n+1}, p_h^{n+1}) \in V_h \times Q_h$ such that for all $y_h = (d_{s,h}, w_{s,h}, w_{f,h}) \in V_h$ and $q_h \in Q_h$,

$$\begin{aligned} 2(\Delta t)^{-1} \int_\Omega \sigma_s(u_{s,h}^{n+\frac{1}{2}}) : \varepsilon(d_{s,h})\,\mathrm{d}x + 2(\Delta t)^{-1} \int_\Omega \rho_s(1-\phi) v_{s,h}^{n+\frac{1}{2}} \cdot w_{s,h}\,\mathrm{d}x + (\Delta t)^{-1} \int_\Omega \rho_f \phi v_{f,h}^{n+1} \cdot w_{f,h}\,\mathrm{d}x \\ + \mathcal{A}\big((u_{s,h}^{n+\frac{1}{2}}, v_{s,h}^{n+\frac{1}{2}}, v_{f,h}^{n+1}, p_h^{n+1}), (y_h, q_h)\big) = \ell(y_h), \end{aligned} \tag{3.37}$$

where ℓ is a continuous linear form depending only on the prescribed body force $f^{n+\frac{1}{2}}$ and the solution at time t^n. As for the Crank-Nicolson scheme, the bilinear form appearing in the left-hand side of (3.37) is a perturbation of the bilinear form $\mathcal{A}$. However, this perturbation does not exactly correspond to the scalar product $(\cdot,\cdot)_H$ because the coefficients in front of the solid and fluid terms – namely $2(\Delta t)^{-1}$ and $(\Delta t)^{-1}$ – are different, so that we cannot directly apply Lemma 3.10.

Nevertheless, we can reproduce its proof with this modified perturbation, which amounts to replacing (λ, μ, ρ_s) by $(2\lambda, 2\mu, 2\rho_s)$ and adapting the choice of the constants α, β and γ. But the restriction on the parameter λ_0 comes from the coefficient $\big(\lambda_0 \rho_f \phi_{\min} - \|\theta\|_{\mathrm{L}^\infty(\Omega)}\big)$ arising in front of the fluid term, which is not affected by this modification. Therefore, we conclude that the discrete problem is well-posed for $(\Delta t)^{-1} > (\rho_f \phi_{\min})^{-1} \|\theta\|_{\mathrm{L}^\infty(\Omega)}$, which corresponds to (3.36). □

Remark 3.18. If θ satisfies the smallness condition (3.6), then the assumptions made on the time step in Theorems 3.12 and 3.17 are not necessary. Indeed, if (3.6) holds true, then the discrete problems (3.23) and (3.24) are well-posed irrespectively of the time step Δt.

3.2.3 Discrete energy balances

The two schemes (3.15) and (3.16) satisfy fundamental energy balances at the discrete level. As a matter of fact, choosing $(d_{s,h}, w_{s,h}, w_{f,h}, q_h) = (\Delta t\, u_{s,h}^{n+\frac{1}{2}}, \Delta t\, v_{s,h}^{n+\frac{1}{2}}, \Delta t\, v_{f,h}^{n+\frac{1}{2}}, \Delta t\, p_h^{n+\frac{1}{2}})$ in (3.21), we obtain

$$\begin{aligned} &\int_\Omega \sigma_s(u_{s,h}^{n+1} - u_{s,h}^n) : \varepsilon(u_{s,h}^{n+\frac{1}{2}})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,(v_{s,h}^{n+1} - v_{s,h}^n) \cdot v_{s,h}^{n+\frac{1}{2}}\,\mathrm{d}x + \int_\Omega \rho_f \phi\,(v_{f,h}^{n+1} - v_{f,h}^n) \cdot v_{f,h}^{n+\frac{1}{2}}\,\mathrm{d}x \\ &\qquad - \Delta t \int_\Omega \sigma_s(v_{s,h}^{n+\frac{1}{2}}) : \varepsilon(u_{s,h}^{n+\frac{1}{2}})\,\mathrm{d}x + \Delta t \int_\Omega \sigma_s(u_{s,h}^{n+\frac{1}{2}}) : \varepsilon(v_{s,h}^{n+\frac{1}{2}})\,\mathrm{d}x \\ &+ \Delta t \int_\Omega \phi\, \sigma_f(v_{f,h}^{n+\frac{1}{2}}) : \varepsilon(v_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x - \Delta t \int_\Omega \theta v_{f,h}^{n+\frac{1}{2}} \cdot v_{f,h}^{n+\frac{1}{2}}\,\mathrm{d}x + \Delta t \int_\Omega \phi^2\, k_f^{-1} (v_{f,h}^{n+\frac{1}{2}} - v_{s,h}^{n+\frac{1}{2}})^2\,\mathrm{d}x \\ &- \Delta t \int_\Omega p_h^{n+\frac{1}{2}} \operatorname{div}\big((1-\phi)\, v_{s,h}^{n+\frac{1}{2}} + \phi\, v_{f,h}^{n+\frac{1}{2}}\big)\,\mathrm{d}x + \Delta t \int_\Omega \operatorname{div}\big((1-\phi)\, v_{s,h}^{n+\frac{1}{2}} + \phi\, v_{f,h}^{n+\frac{1}{2}}\big)\, p_h^{n+\frac{1}{2}}\,\mathrm{d}x \\ &\qquad = \Delta t \int_\Omega \rho_s(1-\phi)\, f^{n+\frac{1}{2}} \cdot v_{s,h}^{n+\frac{1}{2}}\,\mathrm{d}x + \Delta t \int_\Omega \rho_f \phi\, f^{n+\frac{1}{2}} \cdot v_{f,h}^{n+\frac{1}{2}}\,\mathrm{d}x, \end{aligned}$$

where $\phi^2 k_f^{-1}\Big(v_{f,h}^{n+\frac{1}{2}} - v_{s,h}^{n+\frac{1}{2}}\Big)^2$ is a shortcut notation for $\phi^2 k_f^{-1}\big(v_{f,h}^{n+\frac{1}{2}} - v_{s,h}^{n+\frac{1}{2}}\big)\cdot\big(v_{f,h}^{n+\frac{1}{2}} - v_{s,h}^{n+\frac{1}{2}}\big)$. Therefore, using that $(v^{n+1}-v^n)\cdot v^{n+\frac{1}{2}} = \frac{1}{2}\big(|v^{n+1}|^2 - |v^n|^2\big)$ and introducing the discrete energy

$$\mathcal{E}_h^n = \underbrace{\frac{1}{2}\int_\Omega \sigma_s(u_{s,h}^n) : \varepsilon(u_{s,h}^n)\,\mathrm{d}x}_{\text{Structure discrete mechanical energy}} + \underbrace{\frac{1}{2}\int_\Omega \rho_s(1-\phi)\,\big|v_{s,h}^n\big|^2\,\mathrm{d}x}_{\text{Structure discrete kinetic energy}} + \underbrace{\frac{1}{2}\int_\Omega \rho_f\phi\,\big|v_{f,h}^n\big|^2\,\mathrm{d}x}_{\text{Fluid discrete kinetic energy}},$$

we find

$$\begin{aligned}(\mathcal{E}_h^{n+1} - \mathcal{E}_h^n) + \Delta t\int_\Omega \phi\,\sigma_f(v_{f,h}^{n+\frac{1}{2}}) : \varepsilon(v_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(v_{f,h}^{n+\frac{1}{2}} - v_{s,h}^{n+\frac{1}{2}}\Big)^2\,\mathrm{d}x \\ = \Delta t\left(\int_\Omega \theta\left|v_{f,h}^{n+\frac{1}{2}}\right|^2\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,f^{n+\frac{1}{2}}\cdot v_{s,h}^{n+\frac{1}{2}}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,f^{n+\frac{1}{2}}\cdot v_{f,h}^{n+\frac{1}{2}}\,\mathrm{d}x\right),\end{aligned} \tag{3.38}$$

which corresponds to the discrete counterpart of the energy balance (3.4). Note that in absence of external forces and if the mass input term θ is negative, namely if this term *removes* fluid mass from the system, (3.38) directly implies the stability of the system since we then have

$$\mathcal{E}_h^{n+1} + \Delta t\int_\Omega \phi\,\sigma_f(v_{f,h}^{n+\frac{1}{2}}) : \varepsilon(v_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(v_{f,h}^{n+\frac{1}{2}} - v_{s,h}^{n+\frac{1}{2}}\Big)^2\,\mathrm{d}x + \Delta t\int_\Omega |\theta|\left|v_{f,h}^{n+\frac{1}{2}}\right|^2\,\mathrm{d}x \le \mathcal{E}_h^n.$$

The general case requires an application of a discrete version of Grönwall Lemma, as it will be detailed in the next section.

Proceeding similarly for the backward Euler scheme, namely taking $(d_{s,h}, w_{s,h}, w_{f,h}, q_h) = (\Delta t\, u_{s,h}^{n+\frac{1}{2}}, \Delta t\, v_{s,h}^{n+\frac{1}{2}}, \Delta t\, v_{f,h}^{n+1}, \Delta t\, p_h^{n+1})$ in (3.22) and using the identity

$$(v^{n+1}-v^n)\cdot v^{n+1} = \frac{1}{2}\Big(|v^{n+1}|^2 - |v^n|^2 + |v^{n+1}-v^n|^2\Big),$$

we get the discrete energy balance

$$\begin{aligned}(\mathcal{E}_h^{n+1} - \mathcal{E}_h^n) + \frac{1}{2}\int_\Omega \rho_f\phi\left|v_{f,h}^{n+1} - v_{f,h}^n\right|^2\,\mathrm{d}x \\ + \Delta t\int_\Omega \phi\,\sigma_f(v_{f,h}^{n+1}) : \varepsilon(v_{f,h}^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(v_{f,h}^{n+1} - v_{s,h}^{n+\frac{1}{2}}\Big)^2\,\mathrm{d}x \\ = \Delta t\left(\int_\Omega \theta\left|v_{f,h}^{n+1}\right|^2\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,f^{n+\frac{1}{2}}\cdot v_{s,h}^{n+\frac{1}{2}}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,f^{n+\frac{1}{2}}\cdot v_{f,h}^{n+1}\,\mathrm{d}x\right).\end{aligned} \tag{3.39}$$

This is almost the same energy balance as for the Crank-Nicolson scheme, the principal difference being the presence of an additional fluid term $\frac{1}{2}\int_\Omega \rho_f\phi\left|v_{f,h}^{n+1} - v_{f,h}^n\right|^2\,\mathrm{d}x$ inducing numerical dissipation.

3.3 Convergence analysis

The goal of this section is to compare the solution of the continuous problem to the solution of the fully-discrete schemes (3.21) or (3.22). To do so, we are first going to build a projector from the continuous to the discrete space that is adapated to the bilinear form appearing in our problem.

3.3.1 Choosing the finite element spaces

Let us assume that the discrete inf-sup condition (3.18) is fulfilled, and choose a parameter $\lambda_0 > (\rho_f \phi_{\min})^{-1} \|\theta\|_{L^\infty(\Omega)}$. Then, Lemma 3.10 implies that for any $(z,p) \in V \times Q$, there exists a unique $P_h(z,p) \in V_h \times Q_h$ such that

$$\mathcal{A}_{\lambda_0}\big(P_h(z,p),(y_h,q_h)\big) = \mathcal{A}_{\lambda_0}\big((z,p),(y_h,q_h)\big), \quad \forall (y_h,q_h) \in V_h \times Q_h. \tag{3.40}$$

This defines a projector P_h from $[\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times \mathrm{L}^2_0(\Omega)$ to $X_h \times X_h \times X_h \times Q_h$. The projector P_h can be seen as a Riesz projector [Wheeler, 1973; Baker, 1976] adapted to the hyperbolic – parabolic structure of the problem, shifted by the additional mass term with a scaling of λ_0. Let us denote by P_h^u, P_h^s, P_h^f and P_h^p the solid displacement, solid velocity, fluid velocity and pressure components of P_h. The four corresponding projectors act on an element of $[\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times [\mathrm{H}^1_0(\Omega)]^d \times \mathrm{L}^2_0(\Omega)$, but when $z = (u_s, v_s, v_f)$ we will use the notation

$$P_h(z,p) = \big(P_h^u u_s, P_h^s v_s, P_h^f v_f, P_h^p p\big).$$

Similarly, we will condense the three vectorial components of P_h in an operator P_h^z and make the abuse of notation $P_h(z,p) = (P_h^z z, P_h^p p)$, so that (3.40) is equivalent to

$$\mathcal{A}\big(P_h(z,p),(y_h,q_h)\big) + \lambda_0 (P_h^z z, y_h)_H = \mathcal{A}\big((z,p),(y_h,q_h)\big) + \lambda_0 (z,y_h)_H, \quad \forall (y_h,q_h) \in V_h \times Q_h. \tag{3.41}$$

Moreover, in view of property (3.25), it holds

$$\|(z,p) - P_h(z,p)\|_{V\times Q} \leq C \inf_{(y_h,q_h)\in V_h \times q_h} \|(z,p) - (y_h,q_h)\|_{V_h \times Q_h}, \tag{3.42}$$

with $C > 0$ a constant independent of h. If z and p are regular enough, the right-hand side of the previous estimate behaves as a power of the mesh size h. More precisely, denoting by $H^{\ell+1}(\Omega)$ the space $[H^{\ell+1}(\Omega)]^d \times [H^{\ell+1}(\Omega)]^d \times [H^{\ell+1}(\Omega)]^d$, we have

$$\inf_{y_h \in V_h} \|z - y_h\|_V \leq C h^\ell \|z\|_{H^{\ell+1}(\Omega)}, \quad \forall z \in H^{\ell+1}(\Omega) \cap V,$$

and

$$\inf_{q_h \in Q_h} \|p - q_h\| \leq C h^r \|p\|_{H^r(\Omega)}, \quad \forall p \in H^r(\Omega) \cap Q,$$

where the convergence orders ℓ and $r \leq \ell$ depend on the choice of X_h and Q_h. For instance, if (X_h, Q_h) correspond to the so-called Taylor-Hood elements, then $\ell = r = 2$.

Since $\|z - P_h^z z\|_H \leq C \|z - P_h^z z\|_V$ owing to Korn inequality (3.5), we deduce that

$$\|z - P_h^z z\|_H \leq C\big(h^\ell \|z\|_{H^{\ell+1}(\Omega)} + h^r \|p\|_{H^r(\Omega)}\big), \tag{3.43}$$

and

$$\big\|p - P_h^p p\big\| \leq C\big(h^\ell \|z\|_{H^{\ell+1}(\Omega)} + h^r \|p\|_{H^r(\Omega)}\big). \tag{3.44}$$

These two estimates will play a central role to control the space consistency terms arising in the error analysis.

3.3.2 Error analysis for the Crank-Nicolson scheme

We recall that the continuous solution $(z,p) = (u_s, v_s, v_f, p)$ from (3.10) satisfies

$$\big(\dot{z}(t), y\big)_H + \mathcal{A}\big((z(t),p(t)),(y,q)\big) = (g(t),y)_H, \quad \forall y \in V, \forall q \in Q.$$

In particular, since we consider *conforming* finite element approximations $V_h \subset V$ and $Q_h \subset Q$, we have

$$\big(\dot{z}(t), y_h\big)_H + \mathcal{A}\big((z(t), p(t)), (y_h, q_h)\big) = (g(t), y_h)_H, \quad \forall y_h \in V_h, \forall q_h \in Q_h. \tag{3.45}$$

In what follows, we assume that (z, p) is regular enough. In order to quantify the convergence of the fully discrete solution towards the solution of the continuous problem above, for k an integer or a half-integer, we introduce the error $\epsilon_h^k = z(t^k) - z_h^k = (\epsilon_{u,h}^k, \epsilon_{s,h}^k, \epsilon_{f,h}^k)$ with

$$\begin{aligned}
\epsilon_{u,h}^k &= u_s(t^k) - u_{s,h}^k, \\
\epsilon_{s,h}^k &= v_s(t^k) - v_{s,h}^k, \\
\epsilon_{f,h}^k &= v_f(t^k) - v_{f,h}^k.
\end{aligned}$$

We are now ready to state the following error estimate for the Crank-Nicolson scheme, which is the main result of this paper.

Theorem 3.19. *Assume that $(h1) - (h5)$ hold, and that the solution of the continuous problem (3.10) has the additional regularity*

$$\begin{aligned}
(u_s, v_s, v_f) &\in C^1\big([0,T]; H^{\ell+1}(\Omega)\big), \qquad & p &\in C^1\big([0,T]; H^r(\Omega)\big), \\
\big(\partial_{tt}^2 u_s, \partial_{tt}^2 v_s, \partial_{tt}^2 v_f\big) &\in C^1\big([0,T]; H\big), & \partial_{tt}^2 v_f &\in L^2\big(0,T; [\mathrm{H}_0^1(\Omega)]^d\big).
\end{aligned} \tag{3.46}$$

If we have in addition

$$\Delta t < \frac{\rho_f \phi_{\min}}{4 \left\|\theta\right\|_{L^\infty(\Omega)}}, \tag{3.47}$$

then for all $0 \le N \le n_T$, it holds that

$$\begin{aligned}
\frac{1}{2}\int_\Omega \sigma_s(\epsilon_{u,h}^N) : \varepsilon(\epsilon_{u,h}^N)\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\,\left|\epsilon_{s,h}^N\right|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,\left|\epsilon_{f,h}^N\right|^2\,\mathrm{d}x \\
+ \Delta t \sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\epsilon_{f,h}^{n+\frac{1}{2}}) : \varepsilon(\epsilon_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + \Delta t \sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\big(\epsilon_{f,h}^{n+\frac{1}{2}} - \epsilon_{s,h}^{n+\frac{1}{2}}\big)^2\,\mathrm{d}x \\
\le C \exp\left(\frac{4(\rho_f\phi_{\min})^{-1}\left\|\theta\right\|_{L^\infty(\Omega)} T}{1 - 4\Delta t(\rho_f\phi_{\min})^{-1}\left\|\theta\right\|_{L^\infty(\Omega)}}\right)(\Delta t^2 + h^\ell + h^r)^2,
\end{aligned} \tag{3.48}$$

with C a constant independent of h and Δt.

Proof. The proof is divided into four steps. First, we derive a suitable error equation by using the definition of the specific projector P_h introduced earlier and by gathering time and space consistency terms. Then, after exploiting the stability of the scheme, these terms are estimated and the conclusion is obtained by an application of a discrete version of Grönwall Lemma.

Step 1: derivation of the error equation. First, we want to inject the continuous solution into the semi-discretized in time scheme (3.15), namely compute

$$\left(\frac{z(t^{n+1}) - z(t^n)}{\Delta t}, y_h\right)_H + \mathcal{A}\left(\left(\frac{z(t^{n+1}) + z(t^n)}{2}, \frac{p(t^{n+1}) + p(t^n)}{2}\right), (y_h, q_h)\right),$$

for all $(y_h, q_h) \in V_h \times Q_h$. To do so, we observe that averaging (3.45) at times t^n and t^{n+1} leads to

$$\left(\frac{\dot{z}(t^{n+1}) + \dot{z}(t^n)}{2}, y_h\right)_H + \mathcal{A}\left(\left(\frac{z(t^{n+1}) + z(t^n)}{2}, \frac{p(t^{n+1}) + p(t^n)}{2}\right), (y_h, q_h)\right) = \left(\frac{g(t^{n+1}) + g(t^n)}{2}, y_h\right)_H.$$

Therefore, writing

$$\frac{\dot{z}(t^{n+1})+\dot{z}(t^n)}{2}=\frac{\dot{z}(t^{n+1})+\dot{z}(t^n)}{2}-\frac{z(t^{n+1})-z(t^n)}{\Delta t}+\frac{z(t^{n+1})-z(t^n)}{\Delta t},$$

we obtain

$$\left(\frac{z(t^{n+1})-z(t^n)}{\Delta t},y_h\right)_H+\mathcal{A}\left(\left(\frac{z(t^{n+1})+z(t^n)}{2},\frac{p(t^{n+1})+p(t^n)}{2}\right),(y_h,q_h)\right)$$
$$=\left(\frac{g(t^{n+1})+g(t^n)}{2},y_h\right)_H+\left(\mathcal{R}^{n+\frac{1}{2}},y_h\right)_H, \quad (3.49)$$

where $\mathcal{R}^{n+\frac{1}{2}}$ gathers the time consistency error defined by

$$\mathcal{R}^{n+\frac{1}{2}}=\frac{z(t^{n+1})-z(t^n)}{\Delta t}-\frac{\dot{z}(t^{n+1})+\dot{z}(t^n)}{2},$$

which will hereafter be controled using a Taylor expansion.

Next, for a given $\lambda_0>(\rho_f\phi_{\min})^{-1}\left\|\theta\right\|_{L^\infty(\Omega)}$, we are going to approximate the continuous solution by means of the discrete projector P_h defined in (3.41). We recall that P_h satisfies, for any $(z,p)\in V\times Q$,

$$\mathcal{A}\big((z,p),(y_h,q_h)\big)=\mathcal{A}\big(P_h(z,p),(y_h,q_h)\big)+\lambda_0\big(P_h^z z-z,y_h\big)_H,\quad \forall y_h\in V_h,\forall q_h\times Q_h.$$

Averaging this relation for the choices $(z,p)=\big(z(t^n),p(t^n)\big)$ and $(z,p)=\big(z(t^{n+1}),p(t^{n+1})\big)$, we get

$$\mathcal{A}\left(\left(\frac{z(t^{n+1})+z(t^n)}{2},\frac{p(t^{n+1})+p(t^n)}{2}\right),(y_h,q_h)\right)$$
$$=\mathcal{A}\left(\left(\frac{P_h^z z(t^{n+1})+P_h^z z(t^n)}{2},\frac{P_h^p p(t^{n+1})+P_h^p p(t^n)}{2}\right),(y_h,q_h)\right)$$
$$+\lambda_0\left(\frac{P_h^z z(t^{n+1})+P_h^z z(t^n)}{2}-\frac{z(t^{n+1})+z(t^n)}{2},y_h\right)_H.$$

Plugging this result into (3.49), it follows that

$$\left(\frac{z(t^{n+1})-z(t^n)}{\Delta t},y_h\right)_H+\mathcal{A}\left(\left(\frac{P_h^z z(t^{n+1})+P_h^z z(t^n)}{2},\frac{P_h^p p(t^{n+1})+P_h^p p(t^n)}{2}\right),(y_h,q_h)\right)$$
$$=\left(\frac{g(t^{n+1})+g(t^n)}{2},y_h\right)_H+\left(\mathcal{R}^{n+\frac{1}{2}},y_h\right)_H+\lambda_0\left(\mathcal{S}_h^{n+\frac{1}{2}},y_h\right)_H, \quad (3.50)$$

where $\mathcal{S}_h^{n+\frac{1}{2}}$ is a space consistency term given by

$$\mathcal{S}_h^{n+\frac{1}{2}}=\frac{z(t^{n+1})+z(t^n)}{2}-\frac{P_h^z z(t^{n+1})+P_h^z z(t^n)}{2},$$

that will further be estimated using the approximability properties of the operator P_h. Decomposing the first term of (3.50) as

$$\frac{z(t^{n+1})-z(t^n)}{\Delta t}=\frac{z(t^{n+1})-z(t^n)}{\Delta t}-P_h^z\left(\frac{z(t^{n+1})-z(t^n)}{\Delta t}\right)+P_h^z\left(\frac{z(t^{n+1})-z(t^n)}{\Delta t}\right),$$

and using the linearity of P_h, we end up with

$$\left(\frac{P_h^z z(t^{n+1}) - P_h^z z(t^n)}{\Delta t}, y_h\right)_H + \mathcal{A}\left(\left(\frac{P_h^z z(t^{n+1}) + P_h^z z(t^n)}{2}, \frac{P_h^p p(t^{n+1}) + P_h^p p(t^n)}{2}\right), (y_h, q_h)\right)$$
$$= \left(\frac{g(t^{n+1}) + g(t^n)}{2}, y_h\right)_H + \left(\mathcal{R}^{n+\frac{1}{2}}, y_h\right)_H + \lambda_0\left(\mathcal{S}_h^{n+\frac{1}{2}}, y_h\right)_H + \left(\mathcal{T}_h^{n+\frac{1}{2}}, y_h\right)_H, \quad (3.51)$$

where

$$\mathcal{T}_h^{n+\frac{1}{2}} = P_h^z\left(\frac{z(t^{n+1}) - z(t^n)}{\Delta t}\right) - \frac{z(t^{n+1}) - z(t^n)}{\Delta t},$$

is another space consistency term coming from the spatial approximation of the discrete derivative of the solution.

Now, let us denote by (e_h^n, δ_h^n) the error between the projection of the continuous solution and the discrete solution at time t^n, namely $e_h^n = P_h^z z(t^n) - z_h^n = (e_{u,h}^n, e_{s,h}^n, e_{f,h}^n)$ with

$$e_{u,h}^n = P_h^u u_s(t^n) - u_{s,h}^n,$$
$$e_{s,h}^n = P_h^s v_s(t^n) - v_{s,h}^n,$$
$$e_{f,h}^n = P_h^f v_f(t^n) - v_{f,h}^n,$$

and

$$\delta_h^n = P_h^p p(t^n) - p_h^n.$$

From (3.23), we know that the fully-discrete solution $(z_h^{n+\frac{1}{2}}, p_h^{n+\frac{1}{2}}) = (u_{s,h}^{n+\frac{1}{2}}, v_{s,h}^{n+\frac{1}{2}}, v_{f,h}^{n+\frac{1}{2}}, p_h^{n+\frac{1}{2}})$ satisfies

$$\left(\frac{z_h^{n+1} - z_h^n}{\Delta t}, y_h\right)_H + \mathcal{A}\left((z_h^{n+\frac{1}{2}}, p_h^{n+\frac{1}{2}}), (y_h, q_h)\right) = \left(\frac{g(t^n) + g(t^{n+1})}{2}, y_h\right)_H, \quad \forall y_h \in V_h, \forall q_h \in Q_h. \quad (3.52)$$

Substracting (3.52) from (3.51), we obtain

$$\left(\frac{e_h^{n+1} - e_h^n}{\Delta t}, y_h\right)_H + \mathcal{A}\left((e_h^{n+\frac{1}{2}}, \delta_h^{n+\frac{1}{2}}), (y_h, q_h)\right)$$
$$= \left(\mathcal{R}^{n+\frac{1}{2}}, y_h\right)_H + \lambda_0\left(\mathcal{S}_h^{n+\frac{1}{2}}, y_h\right)_H + \left(\mathcal{T}_h^{n+\frac{1}{2}}, y_h\right)_H, \quad \forall y_h \in V_h, \forall q_h \in Q_h. \quad (3.53)$$

where we have adopted the notation

$$e_h^{n+\frac{1}{2}} = \frac{P_h^z z(t^{n+1}) + P_h^z z(t^n)}{2} - z_h^{n+\frac{1}{2}} = \frac{e_h^{n+1} + e_h^n}{2},$$

and

$$\delta_h^{n+\frac{1}{2}} = \frac{P_h^p p(t^{n+1}) + P_h^p p(t^n)}{2} - p_h^{n+\frac{1}{2}} = \frac{\delta_h^{n+1} + \delta_h^n}{2}.$$

Step 2: stability estimate in the discrete energy norm. Choosing $(y_h, q_h) = (e_h^{n+\frac{1}{2}}, \delta_h^{n+\frac{1}{2}})$ as test function in (3.53) yields

$$\left(\frac{e_h^{n+1} - e_h^n}{\Delta t}, e_h^{n+\frac{1}{2}}\right)_H + \mathcal{A}\left((e_h^{n+\frac{1}{2}}, \delta_h^{n+\frac{1}{2}}), (e_h^{n+\frac{1}{2}}, \delta_h^{n+\frac{1}{2}})\right)$$
$$= \left(\mathcal{R}^{n+\frac{1}{2}}, e_h^{n+\frac{1}{2}}\right)_H + \lambda_0\left(\mathcal{S}_h^{n+\frac{1}{2}}, e_h^{n+\frac{1}{2}}\right)_H + \left(\mathcal{T}_h^{n+\frac{1}{2}}, e_h^{n+\frac{1}{2}}\right)_H.$$

With the stability identity (3.38), this implies

$$\frac{1}{2}\left\|e_h^{n+1}\right\|_H^2 - \frac{1}{2}\left\|e_h^{n}\right\|_H^2 + \Delta t\int_\Omega \phi\,\sigma_f(e_{f,h}^{n+\frac12}) : \varepsilon(e_{f,h}^{n+\frac12})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\big(e_{f,h}^{n+\frac12} - e_{s,h}^{n+\frac12}\big)^2\,\mathrm{d}x$$
$$= \Delta t\int_\Omega \theta\left|e_{f,h}^{n+\frac12}\right|^2\,\mathrm{d}x + \Delta t\big(\mathcal{R}^{n+\frac12} + \lambda_0\mathcal{S}_h^{n+\frac12} + \mathcal{T}_h^{n+\frac12}, e_h^{n+\frac12}\big)_H.$$

Applying Young inequality $ab \le \frac{\xi}{2}a^2 + \frac{1}{2\xi}b^2$ for a generic parameter $\xi > 0$, we get

$$\frac{1}{2}\left\|e_h^{n+1}\right\|_H^2 - \frac{1}{2}\left\|e_h^{n}\right\|_H^2 + \Delta t\int_\Omega \phi\,\sigma_f(e_{f,h}^{n+\frac12}) : \varepsilon(e_{f,h}^{n+\frac12})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\big(e_{f,h}^{n+\frac12} - e_{s,h}^{n+\frac12}\big)^2\,\mathrm{d}x$$
$$\le \frac{\Delta t\,\|\theta\|_{L^\infty(\Omega)}}{\rho_f\phi_{\min}}\left\|e_h^{n+\frac12}\right\|_H^2 + \frac{\xi\Delta t}{2}\left\|\mathcal{R}^{n+\frac12} + \lambda_0\mathcal{S}_h^{n+\frac12} + \mathcal{T}_h^{n+\frac12}\right\|_H^2 + \frac{\Delta t}{2\xi}\left\|e_h^{n+\frac12}\right\|_H^2. \quad (3.54)$$

Step 3: estimation of the consistency terms. Let us now estimate the consistency terms $\mathcal{R}^{n+\frac12}$, $\mathcal{S}^{n+\frac12}$ and $\mathcal{T}^{n+\frac12}$ appearing in the right-hand side of the previous inequality.

The time consistency error $\mathcal{R}^{n+\frac12}$ is controled using a Taylor expansion. As a matter of fact, we easily verify that

$$\left\|\mathcal{R}^{n+\frac12}\right\|_H = \left\|\frac{z(t^{n+1}) - z(t^n)}{\Delta t} - \frac{\dot z(t^{n+1}) + \dot z(t^n)}{2}\right\|_H \le C\Delta t^2\,\|z\|_{C^3([0,T];H)}\,. \quad (3.55)$$

The space consistency term $\mathcal{S}_h^{n+\frac12}$ and $\mathcal{T}_h^{n+\frac12}$ can be handled with the approximability property (3.42) under suitable regularity assumptions. Indeed, as $z \in C^0\big([0,T];H^{\ell+1}(\Omega)\big)$ and $p \in C^0\big([0,T];H^r(\Omega)\big)$, we infer from (3.43) that

$$\left\|\mathcal{S}_h^{n+\frac12}\right\|_H = \left\|\frac{z(t^{n+1}) + z(t^n)}{2} - P_h^z\left(\frac{z(t^{n+1}) + z(t^n)}{2}\right)\right\|_H \le C(h^\ell + h^r). \quad (3.56)$$

For the second term $\mathcal{T}_h^{n+\frac12}$, we observe that

$$\left\|\mathcal{T}_h^{n+\frac12}\right\|_H = \left\|\frac{z(t^{n+1}) - z(t^n)}{\Delta t} - P_h^z\left(\frac{z(t^{n+1}) - z(t^n)}{\Delta t}\right)\right\|_H$$
$$\le C\left(h^\ell\left\|\frac{z(t^{n+1}) - z(t^n)}{\Delta t}\right\|_{H^{\ell+1}(\Omega)} + h^r\left\|\frac{p(t^{n+1}) - p(t^n)}{\Delta t}\right\|_{H^r(\Omega)}\right)$$
$$= C\left(h^\ell\left\|\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\dot z(t)\,\mathrm{d}t\right\|_{H^{\ell+1}(\Omega)} + h^r\left\|\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\dot p(t)\,\mathrm{d}t\right\|_{H^r(\Omega)}\right),$$

and hence

$$\left\|\mathcal{T}_h^{n+\frac12}\right\|_H \le C(h^\ell + h^r), \quad (3.57)$$

by virtue of (3.46), with C a constant depending on z and p.

Step 4: final error analysis. Putting the consistency errors (3.55), (3.56) and (3.57) together with (3.54), we deduce

$$\frac{1}{2}\left\|e_h^{n+1}\right\|_H^2 - \frac{1}{2}\left\|e_h^{n}\right\|_H^2 + \Delta t\int_\Omega \phi\,\sigma_f(e_{f,h}^{n+\frac12}) : \varepsilon(e_{f,h}^{n+\frac12})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\big(e_{f,h}^{n+\frac12} - e_{s,h}^{n+\frac12}\big)^2\,\mathrm{d}x$$
$$\le \Delta t\left(\frac{\|\theta\|_{L^\infty(\Omega)}}{\rho_f\phi_{\min}} + \frac{1}{2\xi}\right)\left\|e_h^{n+\frac12}\right\|_H^2 + \frac{\xi\Delta t}{2}C(\Delta t^2 + h^\ell + h^r)^2.$$

Multiplying by two and choosing for instance $\xi = \frac{\rho_f \phi_{\min}}{2\|\theta\|_{L^\infty(\Omega)}}$, we get

$$\left\|e_h^{n+1}\right\|_H^2 - \|e_h^n\|_H^2 + 2\Delta t \int_\Omega \phi\, \sigma_f(e_{f,h}^{n+\frac{1}{2}}) : \varepsilon(e_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + 2\Delta t \int_\Omega \phi^2 k_f^{-1}\left(e_{f,h}^{n+\frac{1}{2}} - e_{s,h}^{n+\frac{1}{2}}\right)^2 \mathrm{d}x \\ \leq \frac{4\Delta t\, \|\theta\|_{L^\infty(\Omega)}}{\rho_f \phi_{\min}} \left\|e_h^{n+\frac{1}{2}}\right\|_H^2 + C(\Delta t^2 + h^\ell + h^r)^2,$$

where the constant C now also depends on θ.

Let $N \leq n_T$ be an arbitrary integer. Summing from 0 to $N-1$ and noting that $N\Delta t \leq T$ yields

$$\left\|e_h^N\right\|_H^2 + 2\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f(e_{f,h}^{n+\frac{1}{2}}) : \varepsilon(e_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + 2\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi^2 k_f^{-1}\left(e_{f,h}^{n+\frac{1}{2}} - e_{s,h}^{n+\frac{1}{2}}\right)^2 \mathrm{d}x \\ \leq \left\|e_h^0\right\|_H^2 + \frac{4\Delta t\, \|\theta\|_{L^\infty(\Omega)}}{\rho_f \phi_{\min}} \sum_{n=0}^{N-1} \left\|e_h^{n+\frac{1}{2}}\right\|_H^2 + C(\Delta t^2 + h^\ell + h^r)^2,$$

with C another constant, which also depends on T. Thanks to the chosen initial conditions we have

$$\left\|e_h^0\right\|_H = \|P_h^z z(0) - I_h z(0)\|_H \leq \|P_h^z z(0) - z(0)\|_H + \|z(0) - I_h z(0)\|_H \leq C(h^\ell + h^r).$$

Moreover, since

$$\sum_{n=0}^{N-1} \left\|e_h^{n+\frac{1}{2}}\right\|_H^2 = \sum_{n=0}^{N-1} \left\|\frac{e_h^{n+1} + e_h^n}{2}\right\|_H^2 \leq \frac{1}{2} \sum_{n=0}^{N-1} \left(\left\|e_h^{n+1}\right\|_H^2 + \|e_h^n\|_H^2\right) \leq \sum_{n=0}^{N} \|e_h^n\|_H^2\,,$$

we find

$$\left\|e_h^N\right\|_H^2 + 2\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f(e_{f,h}^{n+\frac{1}{2}}) : \varepsilon(e_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + 2\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi^2 k_f^{-1}\left(e_{f,h}^{n+\frac{1}{2}} - e_{s,h}^{n+\frac{1}{2}}\right)^2 \mathrm{d}x \\ \leq C(\Delta t^2 + h^\ell + h^r)^2 + \frac{4\Delta t\, \|\theta\|_{L^\infty(\Omega)}}{\rho_f \phi_{\min}} \sum_{n=0}^{N} \|e_h^n\|_H^2\,. \quad (3.58)$$

To conclude, we use a discrete version of Grönwall Lemma, recalled below for the sake of completeness. For a proof of this result, we refer the reader to [Heywood and Rannacher, 1990, Lemma 5.1].

Lemma 3.20. *Let $C > 0$ and $\delta > 0$. Let (a_n), (b_n) and (γ_n) be sequences of positive numbers such that*

$$a_N + \delta \sum_{n=0}^{N} b_n \leq C + \delta \sum_{n=0}^{N} \gamma_n a_n.$$

Assume that $\delta\gamma_n < 1$ for all n, and set $\sigma_n = (1 - \delta\gamma_n)^{-1}$. Then, for all $N \geq 0$, it holds that

$$a_N + \delta \sum_{n=0}^{N} b_n \leq C \exp\left(\delta \sum_{n=0}^{N} \sigma_n \gamma_n\right).$$

Let us define $\gamma = \frac{4\|\theta\|_{L^\infty(\Omega)}}{\rho_f \phi_{\min}}$. Recalling (3.47), we have $\gamma\Delta t < 1$. Therefore, Lemma 3.20 implies

that

$$\left\|e_h^N\right\|_H^2 + 2\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f(e_{f,h}^{n+\frac{1}{2}}) : \varepsilon(e_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + 2\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi^2 k_f^{-1} \left(e_{f,h}^{n+\frac{1}{2}} - e_{s,h}^{n+\frac{1}{2}}\right)^2 \mathrm{d}x$$
$$\leq C(\Delta t^2 + h^\ell + h^r)^2 \exp\left((N+1)\Delta t \frac{\gamma}{1-\gamma\Delta t}\right)$$
$$\leq C(\Delta t^2 + h^\ell + h^r)^2 \exp\left(\frac{\gamma T}{1-\gamma\Delta t}\right) \quad \text{since } N\Delta t \leq T.$$

Finally, writing

$$\epsilon_h^N = z(t^N) - P_h^z z(t^N) + P_h^z z(t^N) - z_h^N = z(t^N) - P_h^z z(t^N) + e_h^N,$$

and using (3.43), we obtain

$$\frac{1}{2}\left\|\epsilon_h^N\right\|_H^2 + \Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f(e_{f,h}^{n+\frac{1}{2}}) : \varepsilon(e_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x + \Delta t \sum_{n=0}^{N-1} \int_\Omega \phi^2 k_f^{-1} \left(e_{f,h}^{n+\frac{1}{2}} - e_{s,h}^{n+\frac{1}{2}}\right)^2 \mathrm{d}x$$
$$\leq C(\Delta t^2 + h^\ell + h^r)^2 \exp\left(\frac{\gamma T}{1-\gamma\Delta t}\right).$$

In order to derive (3.48), we rewrite the viscous part as

$$v_f(t^{n+\frac{1}{2}}) - v_{f,h}^{n+\frac{1}{2}} = v_f(t^{n+\frac{1}{2}}) - \frac{v_f(t^{n+1}) + v_f(t^n)}{2}$$
$$+ \frac{v_f(t^{n+1}) + v_f(t^n)}{2} - P_h^f\left(\frac{v_f(t^{n+1}) + v_f(t^n)}{2}\right) + P_h^f\left(\frac{v_f(t^{n+1}) + v_f(t^n)}{2}\right) - v_{f,h}^{n+\frac{1}{2}},$$

namely

$$\epsilon_{f,h}^{n+\frac{1}{2}} = v_f(t^{n+\frac{1}{2}}) - \frac{v_f(t^{n+1}) + v_f(t^n)}{2} + \mathcal{S}_{f,h}^{n+\frac{1}{2}} + e_{f,h}^{n+\frac{1}{2}}.$$

The second term of the above expression is controled thanks to (3.56), and the first one can be estimated as follows

$$\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f\left(v_f(t^{n+\frac{1}{2}}) - \frac{v_f(t^{n+1}) + v_f(t^n)}{2}\right) : \varepsilon\left(v_f(t^{n+\frac{1}{2}}) - \frac{v_f(t^{n+1}) + v_f(t^n)}{2}\right)\mathrm{d}x$$
$$\leq C\Delta t^4 \left\|\partial_{tt}^2 v_f\right\|^2_{L^2(0,T;[\mathrm{H}_0^1(\Omega)]^d)}$$

using a Taylor expansion. □

Remark 3.21. Here, we prove convergence under the time step restriction (3.47), which is slightly more restrictive than the condition found for the well-posedness of the discrete problem, see (3.34). Note however that it may not be optimal.

Remark 3.22. If the smallness condition (3.6) is fulfilled, namely if

$$\frac{C_d \left\|\theta\right\|_{\mathrm{L}^\infty(\Omega)}}{2\mu_f \phi_{\min}} \leq 1,$$

another strategy would be to absorb the additional fluid mass term by the viscous fluid dissipation. Indeed, we then have

$$\Delta t \int_\Omega \theta \left|e_{f,h}^{n+\frac{1}{2}}\right|^2 \mathrm{d}x \leq \Delta t \frac{C_d \left\|\theta\right\|_{\mathrm{L}^\infty(\Omega)}}{2\mu_f \phi_{\min}} \int_\Omega \phi\, \sigma_f(e_{f,h}^{n+\frac{1}{2}}) : \varepsilon(e_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x \leq \Delta t \int_\Omega \phi\, \sigma_f(e_{f,h}^{n+\frac{1}{2}}) : \varepsilon(e_{f,h}^{n+\frac{1}{2}})\,\mathrm{d}x,$$

which indicates that the condition (3.47) may be dropped if the fluid mass input is small enough or if the fluid viscosity is large enough.

3.3.3 Error analysis for the backward Euler scheme

Now, we move to the analysis of the backward Euler scheme, for which we establish a similar result than the one found in [Burtschell et al., 2019] for a compressible material (case $\kappa < +\infty$).

Theorem 3.23. *Assume that $(h1)-(h5)$ hold, and that the solution of the continuous problem (3.10) has the additional regularity*

$$\begin{aligned} (u_s, v_s, v_f) &\in C^1\big([0,T]; H^{\ell+1}(\Omega)\big), \qquad p \in C^1\big([0,T]; H^r(\Omega)\big), \\ \big(\partial_{tt}^2 u_s, \partial_{tt}^2 v_s, \partial_{tt}^2 v_f\big) &\in C^1\big([0,T]; H\big), \qquad \partial_t v_f \in L^2\big(0,T; [\mathrm{H}_0^1(\Omega)]^d\big). \end{aligned}$$

If we have in addition

$$\Delta t < \frac{\rho_f \phi_{\min}}{4\,\|\theta\|_{L^\infty(\Omega)}},$$

then for all $0 \le N \le n_T$, it holds that

$$\begin{aligned} &\frac{1}{2}\int_\Omega \sigma_s(\epsilon_{u,h}^N) : \varepsilon(\epsilon_{u,h}^N)\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\,\left|\epsilon_{s,h}^N\right|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,\left|\epsilon_{f,h}^N\right|^2\,\mathrm{d}x \\ &\qquad + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\epsilon_{f,h}^{n+1}) : \varepsilon(\epsilon_{f,h}^{n+1})\,\mathrm{d}x + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\big(\epsilon_{f,h}^{n+1} - \epsilon_{s,h}^{n+\frac{1}{2}}\big)^2\,\mathrm{d}x \\ &\quad + \frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|\epsilon_{f,h}^{n+1} - \epsilon_{f,h}^{n}\right|^2\,\mathrm{d}x \le C\exp\left(\frac{4(\rho_f\phi_{\min})^{-1}\,\|\theta\|_{L^\infty(\Omega)}\,T}{1 - 4\Delta t(\rho_f\phi_{\min})^{-1}\,\|\theta\|_{L^\infty(\Omega)}}\right)(\Delta t + h^\ell + h^r)^2, \end{aligned} \tag{3.59}$$

with C a constant independent of h and Δt.

Proof. The difficulty of the backward Euler scheme is that it includes a shift between the solid quantities, which are approximated at time $t^{n+\frac{1}{2}}$, and the fluid and pressure quantities, which are approximated at time t^{n+1}. Therefore, we cannot project the continuous solution on the discrete space at the same time as we did for the Crank-Nicolson scheme, since our projector P_h acts simultaneously on solid, fluid and pressure quantities. To overcome this issue, our strategy is to be as close as possible to the analysis performed for the Crank-Nicolson scheme by changing the definitions of the errors to take into account the time shifting rather than handling this shift at the projection level.

To do so, we start the proof from equation (3.50), that reads

$$\begin{aligned} &\left(\frac{z(t^{n+1}) - z(t^n)}{\Delta t}, y_h\right)_H + \mathcal{A}\left(\left(\frac{P_h^z z(t^{n+1}) + P_h^z z(t^n)}{2}, \frac{P_h^p p(t^{n+1}) + P_h^p p(t^n)}{2}\right), (y_h, q_h)\right) \\ &\qquad = \left(\frac{g(t^{n+1}) + g(t^n)}{2}, y_h\right)_H + \big(\mathcal{R}^{n+\frac{1}{2}}, y_h\big)_H + \lambda_0\big(\mathcal{S}_h^{n+\frac{1}{2}}, y_h\big)_H. \end{aligned}$$

From (3.24), the fully-discrete solution $(z_h^{n+1}, p_h^{n+1}) = (u_{s,h}^{n+1}, v_{s,h}^{n+1}, v_{f,h}^{n+1}, p_h^{n+1})$ satisfies

$$\left(\frac{z_h^{n+1} - z_h^n}{\Delta t}, y_h\right)_H + \mathcal{A}\big((u_{s,h}^{n+\frac{1}{2}}, v_{s,h}^{n+\frac{1}{2}}, v_{f,h}^{n+1}, p_h^{n+1}), (y_h, q_h)\big) = \left(\frac{g(t^{n+1}) + g(t^n)}{2}, y_h\right)_H.$$

Substracting these two relations, we obtain

$$\begin{aligned} &\left(\frac{z(t^{n+1}) - z(t^n)}{\Delta t} - \frac{z_h^{n+1} - z_h^n}{\Delta t}, y_h\right)_H + \mathcal{A}\big((e_{u,h}^{n+\frac{1}{2}}, e_{s,h}^{n+\frac{1}{2}}, \tilde{e}_{f,h}^{n+1}, \tilde{\delta}_h^{n+1}), (y_h, q_h)\big) \\ &\qquad = \big(\mathcal{R}^{n+\frac{1}{2}}, y_h\big)_H + \lambda_0\big(\mathcal{S}_h^{n+\frac{1}{2}}, y_h\big)_H, \quad \forall y_h \in V_h, \forall q_h \in Q_h, \end{aligned} \tag{3.60}$$

where the solid quantities errors

$$e_{u,h}^{n+\frac{1}{2}} = \frac{P_h^u u_s(t^{n+1}) + P_h^u u_s(t^n)}{2} - u_{s,h}^{n+\frac{1}{2}} \quad \text{and} \quad e_{s,h}^{n+\frac{1}{2}} = \frac{P_h^s v_s(t^{n+1}) + P_h^s v_s(t^n)}{2} - v_{s,h}^{n+\frac{1}{2}},$$

are defined as in the Crank-Nicolson scheme, whereas for the fluid and pressure quantities we consider the new errors

$$\tilde{e}_{f,h}^{n+1} = \frac{P_h^f v_f(t^{n+1}) + P_h^f v_f(t^n)}{2} - v_{f,h}^{n+1} \quad \text{and} \quad \tilde{\delta}_h^{n+1} = \frac{P_h^p p(t^{n+1}) + P_h^p p(t^n)}{2} - p_h^{n+1}.$$

In order to derive a system satisfied by the error $(e_{u,h}^{n+\frac{1}{2}}, e_{s,h}^{n+\frac{1}{2}}, \tilde{e}_{f,h}^{n+1}, \tilde{\delta}_h^{n+1})$, we compute

$$\begin{aligned}
\frac{e_{u,h}^{n+1} - e_{u,h}^n}{\Delta t} &= \frac{P_h^u u_s(t^{n+1}) - P_h^u u_s(t^n)}{\Delta t} - \frac{u_{s,h}^{n+1} - u_{s,h}^n}{\Delta t}, \\
\frac{e_{s,h}^{n+1} - e_{s,h}^n}{\Delta t} &= \frac{P_h^s v_s(t^{n+1}) - P_h^s v_s(t^n)}{\Delta t} - \frac{v_{s,h}^{n+1} - v_{s,h}^n}{\Delta t}, \\
\frac{\tilde{e}_{f,h}^{n+1} - \tilde{e}_{f,h}^n}{\Delta t} &= \frac{P_h^f v_f(t^{n+1}) - P_h^f v_f(t^{n-1})}{2\Delta t} - \frac{v_{f,h}^{n+1} - v_{f,h}^n}{\Delta t}.
\end{aligned}$$

Plugging these results into (3.60), it follows that for any $y_h = (d_{s,h}, w_{s,h}, w_{f,h}) \in V_h$ and $q_h \in Q_h$, we have

$$\begin{aligned}
&\int_\Omega \sigma_s\left(\frac{e_{u,h}^{n+1} - e_{u,h}^n}{\Delta t}\right) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\left(\frac{e_{s,h}^{n+1} - e_{s,h}^n}{\Delta t}\right)\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f \phi\left(\frac{\tilde{e}_{f,h}^{n+1} - \tilde{e}_{f,h}^n}{\Delta t}\right)\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad + \mathcal{A}\big((e_{u,h}^{n+\frac{1}{2}}, e_{s,h}^{n+\frac{1}{2}}, \tilde{e}_{f,h}^{n+1}, \tilde{\delta}_h^{n+1}), (y_h, q_h)\big) = \big(\mathcal{R}^{n+\frac{1}{2}}, y_h\big)_H + \lambda_0\big(\mathcal{S}_h^{n+\frac{1}{2}}, y_h\big)_H \\
&\qquad + \int_\Omega \sigma_s\big(\mathcal{T}_{u,h}^{n+\frac{1}{2}}\big) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,\mathcal{T}_{s,h}^{n+\frac{1}{2}} \cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f \phi\,\mathcal{T}_{f,h}^{n+1} \cdot w_{f,h}, \quad (3.61)
\end{aligned}$$

with

$$\begin{aligned}
\mathcal{T}_{u,h}^{n+\frac{1}{2}} &= \frac{P_h^u u_s(t^{n+1}) - P_h^u u_s(t^n)}{\Delta t} - \frac{u_s(t^{n+1}) - u_s(t^n)}{\Delta t}, \\
\mathcal{T}_{s,h}^{n+\frac{1}{2}} &= \frac{P_h^s v_s(t^{n+1}) - P_h^s v_s(t^n)}{\Delta t} - \frac{v_s(t^{n+1}) - v_s(t^n)}{\Delta t}, \\
\mathcal{T}_{f,h}^{n+1} &= \frac{P_h^f v_f(t^{n+1}) - P_h^f v_f(t^{n-1})}{2\Delta t} - \frac{v_f(t^{n+1}) - v_f(t^n)}{\Delta t}.
\end{aligned}$$

Note that the two first terms $\mathcal{T}_{u,h}^{n+\frac{1}{2}}$ and $\mathcal{T}_{s,h}^{n+\frac{1}{2}}$ correspond exactly to the solid components of the term $\mathcal{T}_h^{n+\frac{1}{2}}$ that have already been studied for the Crank-Nicolson scheme, while the third term $\mathcal{T}_{f,h}^{n+1}$ is different.

Choosing $(y_h, q_h) = (e_{u,h}^{n+\frac{1}{2}}, e_{s,h}^{n+\frac{1}{2}}, \tilde{e}_{f,h}^{n+1}, \tilde{\delta}_h^{n+1})$ as test function in (3.61) and exploiting the stability identity (3.39), it follows

$$\begin{aligned}
&\frac{1}{2}\left\|\big(e_{u,h}^{n+1}, e_{s,h}^{n+1}, \tilde{e}_{f,h}^{n+1}\big)\right\|_H^2 - \frac{1}{2}\left\|\big(e_{u,h}^n, e_{s,h}^n, \tilde{e}_{f,h}^n\big)\right\|_H^2 \\
&+ \frac{1}{2}\int_\Omega \rho_f \phi \left|\tilde{e}_{f,h}^{n+1} - \tilde{e}_{f,h}^n\right|^2 \mathrm{d}x + \Delta t \int_\Omega \phi\,\sigma_f(\tilde{e}_{f,h}^{n+1}) : \varepsilon(\tilde{e}_{f,h}^{n+1})\,\mathrm{d}x + \Delta t \int_\Omega \phi^2 k_f^{-1}\big(\tilde{e}_{f,h}^{n+1} - e_{s,h}^{n+\frac{1}{2}}\big)^2\,\mathrm{d}x \\
&\qquad = \Delta t \int_\Omega \theta \left|\tilde{e}_{f,h}^{n+1}\right|^2 \mathrm{d}x + \Delta t\big(\mathcal{R}^{n+\frac{1}{2}} + \lambda_0 \mathcal{S}_h^{n+\frac{1}{2}}, (e_{u,h}^{n+\frac{1}{2}}, e_{s,h}^{n+\frac{1}{2}}, \tilde{e}_{f,h}^{n+1})\big)_H \\
&\quad + \int_\Omega \sigma_s\big(\mathcal{T}_{u,h}^{n+\frac{1}{2}}\big) : \varepsilon(e_{u,h}^{n+\frac{1}{2}})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\,\mathcal{T}_{s,h}^{n+\frac{1}{2}} \cdot e_{s,h}^{n+\frac{1}{2}}\,\mathrm{d}x + \int_\Omega \rho_f \phi\,\mathcal{T}_{f,h}^{n+1} \cdot \tilde{e}_{f,h}^{n+1}. \quad (3.62)
\end{aligned}$$

The rest of the proof is almost similar to the one of Theorem 3.19.

The terms $\mathcal{R}^{n+\frac{1}{2}}$, $\mathcal{S}_h^{n+\frac{1}{2}}$, $\mathcal{T}_{u,h}^{n+\frac{1}{2}}$ and $\mathcal{T}_{s,h}^{n+\frac{1}{2}}$ have already been estimated during the analysis of the Crank-Nicolson scheme, see (3.55), (3.56) and (3.57). We only need to deal with the term $\mathcal{T}_{f,h}^{n+1}$, that we decompose as

$$\mathcal{T}_{f,h}^{n+1} = P_h^f\left(\frac{v_f(t^{n+1}) - v_f(t^{n-1})}{2\Delta t}\right) - \frac{v_f(t^{n+1}) - v_f(t^{n-1})}{2\Delta t} + \frac{v_f(t^{n+1}) - v_f(t^{n-1})}{2\Delta t} - \frac{v_f(t^{n+1}) - v_f(t^n)}{\Delta t}$$

The first part of the above expression is a space error term that can be estimated as in (3.57), namely

$$\left\| P_h^f\left(\frac{v_f(t^{n+1}) - v_f(t^{n-1})}{2\Delta t}\right) - \frac{v_f(t^{n+1}) - v_f(t^{n-1})}{2\Delta t}\right\| \leq \left(h^\ell \left\|\dot{z}\right\|_{C^0([0,T];H^{\ell+1}(\Omega))} + h^r \left\|\dot{p}\right\|_{C^0([0,T];H^r(\Omega))}\right).$$

The second part is a time error term coming from the shift between the fluid and solid quantities. Using a Taylor expansion, we easily check that

$$\left\| \frac{v_f(t^{n+1}) - v_f(t^{n-1})}{2\Delta t} - \frac{v_f(t^{n+1}) - v_f(t^n)}{\Delta t}\right\| \leq C\Delta t \left\|v_f\right\|_{C^2([0,T];[\mathrm{L}^2(\Omega)]^d)}.$$

Hence, we deduce that

$$\left\|\mathcal{T}_{f,h}^{n+1}\right\| \leq C(\Delta t + h^\ell + h^r).$$

and it is at this point that we lose the $\mathcal{O}(\Delta t^2)$ accuracy in time.

The rest of the proof is similar to the Step 4 of the proof of Theorem 3.19. In particular, the viscous part of (3.59) is recovered by decomposing the fluid error as

$$\epsilon_{f,h}^{n+1} = v_f(t^{n+1}) - v_{f,h}^{n+1} = v_f(t^{n+1}) - \frac{v_f(t^{n+1}) + v_f(t^n)}{2} + \mathcal{S}_{f,h}^{n+\frac{1}{2}} + \tilde{e}_{f,h}^{n+1},$$

and using that $\partial_t v_f \in L^2\big(0,T;[\mathrm{H}_0^1(\Omega)]^d\big)$. □

Remark 3.24. Note that our strategy of proof requires strong regularity assumptions on the continuous solution, since it is based on a comparison with the error analysis for the Crank-Nicolson scheme. Handling the temporal shift between the fluid and the solid at the projection level would lead to weaker regularity assumptions, in particular on $\partial_{tt}^2 u_s$, $\partial_{tt}^2 v_s$ and $\partial_{tt}^2 v_f$.

3.4 Numerical results

In this section, we present numerical results to illustrate the theoretical statements established previously. All the simulations have been obtained with the finite element software FEniCS [Logg et al., 2012; Alnæs et al., 2015] using a direct LU solver. First, we validate numerically the discrete energy balances for the two schemes under study, and show the influence of the fluid mass source term θ on the schemes stability. Then, the error results of Theorems 3.19 and 3.23 are discussed by means of convergence plots. Finally, we illustrate the importance of the choice of the finite element spaces employed when entering the incompressible regime.

3.4.1 Discrete energy balance and influence of the additional fluid mass input

To numerically recover the discrete energy balance derived in Section 3.2.3, we simulate the evolution of the system starting from a non-zero initial condition, but in absence of external body forces and

fluid mass source term, namely $f = 0$ and $\theta = 0$. According to (3.38) and (3.39), the discrete energy of the scheme then satisfies

$$\mathcal{E}_h^N + \underbrace{\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f(v_{f,h}^{n+\frac{1}{2}}) : \varepsilon(v_{f,h}^{n+\frac{1}{2}})\, \mathrm{d}x}_{\text{Discrete viscous fluid dissipation}} + \underbrace{\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi^2 k_f^{-1} \left(v_{f,h}^{n+\frac{1}{2}} - v_{s,h}^{n+\frac{1}{2}}\right)^2 \mathrm{d}x}_{\text{Discrete friction dissipation}} = \mathcal{E}_h^0, \tag{3.63}$$

for the Crank-Nicolson scheme and

$$\mathcal{E}_h^N + \underbrace{\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f(v_{f,h}^{n+1}) : \varepsilon(v_{f,h}^{n+1})\, \mathrm{d}x}_{\text{Discrete viscous fluid dissipation}} + \underbrace{\Delta t \sum_{n=0}^{N-1} \int_\Omega \phi^2 k_f^{-1} \left(v_{f,h}^{n+1} - v_{s,h}^{n+\frac{1}{2}}\right)^2 \mathrm{d}x}_{\text{Discrete friction dissipation}}$$
$$+ \underbrace{\frac{1}{2} \sum_{n=0}^{N-1} \int_\Omega \rho_f \phi \left|v_{f,h}^{n+1} - v_{f,h}^{n}\right|^2 \mathrm{d}x}_{\text{Discrete numerical dissipation}} = \mathcal{E}_h^0, \tag{3.64}$$

for the backward Euler scheme. The different above contributions are represented on Figure 3.1, for a test case in the domain $\Omega = (0,1)^2$ discretized in space with $[\mathbb{P}^2]^d \times [\mathbb{P}^2]^d \times [\mathbb{P}^2]^d \times \mathbb{P}^1$ finite elements. Since all the dissipation terms are strictly positive, the discrete energy curve (in blue) is strictly decreasing. Apart from the dissipation coming from the viscosity within the fluid and the friction between the two phases, the yellow curve shows an additional numerical dissipation term for the backward Euler scheme, which is not part of the balance for the Crank-Nicolson scheme. Moreover, by summing the energy and the total dissipation in the system (black curve), we see that we recover exactly the initial energy, as predicted by (3.63) and (3.64).

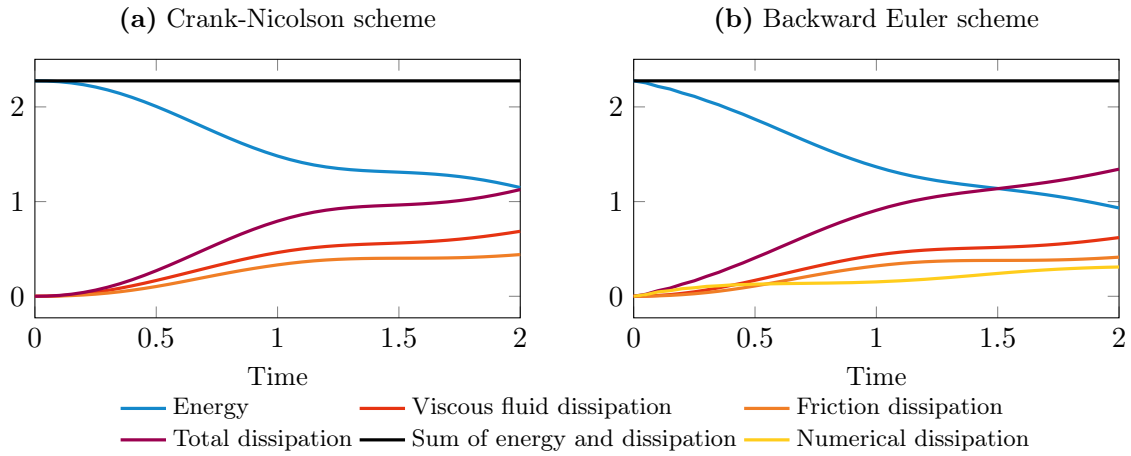


Figure 3.1 – Time evolution of the dissipation terms involved in the discrete energy balance in absence of external body forces and additional fluid mass input for the two schemes under study. Simulations run with $\Delta t = 0.05$, $T = 2$, $\rho_f = 20$, $\rho_s = 1$, $\phi = 0.5$, $\mu_f = 0.1$, $\lambda = \mu = 1$, $k_f^{-1} = 1.5\mathbb{I}$ and $\theta = 0$.

In Figure 3.2, we simulate the same test case as in Figure 3.1a, but with a non-zero fluid mass source term θ. The resulting curves shed light on the influence of the sign of θ on the system dynamics. If θ is negative, the term $-\theta v_f$ supplies the system with an additional dissipation term, so that the energy in Figure 3.2a decreases faster than in Figure 3.1a. If θ is positive, then the term $-\theta v_f$ brings fluid kinetic energy to the system. In this case, if this incoming rate of fluid kinetic energy is not compensated by the viscous and friction dissipation terms, then the total energy increases, as it is the case in Figure 3.2b.

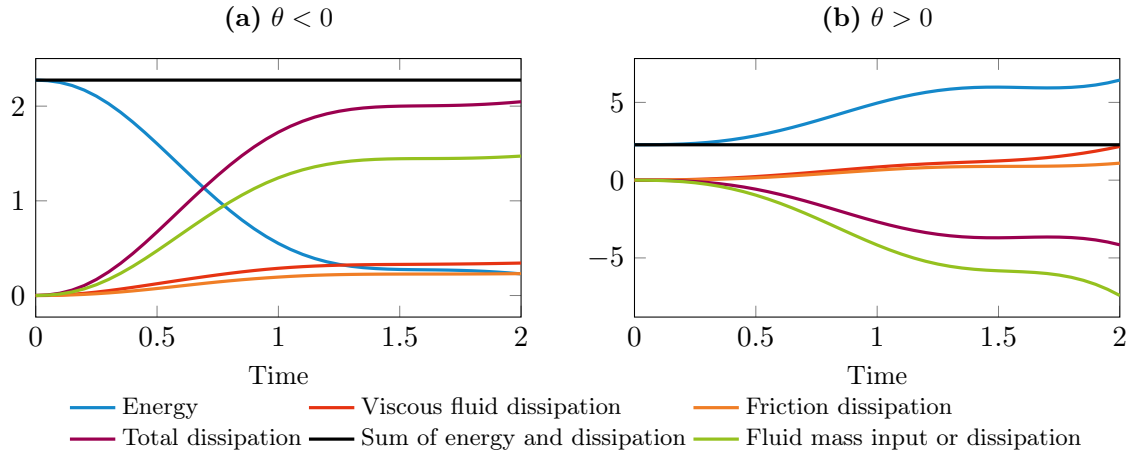
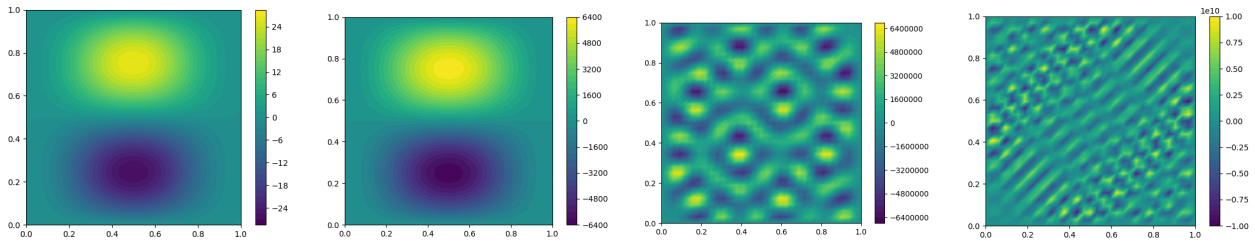


Figure 3.2 – Time evolution of the dissipation terms involved in the discrete energy balance with $f = 0$ for the Crank-Nicolson scheme, but with an additional fluid mass source term. Simulations run with $\Delta t = 0.05$, $T = 2$, $\rho_f = 20$, $\rho_s = 1$, $\phi = 0.5$, $\mu_f = 0.1$, $\lambda = \mu = 1$, $k_f^{-1} = 1.5\mathbb{I}$ and $\theta = -10$ (left) or $\theta = 10$ (right).

Another implication of the additional fluid mass source term θ – when it does not satisfy the smallness condition (3.6) – is that it imposes a restriction on the time step. Indeed, from Theorems 3.12 and 3.17, the existence of the discrete solution associated with the Crank-Nicolson or backward Euler schemes is respectively ensured under the sufficient condition (3.34) or (3.36), namely

$$\Delta t < \frac{2\rho_f \phi_{\min}}{\|\theta\|_{L^\infty(\Omega)}} \qquad \text{or} \qquad \Delta t < \frac{\rho_f \phi_{\min}}{\|\theta\|_{L^\infty(\Omega)}}.$$

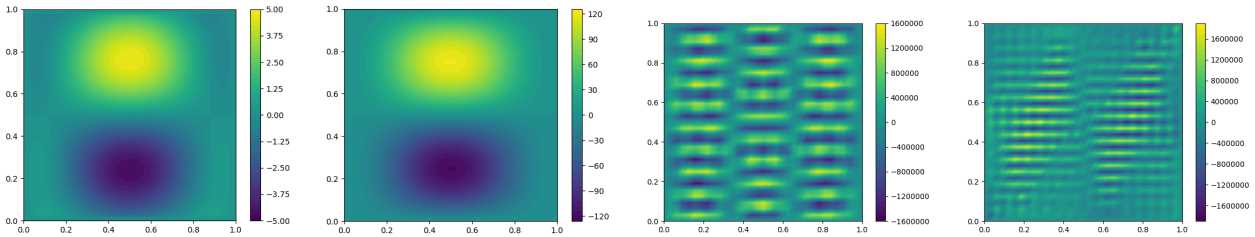
Figures 3.3 and 3.4 highlight the unstability of the schemes when these conditions are not respected, that thus appear to be necessary for the considered test case. For the Crank-Nicolson scheme, Figures 3.3c and 3.3d show that the computed fluid velocity diverges after a few iterations in time when (3.34) is not satisfied, whereas the fluid velocity profile is close to the initial condition profile when (3.34) is satisfied, see Figures 3.3a and 3.3b. The same phenomenon occurs for the backward Euler scheme in Figure 3.4, but with a time step restriction that is twice more restrictive than for the Crank-Nicolson scheme, in accordance with (3.36).



(a) $\Delta t = 0.05$ (b) $\Delta t = 0.09$ (c) $\Delta t = 0.11$ (d) $\Delta t = 0.15$

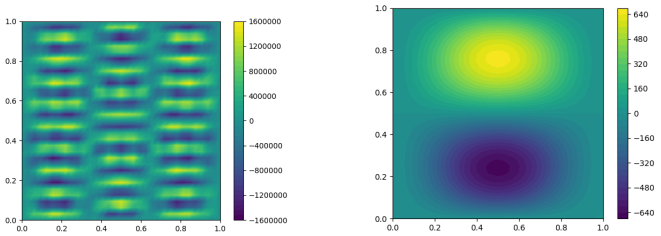
Figure 3.3 – Fluid velocity x-component profile computed with the Crank-Nicolson scheme after three iterations in time, for different time steps close to the threshold $\frac{2\rho_f \phi_{\min}}{\|\theta\|_{L^\infty(\Omega)}} = 0.1$. Simulation run with $f = 0$, $\rho_f = \rho_s = 1$, $\phi = 0.5$, $\mu_f = 0.001$, $\lambda = \mu = 1$, $k_f^{-1} = 0$ and $\theta = 10$.

Interestingly, these unstabilities can be removed by increasing the value of the fluid viscosity, as shown in Figure 3.5. Indeed, if μ_f is large enough, then the incoming rate of fluid kinetic energy coming from the θ term can be counterbalanced by the fluid viscous dissipation even if the time step restriction is not fulfilled, as mentioned in Remarks 3.11, 3.18 and 3.22.



(a) $\Delta t = 0.02$ (b) $\Delta t = 0.04$ (c) $\Delta t = 0.06$ (d) $\Delta t = 0.09$

Figure 3.4 – Fluid velocity x-component profile computed with the backward Euler scheme after three iterations in time, for different time steps close to the threshold $\frac{\rho_f \phi_{\min}}{\|\theta\|_{L^\infty(\Omega)}} = 0.05$. Simulation run with $f = 0$, $\rho_f = \rho_s = 1$, $\phi = 0.5$, $\mu_f = 0.001$, $\lambda = \mu = 1$, $k_f^{-1} = 0$ and $\theta = 10$.



(a) $\mu_f = 0.001$ (b) $\mu_f = 0.1$

Figure 3.5 – Fluid velocity x-component profile computed with the backward Euler scheme after three iterations in time, for $\Delta t = 0.06$ and two different values of fluid viscosity.

Finally, Table 3.1 illustrates that even when the time step restriction is satisfied, the error between the discrete and continuous solutions may be large in long time simulations if the time step is not small enough. This is due to the constant $\exp\left(\frac{4(\rho_f \phi_{\min})^{-1}\|\theta\|_{L^\infty(\Omega)} T}{1-4\Delta t(\rho_f \phi_{\min})^{-1}\|\theta\|_{L^\infty(\Omega)}}\right)$ appearing in the error estimates (3.48) and (3.59). In the example of Table 3.1, we see that Δt must be less than 0.001 to get an error that is not polluted by this exponential growth.

Δt	Relative error	Δt	Relative error
0.0002	0.02	0.005	1.8
0.0005	0.08	0.01	8.8
0.001	0.20	0.02	240

Table 3.1 – Relative error $\left\|z^{\text{ref}}(T) - z_h^{n_T}\right\|_H / \left\|z^{\text{ref}}(T)\right\|_H$ at $T = 1$ between the discrete solution computed for different time steps and a reference solution z^{ref} computed for $\Delta t = 0.0001$, obtained with the backward Euler scheme and the same physical parameters as in Figure 3.4.

3.4.2 Convergence rates

Next, we present convergence plots generated using the manufactured solution method in the unit square domain $\Omega = (0,1)^2$. To build an analytical solution, we assume that the porosity ϕ is constant and we pick a function v^{ref} such that $\operatorname{div} v^{\text{ref}} = 0$ in Ω and $v^{\text{ref}} = 0$ on $\partial\Omega$, for instance

$$v^{\text{ref}}(x,y) = \Big(\sin(2\pi y)\big(\cos(2\pi x) - 1\big), \sin(2\pi x)\big(1 - \cos(2\pi y)\big)\Big).$$

Then, we choose the fluid and solid velocities analytical solutions as

$$v_s^{\text{ref}}(x,y,t) = \cos(t)\phi\, v^{\text{ref}}(x,y) \quad \text{and} \quad v_f^{\text{ref}}(x,y,t) = \cos(t)(1-\phi)\, v^{\text{ref}}(x,y),$$

in such a way that, since ϕ is constant, we have

$$\operatorname{div}\left((1-\phi)v_s^{\mathrm{ref}}+\phi v_f^{\mathrm{ref}}\right)=\cos(t)\phi(1-\phi)\operatorname{div} v^{\mathrm{ref}}=0.$$

The solid displacement analytical solution is then obtained by time integration of the solid velocity, namely $u_s^{\mathrm{ref}}(x,y,t)=\sin(t)\phi\, v^{\mathrm{ref}}(x,y)$. Lastly, for the pressure analytical solution, we take $p^{\mathrm{ref}}(x,y,t)=\sin(t)\sin(2\pi x)\sin(2\pi y)$, which satisfies the condition $\int_\Omega p\,\mathrm{d}x=0$. To simplify, we assume that $\theta=0$. The simulation is then run with the source terms and initial conditions associated with the analytical solution, for $\rho_s=\rho_f=\mu_f=\lambda=\mu=1$, $k_f=\mathbb{I}$, $\phi=0.5$ and $T=1$. By comparing the resulting discrete solution to the previous analytical solution, we investigate numerically the spatial and temporal convergence rates of the two proposed schemes.

In Figures 3.6 and 3.7, the simulation is performed with a very small time step $\Delta t=0.005$ and the mesh size is progressively decreased. For the spatial discretization, we use finite element spaces pairs (X_h,Q_h) that are known to be stable for Stokes problem, namely the MINI element or Taylor-Hood elements [Glowinski, 2003; Boffi et al., 2013]. Figure 3.6-left corroborates the statement of Theorem 3.19: it shows a spatial convergence rate of 1 in the energy and fluid viscous dissipation norms for the MINI element, for which $\ell=r=1$. In Figure 3.6-right, the energy norm is decomposed into the three contributions of solid displacement, solid velocity and fluid velocity. We observe that the convergence rate is of order 2 for the velocities, so that the energy norm convergence rate is restricted by the solid displacement term. This extra convergence probably comes from the fact that estimate (3.43) is optimal only for the displacement $[\mathrm{H}_0^1(\Omega)]^d$ norm, but may be improved for the velocities $[\mathrm{L}^2(\Omega)]^d$ norm using a duality argument. This may be one of the drawbacks of the T-coercivity method since it is by essence an *all-in-one* approach handling all the variables together. Note that even if this result is not given by Theorem 3.19, we also recover numerically the pressure convergence, with a convergence rate of 1.5 as found in other studies on the MINI element [Eichel et al., 2011; Cioncolini and Boffi, 2019].

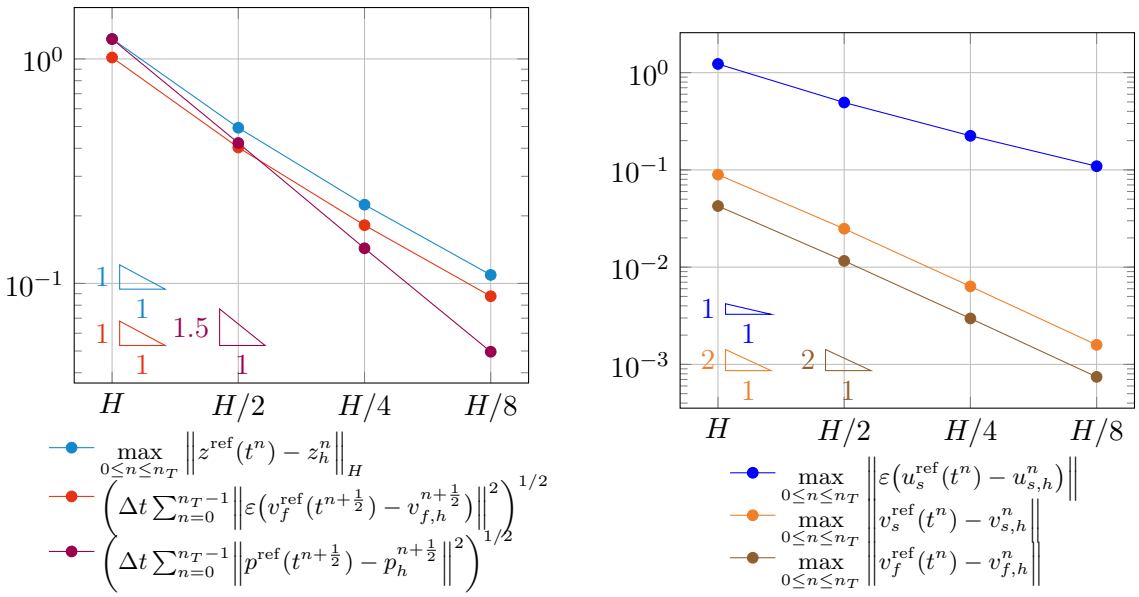


Figure 3.6 – Error curves with respect to the mesh size h for $[\mathbb{P}_b^1]^d\times[\mathbb{P}_b^1]^d\times[\mathbb{P}_b^1]^d\times\mathbb{P}^1$ elements. Simulation run with the Crank-Nicolson scheme for $\Delta t=0.005$, starting from a mesh size H that corresponds to a uniform mesh built with 8 subdivisions along each axis direction.

For Taylor-Hood elements, we have $\ell=r=2$ and Figure 3.7 gives a convergence rate of 3 in the energy norm. This superconvergence is probably due to the C^∞ regularity of our analytical solution. Here again, we find an improved convergence for the velocity variables compared to the displacement one.

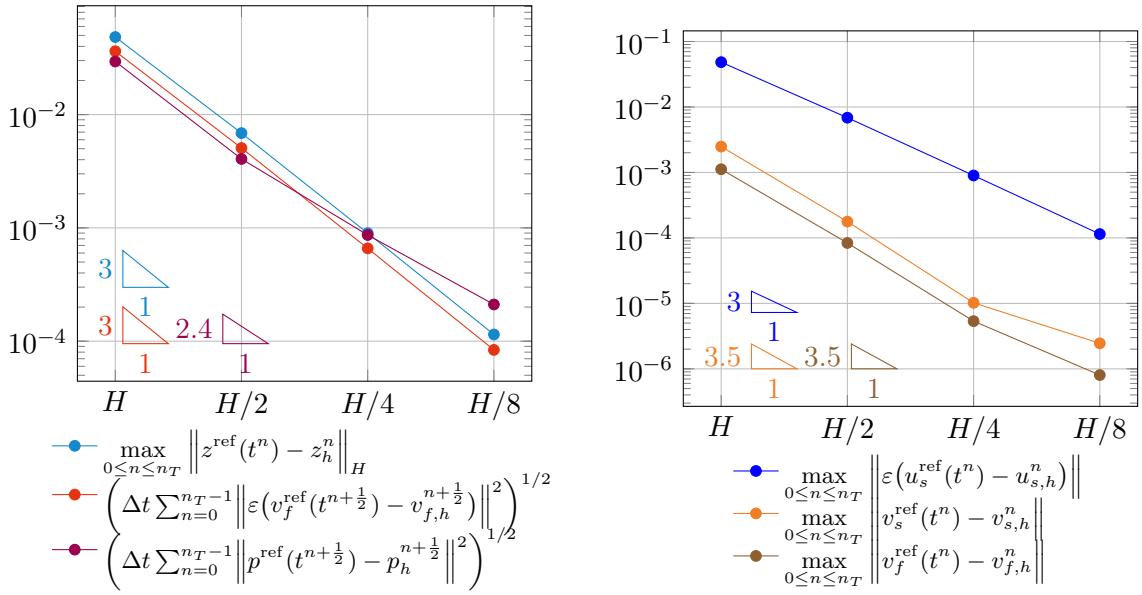


Figure 3.7 – Error curves with respect to the mesh size h for $[\mathbb{P}^2]^d \times [\mathbb{P}^2]^d \times [\mathbb{P}^2]^d \times \mathbb{P}^1$ elements. Simulation run with the Crank-Nicolson scheme for $\Delta t = 0.005$, starting from a mesh size H that corresponds to a uniform mesh built with 8 subdivisions along each axis direction.

Figures 3.6 and 3.7 are obtained for the Crank-Nicolson scheme, but similar results hold for the backward Euler scheme since this scheme does not change the spatial discretization of the problem. However, when it comes to temporal convergence, Figure 3.8 highlights the major difference between the two proposed schemes: the Crank-Nicolson scheme is of second order in time, whereas the backward Euler scheme is of first order. Note however that on Figure 3.8b, the solid velocity still shows a second-order convergence in time in the backward Euler scheme, as if it was not affected by the other variables.

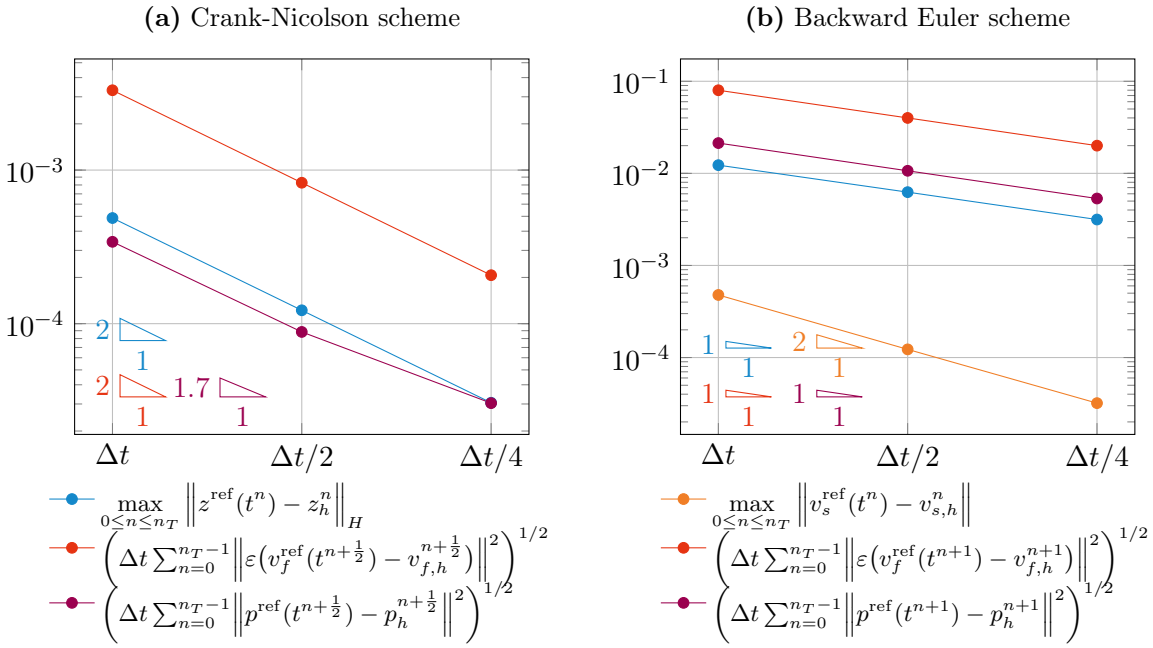


Figure 3.8 – Error curves of the two proposed schemes with respect to the time step. Simulation run for a very refined mesh, starting from the time step $\Delta t = 0.1$.

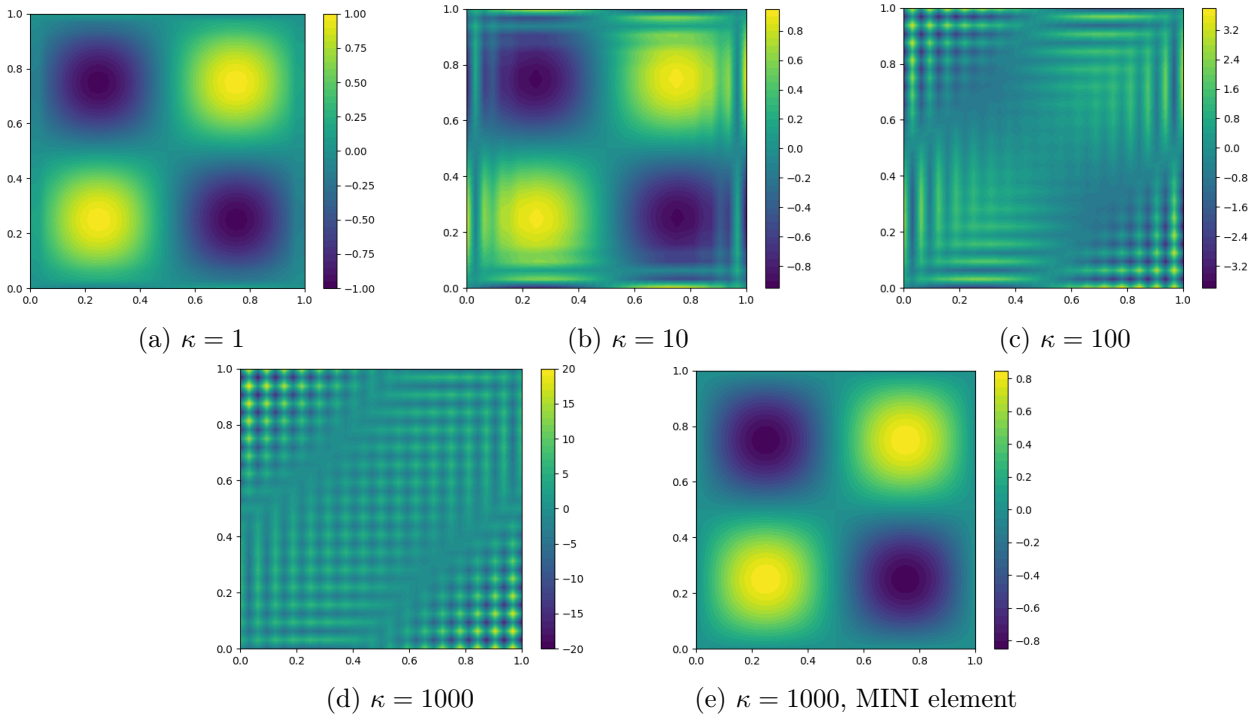
3.4.3 Choosing the finite element spaces in the incompressible limit

In the previous sections, we have focused on the case where the porous material is fully incompressible, namely $\kappa = \infty$. Yet, our analysis also provides guidelines to discretize the system (3.1), in which the mixture divergence equation is penalized by a term of the form $\frac{b-\phi}{\kappa}\partial_t p$. As a matter of fact, it was shown in [Barré et al., 2023, Theorem 4.2] that the solution of the compressible system (3.1) converges towards the solution of the incompressible system (3.2) as the bulk modulus κ goes to infinity. This suggests to use finite elements satisfying the inf-sup condition (3.18) to discretize the system (3.1) when κ is large, namely for nearly incompressible materials. Theorems 3.19 and 3.23 extend the convergence analysis of [Burtschell et al., 2019] and [Barnafi et al., 2021] up to the incompressible limit, which also suggests a discretization of (3.1) that is robust with respect to κ provided that the discrete inf-sup condition (3.18) is fulfilled.

To illustrate numerically what happens if (3.18) is not satisfied, we use the same analytical solution as in the previous section and simulate the solution of (3.1) for different values of κ with $[\mathbb{P}^1]^d \times [\mathbb{P}^1]^d \times [\mathbb{P}^1]^d \times \mathbb{P}^1$ finite elements, which do not satisfy the discrete inf-sup condition. To do so, we use the Crank-Nicolson scheme (3.15) where the mixture divergence equation (3.15d) is replaced by

$$\frac{b-\phi}{\kappa}\frac{p^{n+1}-p^n}{\Delta t} + \operatorname{div}\left((1-\phi)\, v_s^{n+\frac{1}{2}} + \phi\, v_f^{n+\frac{1}{2}}\right) = g^{n+\frac{1}{2}},$$

with $g^{n+\frac{1}{2}}$ a source term corresponding to the pressure analytical solution. The resulting pressure profile is shown in Figure 3.9, where pressure oscillations appear when entering the incompressible regime. The size of these oscillations increases with the bulk modulus κ, leading to a completely incorrect pressure above $\kappa = 100$. Finally, Figure 3.9e shows that these oscillations are removed when using a Stokes-stable pair, as indicated by our theoretical results.



(a) $\kappa = 1$ (b) $\kappa = 10$ (c) $\kappa = 100$

(d) $\kappa = 1000$ (e) $\kappa = 1000$, MINI element

Figure 3.9 – Pressure profile for different values of bulk modulus. Except for Figure 3.9e, the problem is discretized using $[\mathbb{P}^1]^d \times [\mathbb{P}^1]^d \times [\mathbb{P}^1]^d \times \mathbb{P}^1$ finite elements, which do not satisfy the discrete inf-sup condition.

Conclusion

We have derived error estimates for two monolithic schemes: one based on a Crank-Nicolson time discretization for both the fluid and structural parts, the other based on an implicit backward-Euler discretization for the fluid part. For both schemes, the spatial discretization is a well-chosen finite element discretization that satisfies an inf-sup condition that allows one to derive a discrete T-coercivity property, independent of the porosity of the mixture, hence ensuring robustness with respect to it. The T-coercivity property approach provides the existence of the discrete solution, assuming the time step is small enough compared to the additional fluid mass input but without any permeability condition. Moreover, the T-coercivity allows us to define a well-adapted projection operator on the finite element space, which is a key argument of the error derivation. The theoretical results are confirmed by numerical simulations. We believe that the considered strategy paves the way to propose an asymptotically stable scheme with respect to the bulk modulus that will not suffer from poroelastic locking, which occurs in Biot-type systems [Phillips and Wheeler, 2009; Ferronato et al., 2010; Haga et al., 2012; Oyarzúa and Ruiz-Baier, 2016; Yi, 2017; Lee, 2018]. In future work, we expect to use the proposed time schemes as pivot in order to obtain error estimates for splitting strategies commonly used for poromechanical models [Zienkiewicz et al., 1993; Huang et al., 2001; Li et al., 2003; Markert et al., 2009].

CHAPTER 4

A projection scheme for an incompressible soft material poromechanics model

In this chapter, splitting schemes for the linearized poromechanics model (13) are proposed and analyzed. Different kind of boundary conditions are considered and special attention is paid to the treatment of the friction term between the fluid and structure phases. In particular, to avoid the added mass effect in the case of total stress boundary conditions, we use a Robin-Robin strategy inspired from fluid-structure interaction problems. Stability analysis is performed together with a convergence analysis in the case of homogeneous Dirichlet boundary conditions. The chapter is concluded with numerical illustrations. The results of this chapter, obtained in collaboration with Céline Grandmont and Philippe Moireau, will be the object of a forthcoming article. Moreover, in September 2023, I presented this work at the *European Conference on Numerical Mathematics and Advanced Applications (ENUMATH)* in Lisbon, Portugal.

Contents

Introduction

In the previous chapter, we analyzed a monolithic scheme for the linearized poromechanics problem under study in the incompressible case. When the discrete inf-sup condition is satisfied, this scheme has been shown to circumvent pressure oscillations occuring in the incompressible limit and to be robust with respect to porosity and permeability. However, it is a strongly implicit scheme that requires to solve at each time step a large linear system with a saddle-point structure. The main goal of this chapter is to propose a splitting scheme that enables to decouple the solid, fluid and pressure equations at each time step. Our approach is close to the Chorin-Temam projection method [Chorin, 1969; Temam, 1969] but takes into account the specific saddle-point structure of the poromechanics problem involving the mixture divergence constraint. Moreover, it includes the case of total stress boundary conditions thanks to a Robin-Robin coupling technique inspired by fluid-structure interaction problems [Burman et al., 2022a].

Projection methods, also known as fractional-step methods, were originally introduced for the study of incompressible fluid problems. These methods are prediction-correction schemes, in which a tentative velocity is computed without taking into account the incompressibility constraint, and then corrected by projecting it on a divergence-free space. First proposed by Chorin and Temam [Chorin, 1969; Temam, 1969], these schemes were improved in [Goda, 1979] and [Van Kan, 1986], their convergence analysis was performed in [Heywood and Rannacher, 1990; Rannacher, 1992; Shen, 1995; Guermond, 1996; Guermond and Quartapelle, 1998; Badia and Codina, 2007] and general boundary conditions were considered in [Maria Denaro, 2003; Guermond et al., 2005], see the review [Guermond et al., 2006]. More recently, projection schemes were used in fluid-structure interaction problems [Fernández et al., 2007; Guidoboni et al., 2009; Astorino and Grandmont, 2010]. In order to obtain robust schemes with respect to added-mass effects occuring in such problems, they were combined with Nitsche's method [Burman and Fernández, 2009], Robin coupling derived from Nitsche's method [Astorino et al., 2010] or again Robin-Robin coupling [Burman et al., 2022a]. Similarly, projection methods were used in fluid-porous structure interaction problems [Caiazzo et al., 2011; Bukač et al., 2015b].

For Biot-type systems, the most popular splitting procedures are the undrained split and fixed-stress split algorithms [Zienkiewicz et al., 1988; Settari and Mourits, 1998], for which time convergence and space-time convergence were respectively established in [Mikelić and Wheeler, 2013] and [Girault et al., 2019]. These algorithms were applied to the quasi-static Biot equations, namely equation (1.1) with $\rho = 0$, but also to the multiple-network poroelasticity equations [Hong et al., 2020]. For the fixed-stress split, variants and numerical optimizations were described in [Both et al., 2017; Storvik et al., 2019; Both et al., 2019a]. Following this approach, [Both et al., 2022] proposes an alternating minimization splitting scheme for the linearized poromechanics model studied in the previous chapters, which is shown to be convergent by adapting the method from [Both et al., 2019b]. However, when the bulk modulus becomes large, this scheme requires a large number of iterations before convergence is reached. In addition, total stress boundary conditions are not considered.

Less attention has been paid to projection methods for unsteady Biot-type problems. For the unsteady Biot equations, namely equation (1.1) with $\rho > 0$, [Zienkiewicz et al., 1993] developed a staggered time stepping algorithm that was further improved in [Huang et al., 2001] and [Li et al., 2003]. In [Markert et al., 2009], a fractional-step method is studied for the incompressible poroacoustic equations, namely system (1.7) with $c_0 = 0$, which is very close to the poromechanics model considered here. In these studies, equal-order finite elements for the velocities and the pressure are employed, which results in a condition on the time step for the scheme to be stable. Lastly, let us mention that Chorin-Temam-like methods were also used to deal with mixture divergence constraints appearing in biofilms growth modeling, see [Clarelli et al., 2013] and [Polizzi et al., 2017].

In [Burtschell et al., 2017], the non-linear poromechanics model from [Chapelle and Moireau, 2014] is discretized using a partitioned method adapted from fluid-structure interaction [Astorino

et al., 2010]. Using a Newmark scheme for the solid part [Gonzalez, 2000; Le Tallec and Mani, 2000; Hauret and Le Tallec, 2006] together with a projection scheme for the fluid part, the authors propose a splitting scheme in which the fluid viscous sub-step is treated explicitly and that tackles the case of total stress boundary conditions. Yet, the fluid projection sub-step is still coupled implicitly with the solid sub-step through the interstitial pressure.

In this chapter, we present and analyze a projection scheme for the incompressible linearized poromechanics problem from Chapters 1 and 3. Our strategy first consists in computing tentative fluid and solid velocities that do not verify the incompressibility constraint, but take into account the fluid viscous effects, the solid deformation and the friction between the two phases. Then, the pressure is obtained by solving a Poisson-like problem with an homogenized density coefficient and the end-of-step velocities are built from the tentative ones in order to satisfy the mixture divergence constraint. Because the projection is made on the incompressible space common to both phases, this approach does not require any iteration between the different sub-steps as in [Burtschell et al., 2017] and [Both et al., 2022], except if the friction term is treated explicitly. The proposed scheme is close to the one designed in [Markert et al., 2009], for which no stability analysis has been carried out to our knowledge. Here, we provide a complete stability analysis of the scheme depending on the treatment of permeability and of the various boundary conditions under consideration, and show the convergence of the method provided that the discrete inf-sup condition is satisfied.

The chapter is organized as follows. First, Section 4.1 briefly recalls the model under study and presents the different types of boundary conditions. Then, Section 4.2 is devoted to the description of the method and its variants, and the stability analysis is performed with a particular attention made on the friction term. In Section 4.3, we show how to extend the projection scheme to Neumann and total stress boundary conditions while keeping its stability properties. Coming back to the simple case of Dirichlet boundary conditions, an error estimate is provided in Section 4.4. Finally, the theoretical findings are illustrated by numerical results in Section 4.5 and the efficiency of the resulting solver is compared to a monolithic approach.

4.1 Problem setting

We consider the same poromechanics problem as in Chapters 1 and 3. We focus on the incompressible case $\kappa = +\infty$. To simplify, we assume that there is no fluid mass input in the porous medium, namely $\theta = 0$. We refer to Chapter 3 for more details about the numerical treatment of the case $\theta \neq 0$. Written as a first-order evolution equation, the porous problem under study reads: find u_s, v_s, v_f and p such that

$$\begin{cases} \partial_t u_s - v_s = 0, & \text{(4.1a)} \\ \rho_s(1-\phi)\,\partial_t v_s - \operatorname{div}\big(\sigma_s(u_s)\big) - \phi^2\, k_f^{-1}(v_f - v_s) + (1-\phi)\,\nabla p = \rho_s(1-\phi)\, f, & \text{(4.1b)} \\ \rho_f \phi\,\partial_t v_f - \operatorname{div}\big(\phi\,\sigma_f(v_f)\big) + \phi^2\, k_f^{-1}(v_f - v_s) + \phi\,\nabla p = \rho_f \phi\, f, & \text{(4.1c)} \\ \operatorname{div}\big((1-\phi)\, v_s + \phi\, v_f\big) = 0, & \text{(4.1d)} \end{cases}$$

where we used the same notation as in the previous chapters for denoting the different physical quantities involved in the model. The assumptions made on the physical parameters are similar to the ones made previously, except for the permeability tensor k_f for which we assume that there exists $k_{\max}^{-1} > 0$ such that

$$k_f^{-1} v \cdot v \leq k_{\max}^{-1} |v|^2, \quad \forall v \in \mathbb{R}^d, \tag{4.2}$$

in addition to the coercivity condition $k_f^{-1} v \cdot v \geq k_{\min}^{-1} |v|^2$ made before.

To be well-posed, Problem (4.1) has to be complemented with initial conditions for the solid displacement, solid velocity and fluid velocity, and with boundary conditions. Let us denote by Γ_D, Γ_N and Γ_T three subsets (possibly empty) of $\partial\Omega$, such that $\partial\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_T$. In this

chapter, we consider three types of boundary conditions. First, Dirichlet boundary conditions, either homogeneous, namely

$$\begin{aligned} u_s &= 0, \quad \text{on } \Gamma_D, \\ v_s &= 0, \quad \text{on } \Gamma_D, \\ v_f &= 0, \quad \text{on } \Gamma_D, \end{aligned} \tag{4.3}$$

or non-homogeneous, namely

$$\begin{aligned} u_s &= u_{s,bc}, \quad \text{on } \Gamma_D, \\ v_s &= v_{s,bc}, \quad \text{on } \Gamma_D, \\ v_f &= v_{f,bc}, \quad \text{on } \Gamma_D, \end{aligned} \tag{4.4}$$

where $u_{s,bc}$, $v_{s,bc}$ and $v_{f,bc}$ belong to $[\mathrm{L}^2(\Gamma_D)]^d$ and verify the compatibility condition

$$\int_{\Gamma_D} \big((1-\phi)v_{s,bc} + \phi\, v_{f,bc}\big) \cdot n \,\mathrm{d}S = 0. \tag{4.5}$$

Second, Neumann boundary conditions, for which an exterior traction force $b \in [\mathrm{L}^2(\Gamma_N)]^d$ acting on a part of the boundary is distributed between the solid and fluid stresses in the following way:

$$\begin{aligned} \sigma_s(u_s)n - (1-\phi)\, pn &= (1-\phi)\, b, \quad \text{on } \Gamma_N, \\ \phi\, \sigma_f(v_f)n - \phi\, pn &= \phi\, b, \quad \text{on } \Gamma_N, \end{aligned} \tag{4.6}$$

with n the exterior normal of $\partial\Omega$. Third, a total stress boundary condition, which is close to transmission boundary conditions encountered in fluid-solid interaction through the fluid-solid interface. In this case, the repartition of the boundary traction b between the fluid and solid stresses is not precised but the solid and fluid velocities are assumed to match at the interface, namely

$$\begin{aligned} v_f &= v_s, \quad \text{on } \Gamma_T, \\ \sigma^{\mathrm{tot}} n &= b, \quad \text{on } \Gamma_T, \end{aligned} \tag{4.7}$$

where

$$\sigma^{\mathrm{tot}} = \sigma_s(u_s) + \phi\, \sigma_f(v_f) - p\mathbb{I},$$

denotes the total stress tensor of the porous medium. This last kind of boundary conditions has only been considered for Problem (4.1) in [Burtschell et al., 2017] and [Burtschell et al., 2019], where they are treated using a Nitsche's method inspired from [Astorino et al., 2010]. Similar boundary conditions have been largely studied in fluid-structure interaction problems, see for instance [Gerbeau and Vidrascu, 2003; Causin et al., 2005; Fernández et al., 2007; Guidoboni et al., 2009]. In fluid-structure interaction, they occur on the fluid-solid interface between a fluid domain Ω_f and a solid domain Ω_s. The major difference here is that the fluid and the solid phases cohabit in the same domain Ω, so that (4.7) is written on a part of the porous boundary and with the same exterior normal n for the fluid and the solid. Moreover, in fluid-structure interaction problems, the fluid incompressibility is known to cause an *added-mass effect* bringing numerical unstabilities. A similar issue happens here for the porous medium with the incompressibility constraint (4.1d).

Problem (4.1) is a strongly coupled system in which the fluid and solid parts are coupled through the friction term, the interstitial pressure and the incompressibility constraint, and that can be solved numerically using a monolithic time scheme as in the previous chapter. Let us now present our splitting strategy to decouple the different equations of (4.1) and solve the fluid, solid and pressure separately at each time step.

4.2 Time discretization: decoupling strategies

In this section, we will suppose that the porous medium is submitted to homogeneous Dirichlet boundary conditions (4.3), namely that $\Gamma_N = \emptyset$ and $\Gamma_T = \emptyset$. The treatment of Neumann and total stress boundary conditions (4.6) and (4.7) will be the object of the next section.

4.2.1 The fully decoupled projection scheme

The proposed projection scheme is detailed in Scheme 1 below. The first step is a prediction step in which we compute solid and fluid tentative velocities $\widetilde{v}_s^{n+1}$ and $\widetilde{v}_f^{n+1}$ without taking into account the incompressibility constraint and the pressure gradient terms, but only the solid deformation, the fluid viscous effects and the friction between the two phases. Then, these velocities are corrected by reincorporating the pressure gradient term in (4.10a) and (4.10b). This is a *projection step*: to obtain the end-of-step velocities v_s^{n+1} and v_f^{n+1}, the tentative velocities $\widetilde{v}_s^{n+1}$ and $\widetilde{v}_f^{n+1}$ are projected on the mixture divergence-free space

$$H_\phi = \big\{(v_s, v_f) \in [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \,:\, \operatorname{div}\big((1-\phi)\,v_s + \phi\,v_f\big) = 0 \quad \text{in } \mathcal{D}'(\Omega) \\ \text{and } \big((1-\phi)\,v_s + \phi\,v_f\big)\cdot n = 0 \quad \text{on } \Gamma_D\big\}$$

studied in Chapter 1, see Proposition 1.25. Lastly, the solid displacement u_s^{n+1} is computed directly in the structure prediction sub-step, and does not need to be corrected.

Scheme 1 Explicit treatment of permeability (non-incremental version)

Step 1: `(prediction step)`

– **Step 1.1**: `(structure prediction sub-step)`
Find u_s^{n+1} and $\widetilde{v}_s^{n+1}$ such that $u_s^{n+1}|_{\Gamma_D} = 0$, $\widetilde{v}_s^{n+1}|_{\Gamma_D} = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{\widetilde{v}_s^{n+1} - v_s^n}{\Delta t} - \operatorname{div}\left(\sigma_s\Big(\dfrac{u_s^{n+1}+u_s^n}{2}\Big)\right) \\ \qquad\qquad - \phi^2 k_f^{-1}\Big(v_f^n - \dfrac{\widetilde{v}_s^{n+1}+v_s^n}{2}\Big) = \rho_s(1-\phi) f^{n+\frac{1}{2}}, & (4.8a) \\ \dfrac{u_s^{n+1}-u_s^n}{\Delta t} = \dfrac{\widetilde{v}_s^{n+1}+v_s^n}{2}. & (4.8b) \end{cases}$$

– **Step 1.2**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1}$ such that $\widetilde{v}_f^{n+1}|_{\Gamma_D} = 0$ and

$$\rho_f\phi\frac{\widetilde{v}_f^{n+1} - v_f^n}{\Delta t} - \operatorname{div}\big(\phi\,\sigma_f(\widetilde{v}_f^{n+1})\big) + \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1} - \frac{\widetilde{v}_s^{n+1}+v_s^n}{2}\Big) = \rho_f\phi\, f^{n+\frac{1}{2}}. \quad (4.9)$$

Step 2: `(correction step)`
Find v_s^{n+1}, v_f^{n+1} and p^{n+1} such that $\int_\Omega p^{n+1}\,\mathrm{d}x = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{v_s^{n+1} - \widetilde{v}_s^{n+1}}{\Delta t} + (1-\phi)\,\nabla p^{n+1} = 0, & (4.10a) \\ \rho_f\phi\dfrac{v_f^{n+1} - \widetilde{v}_f^{n+1}}{\Delta t} + \phi\,\nabla p^{n+1} = 0, & (4.10b) \\ \operatorname{div}\big((1-\phi)v_s^{n+1} + \phi\,v_f^{n+1}\big) = 0, & (4.10c) \\ \big((1-\phi)v_s^{n+1} + \phi\,v_f^{n+1}\big)\cdot n|_{\Gamma_D} = 0. & (4.10d) \end{cases}$$

Note that in the prediction step, we choose to start by advancing the structure before the fluid. This is motivated by the fact that the friction term has a greater impact on the fluid than on the structure, which is more rigid. As a consequence, it is better to treat the friction term explicitly in the solid sub-step, see the term v_f^n appearing in (4.8a), which allows to treat it implicitly in the fluid sub-step (4.9). Note moreover that as in Chapter 3, the solid part is discretized using a Newmark scheme while the fluid part is discretized with a backward Euler scheme.

From a numerical point of view, the main advantage of this scheme is that in the prediction step, the solid and fluid degrees of freedom are decoupled, which was not the case in the monolithic approach from Chapter 3 that requires to solve a saddle-point problem at each time step. In the projection step, the solid, fluid and pressure degrees of freedom seem coupled at first sight, but this is only because system (4.10) corresponds to the Darcy-formulation of the projection step. In practice, it is more convenient to first solve the pressure with a Poisson-like equation, and then correct the velocities accordingly. As a matter of fact, under suitable regularity assumptions, dividing (4.10a) by ρ_s, (4.10b) by ρ_f, summing and taking the divergence, we find that

$$\frac{1}{\Delta t}(\operatorname{div} v_m^{n+1} - \operatorname{div} \widetilde{v}_m^{n+1}) + \operatorname{div}\left(\left(\frac{1-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right)\nabla p^{n+1}\right) = 0,$$

where we use the notation $\widetilde{v}_m^{n+1}$ and v_m^{n+1} for the tentative and end-of-step mixture velocities, namely

$$\widetilde{v}_m^{n+1} = (1-\phi)\widetilde{v}_s^{n+1} + \phi\,\widetilde{v}_f^{n+1} \quad \text{and} \quad v_m^{n+1} = (1-\phi)v_s^{n+1} + \phi\,v_f^{n+1}.$$

From (4.10c), we know that $\operatorname{div} v_m^{n+1} = 0$. Thus, noting $\rho_{\text{eff}} = \left(\frac{1-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right)^{-1}$, we obtain

$$-\operatorname{div}\left(\rho_{\text{eff}}^{-1}\nabla p^{n+1}\right) = -(\Delta t)^{-1}\operatorname{div}\widetilde{v}_m^{n+1}.$$

Similarly, by dividing (4.10a) by ρ_s, (4.10b) by ρ_f, summing, taking the normal trace on Γ_D and using that $v_m^{n+1}\cdot n|_{\Gamma_D} = 0$ in virtue of (4.10d), we get $\rho_{\text{eff}}^{-1}\nabla p^{n+1}\cdot n|_{\Gamma_D} = 0$. Hence, the Poisson formulation of the correction step reads: find p^{n+1} such that

$$\begin{aligned} -\operatorname{div}\left(\rho_{\text{eff}}^{-1}\nabla p^{n+1}\right) &= -(\Delta t)^{-1}\operatorname{div}\widetilde{v}_m^{n+1}, && \text{in } \Omega, \\ \rho_{\text{eff}}^{-1}\nabla p^{n+1}\cdot n &= 0, && \text{on } \Gamma_D. \end{aligned} \tag{4.11}$$

After solving (4.11), the end-of-step velocities v_s^n and v_f^n can then be directly computed from (4.10a) and (4.10b), without having to couple the solid and fluid degrees of freedom. Note that in Chorin-Temam scheme for Stokes or Navier-Stokes equations, the Poisson correction step involves only Δp^{n+1}. Here, (4.11) involves a coefficient ρ_{eff}^{-1} that depends on space and that is closely related to the values taken by the porosity. More precisely, the effective density

$$\rho_{\text{eff}} = \left(\frac{1-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right)^{-1}, \tag{4.12}$$

corresponds to the harmonic mean of the solid and fluid densities, weighted by the proportions of solid and fluid in the porous medium.

Therefore, Scheme 1 allows to fully decouple the solid, fluid and pressure degrees of freedom. Its main features are that the prediction steps consist in two unconstrained problems, the pressure step can be solved using an efficient Poisson solver, and the solid and fluid correction steps are numerically costless. Finally, let us check its consistency with respect to the initial problem. By

summing the prediction and correction steps, we get

$$\begin{cases} \dfrac{u_s^{n+1}-u_s^n}{\Delta t} = \dfrac{\widetilde{v}_s^{n+1}+v_s^n}{2}, \\ \rho_s(1-\phi)\dfrac{v_s^{n+1}-v_s^n}{\Delta t} - \operatorname{div}\left(\sigma_s\Big(\dfrac{u_s^{n+1}+u_s^n}{2}\Big)\right) \\ \qquad -\phi^2 k_f^{-1}\Big(v_f^n - \dfrac{\widetilde{v}_s^{n+1}+v_s^n}{2}\Big) + (1-\phi)\nabla p^{n+1} = \rho_s(1-\phi) f^{n+\frac{1}{2}}, \\ \rho_f\phi\dfrac{v_f^{n+1}-v_f^n}{\Delta t} - \operatorname{div}\left(\phi\,\sigma_f(\widetilde{v}_f^{n+1})\right) + \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1} - \dfrac{\widetilde{v}_s^{n+1}+v_s^n}{2}\Big) + \phi\,\nabla p^{n+1} = \rho_f\phi\, f^{n+\frac{1}{2}}, \\ \operatorname{div}\left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right) = 0, \end{cases} \tag{4.13}$$

which shows the consistency of the scheme with respect to Problem (4.1). Moreover, we see that (4.13) is very close to the backward Euler monolithic scheme (3.16). In particular, the solid midpoint discretization allows to preserve the discrete mechanical energy, as it will be detailed in the stability analysis of the next subsection. The main difference is the explicit treatment of the fluid velocity in the permeability term, that comes from the decoupling of the solid and fluid prediction steps. However, when it comes to boundary conditions, the solid displacement and the tentative velocities satisfy exactly the Dirichlet boundary conditions (4.3), whereas *a priori* the end-of-step velocities only satisfy

$$\left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right)\cdot n|_{\Gamma_D} = 0.$$

This is one of the drawbacks of the method for Dirichlet boundary conditions, which was also pointed in projection methods for fluid problems [Rannacher, 1992].

Remark 4.1. For non-homogeneous boundary conditions (4.4), the only difference is that we have to impose $u_s^{n+1}|_{\Gamma_D} = u_{s,bc}(t^{n+1})$, $\widetilde{v}_s^{n+1}|_{\Gamma_D} = v_{s,bc}(t^{n+1})$ and $\widetilde{v}_f^{n+1}|_{\Gamma_D} = v_{f,bc}(t^{n+1})$ in the prediction step, with $u_{s,bc}$, $v_{s,bc}$ and $v_{f,bc}$ satisfying the compatibility condition (4.5). Moreover, the Poisson formulation of the correction step becomes: find p^{n+1} such that $\int_\Omega p^{n+1}\,\mathrm{d}x = 0$ and

$$\begin{aligned} -\operatorname{div}\left(\rho_{\text{eff}}^{-1}\nabla p^{n+1}\right) &= -(\Delta t)^{-1}\operatorname{div}\widetilde{v}_m(t^{n+1}), && \text{in } \Omega, \\ \rho_{\text{eff}}^{-1}\nabla p^{n+1}\cdot n &= (\Delta t)^{-1} v_{m,bc}^{n+1}\cdot n, && \text{on } \partial\Omega, \end{aligned}$$

where $v_{m,bc}(t^{n+1}) = (1-\phi)\, v_{s,bc}(t^{n+1}) + \phi\, v_{f,bc}(t^{n+1})$.

Remark 4.2. If we add solid viscosity in the model, namely if $\eta > 0$, the scheme can be modified by simply adding a term $-\operatorname{div}\left(\sigma_s^{\text{vis}}\Big(\frac{\widetilde{v}_s^{n+1}+v_s^n}{2}\Big)\right)$ in the solid prediction step (4.8a). If we consider the compressible or nearly-incompressible case $\kappa < +\infty$ and $\alpha < 1$, the prediction step remains the same but one has to change the correction step as follows: find v_s^{n+1}, v_f^{n+1} and p^{n+1} such that

$$\begin{cases} \rho_s(1-\phi)\dfrac{v_s^{n+1}-\widetilde{v}_s^{n+1}}{\Delta t} + (\alpha-\phi)\,\nabla p^{n+1} = 0, \\ \rho_f\phi\dfrac{v_f^{n+1}-\widetilde{v}_f^{n+1}}{\Delta t} + \phi\,\nabla p^{n+1} = 0, \\ \dfrac{\alpha-\phi}{\kappa}\dfrac{p^{n+1}-p^n}{\Delta t} + \operatorname{div}\left((\alpha-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right) = 0. \end{cases}$$

From the Darcy-formulation above we can retrieve as before a Poisson formulation, that reads

$$(\Delta t)^{-1}\frac{\alpha-\phi}{\kappa}\frac{p^{n+1}-p^n}{\Delta t} - \operatorname{div}\left(\left(\frac{\alpha-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right)\nabla p^{n+1}\right) = -(\Delta t)^{-1}\operatorname{div}\widetilde{v}_m^{n+1},$$

and that has to be complemented with the boundary conditions imposed on the pressure. This formulation is close to (4.11), but involves the pressure discrete derivative and an effective density $\rho_{\text{eff}}^{\kappa} = \left(\frac{\alpha-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right)^{-1}$ that depends on the Biot coefficient α.

For incompressible fluid problems, there exists two versions of projection schemes: the *non-incremental* one, which corresponds to the original scheme introduced by Chorin and Temam [Chorin, 1969; Temam, 1969], and the *incremental* one first proposed by [Goda, 1979], which is known to improve the convergence rate in time of the pressure. Assuming that the interstitial pressure is regular enough, we can propose such an incremental variant of Scheme 1. To do so, the key idea is to take into account the pressure gradients from the previous time step during the prediction step, and to modify the correction step accordingly. This improves the approximation of the tentative velocities by being as close as possible to (4.1b) and (4.1c), and hence the pressure approximation. The corresponding algorithm is presented in Scheme 2.

Scheme 2 Explicit treatment of permeability (incremental version)

Step 1: `(prediction step)`

– **Step 1.1**: `(structure prediction sub-step)`
Find u_s^{n+1} and $\widetilde{v}_s^{n+1}$ such that $u_s^{n+1}|_{\Gamma_D} = 0$, $\widetilde{v}_s^{n+1}|_{\Gamma_D} = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{\widetilde{v}_s^{n+1} - v_s^n}{\Delta t} - \operatorname{div}\left(\sigma_s\Big(\dfrac{u_s^{n+1}+u_s^n}{2}\Big)\right) \\ \qquad - \phi^2 k_f^{-1}(v_f^n - \widetilde{v}_s^{n+1}) + (1-\phi)\nabla p^n = \rho_s(1-\phi) f^{n+\frac{1}{2}}, & (4.14a) \\ \dfrac{u_s^{n+1} - u_s^n}{\Delta t} = \widetilde{v}_s^{n+1}. & (4.14b) \end{cases}$$

– **Step 1.2**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1}$ such that $\widetilde{v}_f^{n+1}|_{\Gamma_D} = 0$ and

$$\begin{cases} \rho_f\phi\dfrac{\widetilde{v}_f^{n+1} - v_f^n}{\Delta t} - \operatorname{div}\left(\phi\,\sigma_f(\widetilde{v}_f^{n+1})\right) \\ \qquad + \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+1}) + \phi\,\nabla p^n = \rho_f\phi\, f^{n+\frac{1}{2}}. & (4.15a) \end{cases}$$

Step 2: `(correction step)`
Find v_s^{n+1}, v_f^{n+1} and p^{n+1} such that $\int_\Omega p^{n+1}\,\mathrm{d}x = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{v_s^{n+1} - \widetilde{v}_s^{n+1}}{\Delta t} + (1-\phi)\,\nabla(p^{n+1} - p^n) = 0, & (4.16a) \\ \rho_f\phi\dfrac{v_f^{n+1} - \widetilde{v}_f^{n+1}}{\Delta t} + \phi\,\nabla(p^{n+1} - p^n) = 0, & (4.16b) \\ \operatorname{div}\left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right) = 0, & (4.16c) \\ \left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right)\cdot n|_{\Gamma_D} = 0. & (4.16d) \end{cases}$$

Here again, the correction step (4.16) can be reformulated as a Poisson problem, by simply replacing p^{n+1} by the pressure increment $p^{n+1}-p^n$ in (4.11). Note that for stability reasons that will be made clearer in the next subsection, the midpoint velocity $\frac{\widetilde{v}_s^{n+1}+v_s^n}{2}$ has been replaced by $\widetilde{v}_s^{n+1}$ in the prediction step.

Scheme 2 is very close to the one proposed in [Markert et al., 2009]. Indeed, the only differences are that in [Markert et al., 2009], the fluid viscosity effects considered in (4.15) are neglected following a magnitude argument given in [Markert, 2007], and that the solid displacement is treated explicitly in (4.14a). However, [Markert et al., 2009] does not include a stability analysis of the scheme, which we are now going to carry out.

4.2.2 Stability analysis

We start with the non-incremental version of the scheme. Let us introduce the notation

$$u_s^{n+\frac{1}{2}} = \frac{u_s^{n+1}+u_s^n}{2},$$

for the solid displacement midpoint velocity, and

$$\widetilde{v}_s^{n+\frac{1}{2}\sharp} = \frac{\widetilde{v}_s^{n+1}+v_s^n}{2}, \tag{4.17}$$

to denote the specific midpoint velocity appearing in the prediction steps. Moreover, we will denote by $\mathrm{H}^1(\Omega)/\mathbb{R}$ the subspace of $\mathrm{H}^1(\Omega)$ composed of functions with zero mean value, namely

$$\mathrm{H}^1(\Omega)/\mathbb{R} = \left\{ p \in \mathrm{H}^1(\Omega)\,,\ \int_\Omega p\,\mathrm{d}x = 0 \right\}.$$

For the stability analysis, we assume that no external body force is applied to the porous medium, namely $f=0$. Note nonetheless that the forthcoming stability analysis can easily be extended to the case where $f \neq 0$. The weak formulation associated with Scheme 1 then reads:

Step 1: `(prediction step)`

– **Step 1.1**: `(structure prediction sub-step)`
Find $u_s^{n+1} \in [\mathrm{H}_0^1(\Omega)]^d$ and $\widetilde{v}_s^{n+1} \in [\mathrm{H}_0^1(\Omega)]^d$ such that

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{\widetilde{v}_s^{n+1}-v_s^n}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s\Big(\frac{u_s^{n+1}-u_s^n}{\Delta t}\Big) : \varepsilon(d_s)\,\mathrm{d}x \\ - \int_\Omega \sigma_s(\widetilde{v}_s^{n+\frac{1}{2}\sharp}) : \varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s^{n+\frac{1}{2}}) : \varepsilon(w_s)\,\mathrm{d}x \\ - \int_\Omega \phi^2 k_f^{-1}(v_f^n - \widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot w_s\,\mathrm{d}x = 0, \end{aligned} \tag{4.18}$$

for all $w_s \in [\mathrm{H}_0^1(\Omega)]^d$ and $d_s \in [\mathrm{H}_0^1(\Omega)]^d$.

– **Step 1.2**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1} \in [\mathrm{H}_0^1(\Omega)]^d$ such that

$$\begin{aligned}\int_\Omega \rho_f\phi\frac{\widetilde{v}_f^{n+1}-v_f^n}{\Delta t}\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(w_f)\,\mathrm{d}x \\ + \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot w_f\,\mathrm{d}x = 0, \end{aligned} \tag{4.19}$$

for all $w_f \in [\mathrm{H}_0^1(\Omega)]^d$.

Step 2: `(correction step)`
Find $v_s^{n+1} \in [\mathrm{L}^2(\Omega)]^d$, $v_f^{n+1} \in [\mathrm{L}^2(\Omega)]^d$ and $p^{n+1} \in \mathrm{H}^1(\Omega)/\mathbb{R}$ such that

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{v_s^{n+1}-\widetilde{v}_s^{n+1}}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \rho_f\phi\frac{v_f^{n+1}-\widetilde{v}_f^{n+1}}{\Delta t}\cdot w_f\,\mathrm{d}x \\ + \int_\Omega (1-\phi)\nabla p^{n+1}\cdot w_s\,\mathrm{d}x + \int_\Omega \phi\,\nabla p^{n+1}\cdot w_f\,\mathrm{d}x \\ - \int_\Omega (1-\phi)v_s^{n+1}\cdot\nabla q\,\mathrm{d}x - \int_\Omega \phi\,v_f^{n+1}\cdot\nabla q\,\mathrm{d}x = 0, \quad (4.20)\end{aligned}$$

for all $w_s \in [\mathrm{L}^2(\Omega)]^d$, $w_f \in [\mathrm{L}^2(\Omega)]^d$ and $q \in \mathrm{H}^1(\Omega)/\mathbb{R}$.

Remark 4.3. An alternative way of formulating weakly the correction step is to use a mixed formulation in the space H_ϕ. From a numerical point of view, such a formulation requires to use an approximation of the space H_ϕ, which may be done with Raviart-Thomas elements. Yet, it is not clear whether this strategy can lead to error estimates that are robust with respect to the porosity field ϕ.

The stability analysis is based on a discrete energy balance for each step of the algorithm. Let us define the discrete kinetic and mechanical energies by

$$\mathcal{E}_c^n = \frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|v_s^n|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,|v_f^n|^2\,\mathrm{d}x \quad \text{and} \quad \mathcal{E}_m^n = \frac{1}{2}\int_\Omega \sigma_s(u_s^n):\varepsilon(u_s^n)\,\mathrm{d}x.$$

Moreover, we introduce the tentative kinetic energy associated with the prediction velocities, namely

$$\widetilde{\mathcal{E}}_c^{n+1} = \frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|\widetilde{v}_s^{n+1}|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,\left|\widetilde{v}_f^{n+1}\right|^2\,\mathrm{d}x.$$

In what follows, we will often use the algebraic identity

$$\big(a^{n+1}-a^n\big)a^{n+1} = \frac{1}{2}\big(a^{n+1}\big)^2 - \frac{1}{2}\big(a^n\big)^2 + \frac{\big(a^{n+1}-a^n\big)^2}{2}. \quad (4.21)$$

For the prediction step, we choose $w_s = \Delta t\,\widetilde{v}_s^{n+\frac{1}{2}\sharp}$, $d_s = \Delta t\,u_s^{n+\frac{1}{2}}$ and $w_f = \Delta t\,\widetilde{v}_f^{n+1}$ in (4.18) and (4.19) to obtain

$$\begin{aligned}\big(\widetilde{\mathcal{E}}_c^{n+1}-\mathcal{E}_c^n\big) + \big(\mathcal{E}_m^{n+1}-\mathcal{E}_m^n\big) + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1}-v_f^n\right|^2\mathrm{d}x + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}):\varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x \\ - \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(v_f^n-\widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)\cdot\widetilde{v}_f^{n+1}\,\mathrm{d}x = 0.\end{aligned}$$

The explicit part of the friction term requires a special treatment. Writing $v_f^n = \widetilde{v}_f^{n+1} + (v_f^n - \widetilde{v}_f^{n+1})$, it can be decomposed as

$$\begin{aligned}-\phi^2 k_f^{-1}\Big(v_f^n-\widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp} + \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)\cdot\widetilde{v}_f^{n+1} \\ = \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp})^2 - \phi^2 k_f^{-1}\big(v_f^n-\widetilde{v}_f^{n+1}\big)\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp},\end{aligned}$$

so that the prediction step energy balance reads:

$$\begin{aligned}\big(\widetilde{\mathcal{E}}_c^{n+1}-\mathcal{E}_c^n\big) + \big(\mathcal{E}_m^{n+1}-\mathcal{E}_m^n\big) + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1}-v_f^n\right|^2\mathrm{d}x + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}):\varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x \\ + \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)^2\mathrm{d}x = \Delta t\int_\Omega \phi^2 k_f^{-1}\big(v_f^n-\widetilde{v}_f^{n+1}\big)\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}x. \quad (4.22)\end{aligned}$$

Testing (4.60) with $w_s = \Delta t\, v_s^{n+1}$, $w_f = \Delta t\, v_f^{n+1}$ and $q = \Delta t\, p^{n+1}$, we get the correction step energy balance

$$\left(\mathcal{E}_c^{n+1} - \widetilde{\mathcal{E}}_c^{n+1}\right) + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|v_s^{n+1} - \widetilde{v}_s^{n+1}\right|^2 \mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^{n+1} - \widetilde{v}_f^{n+1}\right|^2 \mathrm{d}x = 0. \tag{4.23}$$

Summing up the prediction and correction steps contributions (4.22) and (4.23), the tentative kinetic energy $\widetilde{\mathcal{E}}_c^{n+1}$ simplifies and it follows that

$$\begin{aligned}\left(\mathcal{E}^{n+1} - \mathcal{E}^n\right) &+ \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1} - v_f^n\right|^2 \mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|v_s^{n+1} - \widetilde{v}_s^{n+1}\right|^2 \mathrm{d}x \\ &+ \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^{n+1} - \widetilde{v}_f^{n+1}\right|^2 \mathrm{d}x + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x \\ &+ \Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\right)^2 \mathrm{d}x = \Delta t\int_\Omega \phi^2 k_f^{-1}\left(v_f^n - \widetilde{v}_f^{n+1}\right)\cdot \widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}x,\end{aligned} \tag{4.24}$$

where $\mathcal{E}^n$ corresponds to the total energy of the system, namely

$$\mathcal{E}^n = \mathcal{E}_c^n + \mathcal{E}_m^n = \frac{1}{2}\int_\Omega \rho_s(1-\phi)\,|v_s^n|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,|v_f^n|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \sigma_s(u_s^n) : \varepsilon(u_s^n)\,\mathrm{d}x.$$

Remark 4.4. Because of the midpoint discretization chosen for the solid, no numerical dissipation appears in the prediction step for the structure part. However, (4.24) includes a solid numerical dissipation term $\frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2 \mathrm{d}x$ that comes from the correction step. It is possible to get rid of this term by changing the mixture divergence constraint (4.10c) into

$$\operatorname{div}\left((1-\phi)\frac{v_s^{n+1} + \widetilde{v}_s^{n+1}}{2} + \phi\, v_f^{n+1}\right) = 0,$$

and by testing (4.60) with $w_s = \frac{v_s^{n+1} + \widetilde{v}_s^{n+1}}{2}$ instead of $w_s = v_s^{n+1}$. Nonetheless, we will see in the proof of Theorem 4.5 below that this solid numerical dissipation term is useful to control unsigned terms coming from the explicit treatment of the permeability.

With the discrete energy balance (4.24) in hand, we are now ready to establish the following stability result.

Theorem 4.5. *Let u_s^n, v_s^n, v_f^n, p^n, $\widetilde{v}_s^n$ and $\widetilde{v}_f^n$ satisfy Scheme 1 with $f = 0$. If the time step verifies*

$$\Delta t^2 < \frac{\rho_f\rho_s(1-\phi_{\max})}{2\phi_{\max}^3 (k_{\max}^{-1})^2}, \tag{4.25}$$

then for all $0 \le N \le n_T$, it holds

$$\begin{aligned}\mathcal{E}^N &+ \frac{3}{8}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2 \mathrm{d}x + \frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1} - v_f^{n+1}\right|^2 \mathrm{d}x \\ &+ \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\right)^2 \mathrm{d}x \le C_T\mathcal{E}^0,\end{aligned} \tag{4.26}$$

with $C_T > 0$ a constant independent of Δt.

Proof. The idea is to control the right-hand side of (4.24) thanks to the solid and fluid numerical dissipation. To do so, recalling (4.17), we decompose it as

$$\Delta t\int_\Omega \phi^2 k_f^{-1}\left(v_f^n - \widetilde{v}_f^{n+1}\right)\cdot\frac{1}{2}\left[\left(\widetilde{v}_s^{n+1} - v_s^{n+1}\right) + \left(v_s^{n+1} + v_s^n\right)\right]\mathrm{d}x. \tag{4.27}$$

Then, we use Young inequality on each part of the above decomposition to obtain

$$\begin{aligned}\Delta t \int_\Omega \phi^2 k_f^{-1}\big(v_f^n - \widetilde{v}_f^{n+1}\big)\cdot\big(\widetilde{v}_s^{n+1} - v_s^{n+1}\big)\,\mathrm{d}x \\ \leq \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^n - \widetilde{v}_f^{n+1}\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \frac{\Delta t^2\phi^4}{\rho_f\phi\rho_s(1-\phi)}\rho_s(1-\phi)k_f^{-2}\Big(\widetilde{v}_s^{n+1} - v_s^{n+1}\Big)^2\mathrm{d}x \\ \leq \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^n - \widetilde{v}_f^{n+1}\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \Delta t^2\frac{\phi_{\max}^3\big(k_{\max}^{-1}\big)^2}{\rho_f\rho_s(1-\phi_{\max})}\rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2\mathrm{d}x,\end{aligned} \tag{4.28}$$

where we used the new assumption (4.2) made on the permeability tensor. Likewise, we have

$$\begin{aligned}\Delta t \int_\Omega \phi^2 k_f^{-1}\big(v_f^n - \widetilde{v}_f^{n+1}\big)\cdot\big(v_s^{n+1} + v_s^n\big)\,\mathrm{d}x \\ \leq \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^n - \widetilde{v}_f^{n+1}\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \Delta t^2\frac{\phi_{\max}^3(k_{\max}^{-1})^2}{\rho_f\rho_s(1-\phi_{\max})}\rho_s(1-\phi)\left|v_s^{n+1} + v_s^n\right|^2\mathrm{d}x.\end{aligned} \tag{4.29}$$

Taking the half-sum of (4.28) and (4.29) and using that $\left|v_s^{n+1} + v_s^n\right|^2 \leq 2\big(\left|v_s^{n+1}\right|^2 + |v_s^n|^2\big)$, it follows that

$$\begin{aligned}\Delta t \int_\Omega \phi^2 k_f^{-1}\big(v_f^n - \widetilde{v}_f^{n+1}\big)\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}x \leq \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^n - \widetilde{v}_f^{n+1}\right|^2\mathrm{d}x \\ + \frac{1}{4}\int_\Omega \Delta t^2\frac{\phi_{\max}^3\big(k_{\max}^{-1}\big)^2}{\rho_f\rho_s(1-\phi_{\max})}\rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2\mathrm{d}x \\ + \frac{1}{2}\int_\Omega \Delta t^2\frac{\phi_{\max}^3(k_{\max}^{-1})^2}{\rho_f\rho_s(1-\phi_{\max})}\rho_s(1-\phi)\big(\left|v_s^{n+1}\right|^2 + |v_s^n|^2\big)\,\mathrm{d}x.\end{aligned} \tag{4.30}$$

Setting

$$C = \frac{\phi_{\max}^3(k_{\max}^{-1})^2}{\rho_f\rho_s(1-\phi_{\max})}, \tag{4.31}$$

the time step restriction (4.25) is equivalent to $C\Delta t^2 < 1/2$. Therefore, we observe that

$$\begin{aligned}\frac{1}{4}\int_\Omega \Delta t^2\frac{\phi_{\max}^3\big(k_{\max}^{-1}\big)^2}{\rho_f\rho_s(1-\phi_{\max})}\rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2\mathrm{d}x = \frac{C\Delta t^2}{4}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2\mathrm{d}x \\ \leq \frac{1}{8}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2\mathrm{d}x.\end{aligned}$$

Consequently, plugging (4.30) into the discrete energy balance (4.24) yields

$$\begin{aligned}\big(\mathcal{E}^{n+1} - \mathcal{E}^n\big) + \frac{3}{8}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1} - v_f^{n+1}\right|^2\mathrm{d}x \\ + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)^2\mathrm{d}x \\ \leq \frac{1}{2}\int_\Omega \Delta t^2\frac{\phi_{\max}^3(k_{\max}^{-1})^2}{\rho_f\rho_s(1-\phi_{\max})}\rho_s(1-\phi)\big(\left|v_s^{n+1}\right|^2 + |v_s^n|^2\big)\,\mathrm{d}x \leq C\Delta t^2\big(\mathcal{E}^{n+1} + \mathcal{E}^n\big).\end{aligned}$$

Summing this estimate between 0 and $N-1$, we deduce

$$\mathcal{E}^N + \frac{3}{8}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1}-v_s^{n+1}\right|^2\,\mathrm{d}x + \frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1}-v_f^{n+1}\right|^2\,\mathrm{d}x$$
$$+\Delta t\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}):\varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)^2\,\mathrm{d}x$$
$$\leq \mathcal{E}^0 + C\Delta t^2\sum_{n=0}^{N-1}\big(\mathcal{E}^{n+1}+\mathcal{E}^n\big) \leq \mathcal{E}^0 + 2C\Delta t^2\sum_{n=0}^{N}\mathcal{E}^n,$$

Since $2C\Delta t^2 < 1$, the conclusion then follows from discrete Grönwall Lemma 3.20. □

Remark 4.6. In the viscous case $\eta > 0$, an extra term

$$\Delta t\int_\Omega \sigma_s^{\mathrm{vis}}\big(\widetilde{v}_s^{n+\frac{1}{2}\sharp}\big):\varepsilon\big(\widetilde{v}_s^{n+\frac{1}{2}\sharp}\big)\,\mathrm{d}x,$$

appears in the left-hand side of (4.24). This viscous term can then be used to control the unsigned permeability term $\Delta t\int_\Omega \phi^2 k_f^{-1}\big(v_f^n-\widetilde{v}_f^{n+1}\big)\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}x$ without using the decomposition (4.27).

A similar result can be obtained for the incremental version of the scheme, whose weak form in absence of external body forces reads:

Step 1: `(prediction step)`

– **Step 1.1**: `(structure prediction sub-step)`
Find $u_s^{n+1}\in[\mathrm{H}_0^1(\Omega)]^d$ and $\widetilde{v}_s^{n+1}\in[\mathrm{H}_0^1(\Omega)]^d$ such that

$$\int_\Omega \rho_s(1-\phi)\frac{\widetilde{v}_s^{n+1}-v_s^n}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s\Big(\frac{u_s^{n+1}-u_s^n}{\Delta t}\Big):\varepsilon(d_s)\,\mathrm{d}x$$
$$-\int_\Omega \sigma_s(\widetilde{v}_s^{n+1}):\varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s^{n+\frac{1}{2}}):\varepsilon(w_s)\,\mathrm{d}x$$
$$-\int_\Omega \phi^2 k_f^{-1}(v_f^n-\widetilde{v}_s^{n+1})\cdot w_s\,\mathrm{d}x + \int_\Omega (1-\phi)\nabla p^n\cdot w_s\,\mathrm{d}x = 0, \quad (4.32)$$

for all $w_s\in[\mathrm{H}_0^1(\Omega)]^d$ and $d_s\in[\mathrm{H}_0^1(\Omega)]^d$.

– **Step 1.2**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1}\in[\mathrm{H}_0^1(\Omega)]^d$ such that

$$\int_\Omega \rho_f\phi\frac{\widetilde{v}_f^{n+1}-v_f^n}{\Delta t}\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}):\varepsilon(w_f)\,\mathrm{d}x$$
$$+\int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+1})\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\nabla p^n\cdot w_f\,\mathrm{d}x = 0, \quad (4.33)$$

for all $w_f\in[\mathrm{H}_0^1(\Omega)]^d$.

Step 2: `(correction step)`
Find $v_s^{n+1} \in [\mathrm{L}^2(\Omega)]^d$, $v_f^{n+1} \in [\mathrm{L}^2(\Omega)]^d$ and $p^{n+1} \in \mathrm{H}^1(\Omega)/\mathbb{R}$ such that

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{v_s^{n+1}-\widetilde{v}_s^{n+1}}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \rho_f\phi\frac{v_f^{n+1}-\widetilde{v}_f^{n+1}}{\Delta t}\cdot w_f\,\mathrm{d}x \\ + \int_\Omega (1-\phi)\nabla(p^{n+1}-p^n)\cdot w_s\,\mathrm{d}x + \int_\Omega \phi\,\nabla(p^{n+1}-p^n)\cdot w_f\,\mathrm{d}x \\ - \int_\Omega (1-\phi)v_s^{n+1}\cdot\nabla q\,\mathrm{d}x - \int_\Omega \phi\,v_f^{n+1}\cdot\nabla q\,\mathrm{d}x = 0, \quad (4.34)\end{aligned}$$

for all $w_s \in [\mathrm{L}^2(\Omega)]^d$, $w_f \in [\mathrm{L}^2(\Omega)]^d$ and $q \in \mathrm{H}^1(\Omega)/\mathbb{R}$.

Theorem 4.7. *Let u_s^n, v_s^n, v_f^n, p^n, $\widetilde{v}_s^n$ and $\widetilde{v}_f^n$ satisfy Scheme 2, and assume that $p^0 \in \mathrm{H}^1(\Omega)/\mathbb{R}$. If the time step verifies (4.25), then for all $0 \le N \le n_T$, it holds*

$$\begin{aligned}\mathcal{E}^N + \frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla p^N\right|^2\,\mathrm{d}x + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}):\varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x \\ + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+1}\right)^2\mathrm{d}x \le C_T\left(\mathcal{E}^0 + \frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla p^0\right|^2\,\mathrm{d}x\right), \quad (4.35)\end{aligned}$$

with $C_T > 0$ a constant independent of Δt.

Proof. Choosing $d_s = \Delta t\, u_s^{n+\frac{1}{2}}$, $w_s = \Delta t\,\widetilde{v}_s^{n+1}$ and $w_f = \Delta t\,\widetilde{v}_f^{n+1}$ in (4.32) and (4.33), we have

$$\begin{aligned}\left(\widetilde{\mathcal{E}}_c^{n+1}-\mathcal{E}_c^n\right) + \left(\mathcal{E}_m^{n+1}-\mathcal{E}_m^n\right) + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1}-v_s^n\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1}-v_f^n\right|^2\mathrm{d}x \\ + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}):\varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+1}\right)^2\mathrm{d}x \\ + \Delta t\int_\Omega (1-\phi)\nabla p^n\cdot\widetilde{v}_s^{n+1}\,\mathrm{d}x + \Delta t\int_\Omega \phi\,\nabla p^n\cdot\widetilde{v}_f^{n+1}\,\mathrm{d}x \\ = \Delta t\int_\Omega \phi^2 k_f^{-1}\left(v_f^n-\widetilde{v}_f^{n+1}\right)\cdot\widetilde{v}_s^{n+1}\,\mathrm{d}x. \quad (4.36)\end{aligned}$$

Choosing $w_s = \Delta t\, v_s^{n+1}$, $w_f = \Delta t\, v_f^{n+1}$ and $q = \Delta t\, p^{n+1}$ in the correction step (4.34), the terms of the form $\Delta t\int_\Omega(1-\phi)\nabla p^{n+1}\cdot v_s^{n+1}\,\mathrm{d}x$ and $\Delta t\int_\Omega \phi\,\nabla p^{n+1}\cdot v_f^{n+1}\,\mathrm{d}x$ cancel out, so that we obtain

$$\begin{aligned}\left(\mathcal{E}_c^{n+1}-\widetilde{\mathcal{E}}_c^{n+1}\right) + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|v_s^{n+1}-\widetilde{v}_s^{n+1}\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^{n+1}-\widetilde{v}_f^{n+1}\right|^2\mathrm{d}x \\ - \Delta t\int_\Omega (1-\phi)\nabla p^n\cdot v_s^{n+1}\,\mathrm{d}x - \Delta t\int_\Omega \phi\,\nabla p^n\cdot v_f^{n+1}\,\mathrm{d}x = 0. \quad (4.37)\end{aligned}$$

Summing (4.36) and (4.37), we infer

$$\begin{aligned}\left(\mathcal{E}^{n+1}-\mathcal{E}^n\right) + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1}-v_s^n\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1}-v_f^n\right|^2\mathrm{d}x \\ + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|v_s^{n+1}-\widetilde{v}_s^{n+1}\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^{n+1}-\widetilde{v}_f^{n+1}\right|^2\mathrm{d}x \\ + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}):\varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+1}\right)^2\mathrm{d}x \\ + \Delta t\int_\Omega (1-\phi)\nabla p^n\cdot(\widetilde{v}_s^{n+1}-v_s^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi\,\nabla p^n\cdot(\widetilde{v}_f^{n+1}-v_f^{n+1})\,\mathrm{d}x \\ = \Delta t\int_\Omega \phi^2 k_f^{-1}\left(v_f^n-\widetilde{v}_f^{n+1}\right)\cdot\widetilde{v}_s^{n+1}\,\mathrm{d}x. \quad (4.38)\end{aligned}$$

The major difference between (4.38) and the energy balance of the non-incremental scheme (4.24) is the extra pressure terms coming from the explicit treatment of pressure, and the additional solid numerical dissipation term coming from the replacement of $\widetilde{v}_s^{n+\frac{1}{2}\sharp}$ by $\widetilde{v}_s^{n+1}$ in the prediction step.

To control the pressure terms, we test (4.34) with $w_s = \frac{\Delta t^2}{\rho_s}\nabla p^n$, $w_f = \frac{\Delta t^2}{\rho_f}\nabla p^n$ and $q = 0$ to retrieve

$$\Delta t \int_\Omega (1-\phi)(v_s^{n+1} - \widetilde{v}_s^{n+1}) \cdot \nabla p^n \,\mathrm{d}x + \Delta t \int_\Omega \phi \,\nabla p^n \cdot (v_f^{n+1} - \widetilde{v}_f^{n+1}) \,\mathrm{d}x \\ + \Delta t^2 \int_\Omega \left(\frac{1-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right) \nabla(p^{n+1} - p^n) \cdot \nabla p^n \,\mathrm{d}x = 0.$$

Since $\nabla(p^{n+1} - p^n) \cdot \nabla p^n = \frac{1}{2}\Big(\big|\nabla p^{n+1}\big|^2 - |\nabla p^n|^2 - \big|\nabla(p^{n+1} - p^n)\big|^2\Big)$, it follows that

$$\Delta t \int_\Omega (1-\phi) \nabla p^n \cdot (\widetilde{v}_s^{n+1} - v_s^{n+1}) \,\mathrm{d}x + \Delta t \int_\Omega \phi \,\nabla p^n \cdot (\widetilde{v}_f^{n+1} - v_f^{n+1}) \,\mathrm{d}x \\ = \frac{\Delta t^2}{2} \int_\Omega \left(\frac{1-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right) \big|\nabla p^{n+1}\big|^2 \,\mathrm{d}x - \frac{\Delta t^2}{2} \int_\Omega \left(\frac{1-\phi}{\rho_s} + \frac{\phi}{\rho_f}\right) |\nabla p^n|^2 \,\mathrm{d}x \\ - \frac{\Delta t^2}{2} \int_\Omega \frac{1-\phi}{\rho_s} \big|\nabla(p^{n+1} - p^n)\big|^2 \,\mathrm{d}x - \frac{\Delta t^2}{2} \int_\Omega \frac{\phi}{\rho_f} \big|\nabla(p^{n+1} - p^n)\big|^2 \,\mathrm{d}x.$$

But from (4.16a), we know that

$$\frac{\Delta t^2}{2} \left\| \sqrt{\frac{1-\phi}{\rho_s}} \nabla(p^{n+1} - p^n) \right\|^2 = \frac{\Delta t^2}{2} \left\| \sqrt{\rho_s(1-\phi)} \frac{v_s^{n+1} - \widetilde{v}_s^{n+1}}{\Delta t} \right\|^2 = \frac{1}{2} \int_\Omega \rho_s (1-\phi) \big|v_s^{n+1} - \widetilde{v}_s^{n+1}\big|^2 \,\mathrm{d}x. \tag{4.39}$$

Proceeding similarly for the fluid correction step, we find that

$$\frac{\Delta t^2}{2} \left\| \sqrt{\frac{\phi}{\rho_f}} \nabla(p^{n+1} - p^n) \right\|^2 = \frac{1}{2} \int_\Omega \rho_f \phi \left| v_f^{n+1} - \widetilde{v}_f^{n+1} \right|^2 \mathrm{d}x. \tag{4.40}$$

Combining (4.38) with (4.39) and (4.40), the numerical dissipation terms from the correction step vanish and we deduce

$$\big(\mathcal{E}^{n+1} - \mathcal{E}^n\big) + \frac{1}{2} \int_\Omega \rho_s (1-\phi) \big|\widetilde{v}_s^{n+1} - v_s^n\big|^2 \,\mathrm{d}x + \frac{1}{2} \int_\Omega \rho_f \phi \left|\widetilde{v}_f^{n+1} - v_f^n\right|^2 \mathrm{d}x \\ + \Delta t \int_\Omega \phi \,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1}) \,\mathrm{d}x + \Delta t \int_\Omega \phi^2 k_f^{-1} \big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+1}\big)^2 \,\mathrm{d}x \\ + \frac{\Delta t^2}{2} \left(\int_\Omega \rho_{\mathrm{eff}}^{-1} \big|\nabla p^{n+1}\big|^2 \,\mathrm{d}x - \int_\Omega \rho_{\mathrm{eff}}^{-1} |\nabla p^n|^2 \,\mathrm{d}x \right) = \Delta t \int_\Omega \phi^2 k_f^{-1} \big(v_f^n - \widetilde{v}_f^{n+1}\big) \cdot \widetilde{v}_s^{n+1} \,\mathrm{d}x, \tag{4.41}$$

where we recall that $\rho_{\mathrm{eff}}^{-1} = \frac{1-\phi}{\rho_s} + \frac{\phi}{\rho_f}$ as in the Poisson problem (4.11).

The end of the proof is similar to that of Theorem 4.5. Defining C as in (4.31), Young inequality implies that

$$\Delta t \int_\Omega \phi^2 k_f^{-1} \big(v_f^n - \widetilde{v}_f^{n+1}\big) \cdot \widetilde{v}_s^{n+1} \,\mathrm{d}x \\ = \Delta t \int_\Omega \phi^2 k_f^{-1} \big(v_f^n - \widetilde{v}_f^{n+1}\big) \cdot (\widetilde{v}_s^{n+1} - v_s^n) \,\mathrm{d}x + \Delta t \int_\Omega \phi^2 k_f^{-1} \big(v_f^n - \widetilde{v}_f^{n+1}\big) \cdot v_s^n \,\mathrm{d}x \\ \leq \frac{1}{2} \int_\Omega \rho_f \phi \left| v_f^n - \widetilde{v}_f^{n+1} \right|^2 \mathrm{d}x + C\Delta t^2 \int_\Omega \rho_s (1-\phi) \big|\widetilde{v}_s^{n+1} - v_s^n\big|^2 \,\mathrm{d}x + C\Delta t^2 \int_\Omega \rho_s (1-\phi) |v_s^n|^2 \,\mathrm{d}x.$$

Plugging this result into (4.41) and recalling that $C\Delta t^2 < 1/2$, we finally get

$$\left(\mathcal{E}^{n+1} - \mathcal{E}^n\right) + \Delta t \int_\Omega \phi\, \sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t \int_\Omega \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+1}\right)^2 \mathrm{d}x + \frac{\Delta t^2}{2}\left(\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla p^{n+1}\right|^2 \mathrm{d}x - \int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla p^{n}\right|^2 \mathrm{d}x\right) \leq 2C\Delta t^2 \mathcal{E}^n,$$

and we conclude by an application of discrete Grönwall Lemma. □

Remark 4.8. At the end of the proof, we see that the solid numerical dissipation term coming from the prediction step is used to control the permeability unsigned term. This is a reason why we have replaced the midpoint velocity $\widetilde{v}_s^{n+\frac{1}{2}\sharp}$ by $\widetilde{v}_s^{n+1}$ in Scheme 2.

Remark 4.9. Note that the stability of the incremental version of the scheme requires more regularity on the pressure than the non-incremental version since (4.35) requires that $p(0) \in \mathrm{H}^1(\Omega)$.

Therefore, we have proved stability estimates for both the non-incremental and incremental versions of the projection scheme. However, the estimates of Theorems 4.5 and 4.7 are subject to the time step restriction (4.25), namely

$$\Delta t^2 < \frac{\rho_f \rho_s (1-\phi_{\max})}{2\phi_{\max}^3 (k_{\max}^{-1})^2}.$$

If the permeability tensor is small, *i.e.* if $k_{\max}^{-1}$ is large, this condition can be *very restrictive* in practice. As a matter of fact, in biological applications, k_f typically takes values between 10^{-9} and $10^{-12}\,\mathrm{m}^2\,\mathrm{Pa}^{-1}\,\mathrm{s}^{-1}$, so that for realistic density values – say $\rho_s = \rho_f = 10^3\,\mathrm{kg\,m}^{-3}$ – the time step condition (4.25) becomes $\Delta t \lesssim 10^{-6}\,\mathrm{s}$. In order to overcome this time step restriction, we are now going to present two other variants of the scheme.

4.2.3 Other treatments of permeability

The time step condition (4.25) comes from the explicit treatment of the fluid velocity in the permeability term during the solid prediction sub-step (4.8a). It is possible to treat implicitly this term by recoupling the solid and fluid prediction sub-steps, leading to the following scheme.

Scheme 3 Implicit treatment of permeability (non-incremental version)

Step 1: `(prediction step)`
Find u_s^{n+1}, $\widetilde{v}_s^{n+1}$ and $\widetilde{v}_f^{n+1}$ such that $u_s^{n+1}|_{\Gamma_D} = \widetilde{v}_s^{n+1}|_{\Gamma_D} = \widetilde{v}_f^{n+1}|_{\Gamma_D} = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{\widetilde{v}_s^{n+1} - v_s^n}{\Delta t} - \mathrm{div}\left(\sigma_s(u_s^{n+\frac{1}{2}})\right) \\ \qquad\qquad - \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\right) = \rho_s(1-\phi) f^{n+\frac{1}{2}}, \\ \rho_f \phi \dfrac{\widetilde{v}_f^{n+1} - v_f^n}{\Delta t} - \mathrm{div}\left(\phi\, \sigma_f(\widetilde{v}_f^{n+1})\right) + \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\right) = \rho_f \phi\, f^{n+\frac{1}{2}}, \\ \dfrac{u_s^{n+1} - u_s^n}{\Delta t} = \widetilde{v}_s^{n+\frac{1}{2}\sharp}. \end{cases}$$

Step 2: `(correction step)`
Find v_s^{n+1}, v_f^{n+1} and p^{n+1} such that $\int_\Omega p^{n+1}\,\mathrm{d}x = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{v_s^{n+1}-\widetilde{v}_s^{n+1}}{\Delta t} + (1-\phi)\,\nabla p^{n+1} = 0, \\ \rho_f\phi\dfrac{v_f^{n+1}-\widetilde{v}_f^{n+1}}{\Delta t} + \phi\,\nabla p^{n+1} = 0, \\ \mathrm{div}\left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right) = 0, \\ \left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right)\cdot n|_{\Gamma_D} = 0. \end{cases}$$

The drawback of this scheme is that it requires to couple the solid and fluid degrees of freedom to solve the prediction step. Its main advantage is to totally get rid of the time step restriction (4.25). As a matter of fact, assuming that $f = 0$ and reproducing the computations made for the stability analysis of Scheme 1, the discrete energy balance associated with Scheme 3 reads:

$$\begin{aligned} \left(\mathcal{E}^{n+1}-\mathcal{E}^n\right) + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1}-v_f^n\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|v_s^{n+1}-\widetilde{v}_s^{n+1}\right|^2\mathrm{d}x \\ + \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^{n+1}-\widetilde{v}_f^{n+1}\right|^2\mathrm{d}x + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x \\ + \Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp}\right)^2\mathrm{d}x = 0, \end{aligned} \quad (4.42)$$

which directly proves the unconditional stability of the scheme.

Of course, a similar variant can be designed for the incremental version of the scheme. This variant is presented in Scheme 4 below.

Scheme 4 Implicit treatment of permeability (incremental version)

Step 1: `(prediction step)`
Find u_s^{n+1}, $\widetilde{v}_s^{n+1}$ and $\widetilde{v}_f^{n+1}$ such that $u_s^{n+1}|_{\Gamma_D} = \widetilde{v}_s^{n+1}|_{\Gamma_D} = \widetilde{v}_f^{n+1}|_{\Gamma_D} = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{\widetilde{v}_s^{n+1}-v_s^n}{\Delta t} - \mathrm{div}\left(\sigma_s(u_s^{n+\frac{1}{2}})\right) \\ \qquad\qquad - \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+1}\right) + (1-\phi)\nabla p^n = \rho_s(1-\phi)f^{n+\frac{1}{2}}, \\ \rho_f\phi\dfrac{\widetilde{v}_f^{n+1}-v_f^n}{\Delta t} - \mathrm{div}\left(\phi\,\sigma_f(\widetilde{v}_f^{n+1})\right) \\ \qquad\qquad + \phi^2 k_f^{-1}\left(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+1}\right) + \phi\,\nabla p^n = \rho_f\phi\, f^{n+\frac{1}{2}}, \\ \dfrac{u_s^{n+1}-u_s^n}{\Delta t} = \widetilde{v}_s^{n+1}. \end{cases}$$

Step 2: `(correction step)`
Find v_s^{n+1}, v_f^{n+1} and p^{n+1} such that $\int_\Omega p^{n+1}\,\mathrm{d}x = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{v_s^{n+1}-\widetilde{v}_s^{n+1}}{\Delta t} + (1-\phi)\,\nabla(p^{n+1}-p^n) = 0, \\ \rho_f\phi\dfrac{v_f^{n+1}-\widetilde{v}_f^{n+1}}{\Delta t} + \phi\,\nabla(p^{n+1}-p^n) = 0, \\ \operatorname{div}\big((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\big) = 0, \\ \big((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\big)\cdot n|_{\Gamma_D} = 0. \end{cases}$$

Here again, the stability of the scheme is ensured irrespectively of the time step in virtue of the energy balance

$$\begin{aligned} \big(\mathcal{E}^{n+1}-\mathcal{E}^n\big) + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1}-v_s^n\right|^2\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1}-v_f^n\right|^2\mathrm{d}x \\ + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\big(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+1}\big)^2\mathrm{d}x \\ + \frac{\Delta t^2}{2}\left(\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla p^{n+1}\right|^2\mathrm{d}x - \int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla p^n\right|^2\mathrm{d}x\right) = 0, \end{aligned} \tag{4.43}$$

which is directy adapted from (4.41).

Finally, if one is attached to decouple the solid and fluid degrees of freedom during the prediction step, another option is to use a fixed-point strategy, which consists in iterating between the solid and fluid prediction sub-steps until a convergence criterion is reached, for a given tolerance parameter $\varepsilon_{\mathrm{tol}}$. Such an algorithm is summarized in Scheme 5.

Scheme 5 Fixed-point procedure

Initialization: $k=0$, $u_s^{n+1,0}=u_s^n$, $\widetilde{v}_s^{n+1,0}=v_s^n$, $\widetilde{v}_f^{n+1,0}=v_f^n$ and $e=+\infty$.

Step 1: `(prediction step)` While $e>\varepsilon_{\mathrm{tol}}$, go into the following sub-steps:

– **Step 1.1**: `(structure prediction sub-step)`
Find $u_s^{n+1,k+1}$ and $\widetilde{v}_s^{n+1,k+1}$ such that $u_s^{n+1,k+1}|_{\Gamma_D}=0$, $\widetilde{v}_s^{n+1,k+1}|_{\Gamma_D}=0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{\widetilde{v}_s^{n+1,k+1}-v_s^n}{\Delta t} - \operatorname{div}\left(\sigma_s\Big(\dfrac{u_s^{n+1,k+1}+u_s^n}{2}\Big)\right) \\ \qquad\qquad - \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1,k} - \dfrac{\widetilde{v}_s^{n+1,k+1}+v_s^n}{2}\Big) = \rho_s(1-\phi)f^{n+\frac{1}{2}}, \\ \dfrac{u_s^{n+1,k+1}-u_s^n}{\Delta t} = \dfrac{\widetilde{v}_s^{n+1,k+1}+v_s^n}{2}. \end{cases}$$

– **Step 1.2**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1,k+1}$ such that $\widetilde{v}_f^{n+1,k+1}|_{\Gamma_D}=0$ and

$$\begin{cases} \rho_f\phi\dfrac{\widetilde{v}_f^{n+1,k+1}-v_f^n}{\Delta t} - \operatorname{div}\big(\phi\,\sigma_f(\widetilde{v}_f^{n+1,k+1})\big) \\ \qquad\qquad + \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1,k+1} - \dfrac{\widetilde{v}_s^{n+1,k+1}+v_s^n}{2}\Big) = \rho_f\phi\, f^{n+\frac{1}{2}}. \end{cases}$$

– **Step 1.3**: `(fixed-point error update)`
Compute

$$e = \left\|u_s^{n+1,k+1} - u_s^{n+1,k}\right\| + \left\|\widetilde{v}_s^{n+1,k+1} - \widetilde{v}_s^{n+1,k}\right\| + \left\|\widetilde{v}_f^{n+1,k+1} - \widetilde{v}_f^{n+1,k}\right\|,$$

and update $k \leftarrow k+1$.

Prediction step output: set $u_s^{n+1} = u_s^{n+1,k}$, $\widetilde{v}_s^{n+1} = \widetilde{v}_s^{n+1,k}$ and $\widetilde{v}_f^{n+1} = \widetilde{v}_f^{n+1,k}$.

Step 2: `(correction step)`
Find v_s^{n+1}, v_f^{n+1} and p^{n+1} such that $\int_\Omega p^{n+1}\,\mathrm{d}x = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{v_s^{n+1} - \widetilde{v}_s^{n+1}}{\Delta t} + (1-\phi)\,\nabla p^{n+1} = 0, \\ \rho_f \phi \dfrac{v_f^{n+1} - \widetilde{v}_f^{n+1}}{\Delta t} + \phi\,\nabla p^{n+1} = 0, \\ \operatorname{div}\left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right) = 0, \\ \left((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\right)\cdot n|_{\Gamma_D} = 0. \end{cases}$$

Remark 4.10. In order to reduce the number of iterations of the fixed-point procedure, note that Scheme 5 can be solved using a Netwon method as in [Gerbeau and Vidrascu, 2003].

In all this section, we restricted our study to the case of Dirichlet boundary conditions. Let us see what changes for the proposed projection schemes when other types of boundary conditions are considered.

4.3 Neumann and total stress boundary conditions

In many applications, boundary conditions appear on the solid or fluid stresses, either in a distributed way, see (4.6), or with an additional Dirichlet boundary condition between the fluid and solid velocities, see (4.7). The goal of this section is to include such conditions in the projection scheme presented previously. For the sake of conciseness, we will only consider the non-incremental version of the algorithm.

4.3.1 Neumann boundary conditions

We start with the case of Neumann boundary conditions (4.6), namely $\Gamma_D = \Gamma_T = \emptyset$ and

$$\begin{aligned} \sigma_s(u_s)n - (1-\phi)\,pn &= (1-\phi)\,b, \quad && \text{on } \Gamma_N, \\ \phi\,\sigma_f(v_f)n - \phi\,pn &= \phi\,b, && \text{on } \Gamma_N, \end{aligned}$$

with $b \in [\mathrm{L}^2(\Gamma_N)]^d$. Let us denote by $\pi_\tau = \mathbb{I} - n\otimes n$ the tangential plane projection operator. To impose (4.6) in Scheme 1, the boundary traction b applied to the porous medium has to be splitted between the projection and correction steps. To do so, we are going to impose its tangential component in the prediction step, while its normal component will be imposed in the correction step. More precisely, in the solid prediction sub-step (4.8), we impose weakly

$$\sigma_s(u_s^{n+\frac{1}{2}})n = (1-\phi)\,\pi_\tau(b^{n+\frac{1}{2}}), \quad \text{on } \Gamma_N.$$

Similarly, in the fluid prediction sub-step (4.9), we impose weakly

$$\phi\,\sigma_f(\widetilde{v}_f^{n+1})n = \phi\,\pi_\tau(b^{n+\frac{1}{2}}), \quad \text{on } \Gamma_N.$$

Then, the normal component of the external traction is imposed in the correction step (4.10) as a Dirichlet boundary condition on the pressure, namely

$$p^{n+1} = -b^{n+\frac{1}{2}} \cdot n, \quad \text{on } \Gamma_N.$$

The corresponding scheme is weakly written in Scheme 6 below.

Scheme 6 Neumann boundary conditions

Step 1: `(prediction step)`

– **Step 1.1**: `(structure prediction sub-step)`
Find $u_s^{n+1} \in [\mathrm{H}^1(\Omega)]^d$ and $\widetilde{v}_s^{n+1} \in [\mathrm{H}^1(\Omega)]^d$ such that

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{\widetilde{v}_s^{n+1}-v_s^n}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s\Big(\frac{u_s^{n+1}-u_s^n}{\Delta t}\Big) : \varepsilon(d_s)\,\mathrm{d}x \\ - \int_\Omega \sigma_s(\widetilde{v}_s^{n+\frac{1}{2}\sharp}) : \varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s^{n+\frac{1}{2}}) : \varepsilon(w_s)\,\mathrm{d}x - \int_\Omega \phi^2 k_f^{-1}(v_f^n - \widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot w_s\,\mathrm{d}x \\ = \int_\Omega \rho_s(1-\phi) f^{n+\frac{1}{2}}\cdot w_s\,\mathrm{d}x + \int_{\Gamma_N}(1-\phi)\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_s\,\mathrm{d}S, \quad (4.44)\end{aligned}$$

for all $w_s \in [\mathrm{H}^1(\Omega)]^d$ and $d_s \in [\mathrm{H}^1(\Omega)]^d$.

– **Step 1.2**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1} \in [\mathrm{H}^1(\Omega)]^d$ such that

$$\begin{aligned}\int_\Omega \rho_f\phi\frac{\widetilde{v}_f^{n+1}-v_f^n}{\Delta t}\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(w_f)\,\mathrm{d}x \\ + \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot w_f\,\mathrm{d}x \\ = \int_\Omega \rho_f\phi\, f^{n+\frac{1}{2}}\cdot w_f\,\mathrm{d}x + \int_{\Gamma_N}\phi\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_f\,\mathrm{d}S, \quad (4.45)\end{aligned}$$

for all $w_f \in [\mathrm{H}^1(\Omega)]^d$.

Step 2: `(correction step)`
Find $v_s^{n+1} \in [\mathrm{L}^2(\Omega)]^d$, $v_f^{n+1} \in [\mathrm{L}^2(\Omega)]^d$ and $p^{n+1} \in \mathrm{H}^1(\Omega)$ with $p^{n+1}|_{\Gamma_N} = -b^{n+\frac{1}{2}} \cdot n$,

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{v_s^{n+1}-\widetilde{v}_s^{n+1}}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \rho_f\phi\frac{v_f^{n+1}-\widetilde{v}_f^{n+1}}{\Delta t}\cdot w_f\,\mathrm{d}x \\ + \int_\Omega (1-\phi)\nabla p^{n+1}\cdot w_s\,\mathrm{d}x + \int_\Omega \phi\,\nabla p^{n+1}\cdot w_f\,\mathrm{d}x \\ - \int_\Omega (1-\phi)v_s^{n+1}\cdot\nabla q\,\mathrm{d}x - \int_\Omega \phi\,v_f^{n+1}\cdot\nabla q\,\mathrm{d}x = 0, \quad (4.46)\end{aligned}$$

for all $w_s \in [\mathrm{L}^2(\Omega)]^d$, $w_f \in [\mathrm{L}^2(\Omega)]^d$ and $q \in \mathrm{H}^1(\Omega)$.

Let us assume that we use the same test functions in the prediction and correction steps. Summing (4.44), (4.45) and (**??**), the boundary terms appearing in the right-hand side are

$$\int_{\Gamma_N} (1-\phi)\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_s\,\mathrm{d}S + \int_{\Gamma_N} \phi\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_f\,\mathrm{d}S + \int_{\Gamma_N} (b^{n+\frac{1}{2}}\cdot n)\big((1-\phi)\,w_s + \phi w_f\big)\cdot n\,\mathrm{d}S.$$

Using the equality $(b\cdot n)(w\cdot n) = (n\otimes n)b\cdot w$ for all b and w in $\mathbb{R}^d$, these terms can be rewritten as

$$\begin{aligned}
&\int_{\Gamma_N} (1-\phi)\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_s\,\mathrm{d}S + \int_{\Gamma_N} \phi\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_f\,\mathrm{d}S + \int_{\Gamma_N} (b^{n+\frac{1}{2}}\cdot n)\big((1-\phi)\,w_s + \phi w_f\big)\cdot n\,\mathrm{d}S\\
&\quad= \int_{\Gamma_N} (1-\phi)\big(\pi_\tau(b^{n+\frac{1}{2}}) + (n\otimes n)b^{n+\frac{1}{2}}\big)\cdot w_s\,\mathrm{d}S + \int_{\Gamma_N} \phi\big(\pi_\tau(b^{n+\frac{1}{2}}) + (n\otimes n)b^{n+\frac{1}{2}}\big)\cdot w_f\,\mathrm{d}S\\
&\quad= \int_{\Gamma_N} (1-\phi)\,b^{n+\frac{1}{2}}\cdot w_s\,\mathrm{d}S + \int_{\Gamma_N} \phi\,b^{n+\frac{1}{2}}\cdot w_f\,\mathrm{d}S \quad \text{since } \pi_\tau = \mathbb{I} - n\otimes n,
\end{aligned}$$

which shows the consistency of Scheme 6.

In absence of external forces, namely if $f = 0$ and $b = 0$, the energy balance of Scheme 6 is exactly the same as for Scheme 1. Note that if $b \neq 0$, the discrete energy estimate can be extended using a trace inequality. Let us now consider the case of total stress boundary conditions, which requires a specific stability analysis.

4.3.2 Total stress boundary condition

Let us assume that the total stress boundary condition (4.7) holds, namely $\Gamma_D = \Gamma_N = \emptyset$ and

$$\begin{aligned}
v_f &= v_s, \quad &&\text{on } \Gamma_T,\\
\sigma^{\mathrm{tot}} n &= b, \quad &&\text{on } \Gamma_T,
\end{aligned}$$

with

$$\sigma^{\mathrm{tot}} = \sigma_s(u_s) + \phi\,\sigma_f(v_f) - p\mathbb{I}.$$

As mentioned previously, this kind of boundary conditions is very close to the transmission conditions encountered in fluid-structure interaction. Moreover, as in fluid-structure interaction problems, the incompressibility constraint (4.1d) may cause an added-mass effect responsible of numerical unstabilities.

A first option is to treat explicitly the fluid stress part of this condition. This consists in imposing weakly

$$\sigma_s(u_s^{n+\frac{1}{2}})n = \pi_\tau(b^{n+\frac{1}{2}}) - \phi\,\sigma_f(v_f^n)n, \quad \text{on } \Gamma_N$$

in the structure prediction sub-step. Then, knowing $\widetilde{v}_s^{n+1}$, the fluid and solid velocities equality is imposed strongly by setting

$$\widetilde{v}_f^{n+1} = \widetilde{v}_s^{n+1}, \quad \text{on } \Gamma_D$$

in the fluid prediction sub-step. Finally, as for Neumann boundary conditions, the external traction normal component is taken into account by imposing strongly

$$p^{n+1} = -b^{n+\frac{1}{2}}\cdot n, \quad \text{on } \Gamma_N$$

in the correction step. This strategy leads to a consistant formulation. Yet, it introduces in the solid prediction step an additional term of the form $\int_{\Gamma_N} \phi\,\sigma_f(v_f^n)n\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}S$ on the boundary that cannot be controled during the stability analysis and thus may cause numerical unstabilities.

To overcome this difficulty, we are going to use a Robin-Robin coupling approach that was introduced for fluid-structure interaction problems in [Burman et al., 2022b,a], but also analyzed for general parabolic/parabolic and hyperbolic/parabolic problems [Burman et al., 2021]. These studies generalize the Robin method developed in [Burman and Fernández, 2014] and improve the uniformity of the time splitting error with respect to the mesh size h, which is often of order $\mathcal{O}(\Delta t/h)$ in Nitsche's methods [Hansbo et al., 2004; Burman and Fernández, 2009]. Here, we are going to employ the Robin-based algorithm from [Burman et al., 2022b,a] in the prediction step, and show that the correction step can then be handled as in the case of Neumann boundary conditions *without any destabilizing added-mass effect*.

Let us denote by $\alpha > 0$ the Robin coefficient of the method. The proposed Robin-Robin coupling approach hinges on imposing weakly

$$\sigma_s(u_s^{n+\frac{1}{2}})n + \alpha\widetilde{v}_s^{n+\frac{1}{2}\sharp} = (1-\phi)\pi_\tau(b^{n+\frac{1}{2}}) + \alpha\widetilde{v}_f^n - \phi\,\sigma_f(\widetilde{v}_f^n)n, \quad \text{on } \Gamma_T, \tag{4.47}$$

in the solid prediction sub-step (4.8a), and imposing weakly

$$\phi\,\sigma_f(\widetilde{v}_f^{n+1})n + \alpha\widetilde{v}_f^{n+1} = \phi\,\pi_\tau(b^{n+\frac{1}{2}}) + \alpha\widetilde{v}_s^{n+\frac{1}{2}\sharp} + \phi\,\sigma_f(\widetilde{v}_f^n)n \quad \text{on } \Gamma_T, \tag{4.48}$$

in the fluid prediction sub-step. Then, as for Neumann boundary conditions, we impose strongly

$$p^{n+1} = -b^{n+\frac{1}{2}}\cdot n, \quad \text{on } \Gamma_T,$$

in the correction step. Introducing the notation

$$\widetilde{\lambda}^n = \phi\,\sigma_f(\widetilde{v}_f^n)n,$$

for the tentative fluid traction, we respectively get from (4.47) and (4.48) that

$$-\int_{\Gamma_T}\sigma_s(u_s^{n+\frac{1}{2}})n\cdot w_s\,\mathrm{d}S = \alpha\int_{\Gamma_T}(\widetilde{v}_s^{n+\frac{1}{2}\sharp}-\widetilde{v}_f^n)\cdot w_s\,\mathrm{d}S + \int_{\Gamma_T}\widetilde{\lambda}^n\cdot w_s\,\mathrm{d}S - \int_{\Gamma_T}(1-\phi)\pi_\tau(b^{n+\frac{1}{2}})\cdot w_s\,\mathrm{d}S,$$

and

$$-\int_{\Gamma_T}\phi\,\sigma_f(\widetilde{v}_f^{n+1})n\cdot w_f\,\mathrm{d}S = \alpha\int_{\Gamma_T}(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot w_f\,\mathrm{d}S - \int_{\Gamma_T}\widetilde{\lambda}^n\cdot w_f - \int_{\Gamma_T}\phi\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_f\,\mathrm{d}S,$$

which results in the following weak formulation.

Scheme 7 Robin-Robin coupling for total stress boundary conditions

---

Step 1: `(prediction step)`

– **Step 1.1**: `(structure prediction sub-step)`
Find $u_s^{n+1}\in[\mathrm{H}^1(\Omega)]^d$ and $\widetilde{v}_s^{n+1}\in[\mathrm{H}^1(\Omega)]^d$ such that

$$\begin{aligned}
&\int_\Omega \rho_s(1-\phi)\frac{\widetilde{v}_s^{n+1}-v_s^n}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s\Big(\frac{u_s^{n+1}-u_s^n}{\Delta t}\Big):\varepsilon(d_s)\,\mathrm{d}x\\
&-\int_\Omega \sigma_s(\widetilde{v}_s^{n+\frac{1}{2}\sharp}):\varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s^{n+\frac{1}{2}}):\varepsilon(w_s)\,\mathrm{d}x - \int_\Omega \phi^2 k_f^{-1}(v_f^n-\widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot w_s\,\mathrm{d}x\\
&\qquad + \alpha\int_{\Gamma_T}(\widetilde{v}_s^{n+\frac{1}{2}\sharp}-\widetilde{v}_f^n)\cdot w_s\,\mathrm{d}S + \int_{\Gamma_T}\widetilde{\lambda}^n\cdot w_s\,\mathrm{d}S\\
&\qquad = \int_\Omega \rho_s(1-\phi)f^{n+\frac{1}{2}}\cdot w_s\,\mathrm{d}x + \int_{\Gamma_T}(1-\phi)\,\pi_\tau(b^{n+\frac{1}{2}})\cdot w_s\,\mathrm{d}S,
\end{aligned} \tag{4.49}$$

for all $w_s\in[\mathrm{H}^1(\Omega)]^d$ and $d_s\in[\mathrm{H}^1(\Omega)]^d$.

– **Step 1.2**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1} \in [\mathrm{H}^1(\Omega)]^d$ such that

$$\int_\Omega \rho_f \phi \frac{\widetilde{v}_f^{n+1} - v_f^n}{\Delta t} \cdot w_f \,\mathrm{d}x + \int_\Omega \phi\, \sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(w_f)\,\mathrm{d}x + \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}) \cdot w_f \,\mathrm{d}x + \alpha \int_{\Gamma_T} (\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}) \cdot w_f \,\mathrm{d}S - \int_{\Gamma_T} \widetilde{\lambda}^n \cdot w_f = \int_\Omega \rho_f \phi\, f^{n+\frac{1}{2}} \cdot w_f \,\mathrm{d}x + \int_{\Gamma_T} \phi\, \pi_\tau(b^{n+\frac{1}{2}}) \cdot w_f \,\mathrm{d}S, \quad (4.50)$$

for all $w_f \in [\mathrm{H}^1(\Omega)]^d$.

Step 2: `(correction step)`
Find $v_s^{n+1} \in [\mathrm{L}^2(\Omega)]^d$, $v_f^{n+1} \in [\mathrm{L}^2(\Omega)]^d$ and $p^{n+1} \in \mathrm{H}^1(\Omega)$ with $p^{n+1}|_{\Gamma_N} = -b^{n+\frac{1}{2}} \cdot n$,

$$\int_\Omega \rho_s(1-\phi) \frac{v_s^{n+1} - \widetilde{v}_s^{n+1}}{\Delta t} \cdot w_s \,\mathrm{d}x + \int_\Omega \rho_f \phi \frac{v_f^{n+1} - \widetilde{v}_f^{n+1}}{\Delta t} \cdot w_f \,\mathrm{d}x + \int_\Omega (1-\phi) \nabla p^{n+1} \cdot w_s \,\mathrm{d}x + \int_\Omega \phi\, \nabla p^{n+1} \cdot w_f \,\mathrm{d}x - \int_\Omega (1-\phi) v_s^{n+1} \cdot \nabla q \,\mathrm{d}x - \int_\Omega \phi\, v_f^{n+1} \cdot \nabla q \,\mathrm{d}x = 0, \quad (4.51)$$

for all $w_s \in [\mathrm{L}^2(\Omega)]^d$, $w_f \in [\mathrm{L}^2(\Omega)]^d$ and $q \in \mathrm{H}^1(\Omega)$.

Note that from (4.48), we infer

$$\widetilde{\lambda}^{n+1} = \widetilde{\lambda}^n + \alpha\big(\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\big) + \phi\, \pi_\tau(b^{n+\frac{1}{2}}), \quad \text{on } \Gamma_T. \quad (4.52)$$

This relation can be used to update the field $\widetilde{\lambda}^{n+1}$ after the fluid prediction sub-step, so that $\widetilde{\lambda}^n$ does not need to be solved like an additional unknown. Moreover, (4.52) will be crucial for the stability analysis that we are now going to carry out.

Theorem 4.11. *Let u_s^n, v_s^n, v_f^n, p^n, $\widetilde{v}_s^n$ and $\widetilde{v}_f^n$ satisfy Scheme 7 with $f = 0$ and $b = 0$. If the time step verifies the smallness condition (4.25), then for all $0 \le N \le n_T$, it holds*

$$\mathcal{E}^N + \frac{\Delta t}{2\alpha} \int_{\Gamma_T} \left|\widetilde{\lambda}^N\right|^2 \mathrm{d}S + \frac{\alpha \Delta t}{2} \int_{\Gamma_T} \left|\widetilde{v}_f^N\right|^2 \mathrm{d}S + \frac{\alpha \Delta t}{2} \sum_{n=0}^{N-1} \int_{\Gamma_T} \left|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^n\right|^2 \mathrm{d}S + \frac{3}{8} \sum_{n=0}^{N-1} \int_\Omega \rho_s(1-\phi) \left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2 \mathrm{d}x + \frac{1}{2} \sum_{n=0}^{N-1} \int_\Omega \rho_f \phi \left|\widetilde{v}_f^{n+1} - v_f^{n+1}\right|^2 \mathrm{d}x + \Delta t \sum_{n=0}^{N-1} \int_\Omega \phi\, \sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t \sum_{n=0}^{N-1} \int_\Omega \phi^2 k_f^{-1} \left(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\right)^2 \mathrm{d}x \le C_T \left(\mathcal{E}^0 + \frac{\Delta t}{2\alpha} \int_{\Gamma_T} \left|\widetilde{\lambda}^0\right|^2 \mathrm{d}S + \frac{\alpha \Delta t}{2} \int_{\Gamma_T} \left|\widetilde{v}_f^0\right|^2 \mathrm{d}S\right), \quad (4.53)$$

with $C_T > 0$ a constant independent of Δt.

Proof. To obtain the discrete energy balance of Scheme 7, we proceed as for Scheme 1. Testing (4.49) with $w_s = \Delta t\, \widetilde{v}_s^{n+\frac{1}{2}\sharp}$ and $d_s = \Delta t\, u_s^{n+\frac{1}{2}}$, (4.50) with $w_f = \Delta t\, \widetilde{v}_f^{n+1}$ and (**??**) with $w_s = \Delta t\, \widetilde{v}_s^{n+1}$, $w_f = \Delta t\, \widetilde{v}_f^{n+1}$ and $q = \Delta t\, p^{n+1}$, and summing the prediction and correction contributions, we get

$$\begin{aligned}
&\left(\mathcal{E}^{n+1} - \mathcal{E}^n\right) + \frac{1}{2}\int_\Omega \rho_f \phi \left|\widetilde{v}_f^{n+1} - v_f^n\right|^2 \mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s (1-\phi)\left|v_s^{n+1} - \widetilde{v}_s^{n+1}\right|^2 \mathrm{d}x \\
&+ \frac{1}{2}\int_\Omega \rho_f \phi \left|v_f^{n+1} - \widetilde{v}_f^{n+1}\right|^2 \mathrm{d}x + \Delta t \int_\Omega \phi\, \sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t \int_\Omega \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)^2 \mathrm{d}x \\
&+ \alpha\Delta t \int_{\Gamma_T} (\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^n)\cdot \widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}S + \Delta t \int_{\Gamma_T} \widetilde{\lambda}^n \cdot (\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1})\,\mathrm{d}S + \alpha\Delta t \int_{\Gamma_T} (\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot \widetilde{v}_f^{n+1}\,\mathrm{d}S \\
&\qquad = \Delta t \int_\Omega \phi^2 k_f^{-1}\big(v_f^n - \widetilde{v}_f^{n+1}\big)\cdot \widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}x, \quad (4.54)
\end{aligned}$$

The energy balance (4.54) is almost the same than (4.24). The only terms needing a special attention are the ones on the boundary, namely

$$\begin{aligned}
\mathcal{T}_{\Gamma_T} = \alpha\Delta t \int_{\Gamma_T} (\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^n)\cdot \widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}S + \Delta t \int_{\Gamma_T} \widetilde{\lambda}^n \cdot (\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1})\,\mathrm{d}S \\
+ \alpha\Delta t \int_{\Gamma_T} (\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot \widetilde{v}_f^{n+1}\,\mathrm{d}S.
\end{aligned}$$

Splitting $\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^n$ into $(\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}) + (\widetilde{v}_f^{n+1} - \widetilde{v}_f^n)$, $\mathcal{T}_{\Gamma_T}$ can be recast as

$$\mathcal{T}_{\Gamma_T} = \alpha\Delta t \int_{\Gamma_T} \left|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\right|^2 \mathrm{d}S + \alpha\Delta t \int_{\Gamma_T} (\widetilde{v}_f^{n+1} - \widetilde{v}_f^n)\cdot \widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}S + \Delta t \int_{\Gamma_T} \widetilde{\lambda}^n \cdot (\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1})\,\mathrm{d}S$$

Now, since $b = 0$, we know from the key identity (4.52) that $\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1} = \frac{1}{\alpha}(\widetilde{\lambda}^{n+1} - \widetilde{\lambda}^n)$. Thus

$$\mathcal{T}_{\Gamma_T} = \alpha\Delta t \big\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\big\|_{\Gamma_T}^2 + \alpha\Delta t \int_{\Gamma_T} (\widetilde{v}_f^{n+1} - \widetilde{v}_f^n)\cdot \widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}S + \frac{\Delta t}{\alpha}\int_{\Gamma_T} \widetilde{\lambda}^n \cdot (\widetilde{\lambda}^{n+1} - \widetilde{\lambda}^n)\,\mathrm{d}S, \quad (4.55)$$

where we used the notation $\big\|v\big\|_{\Gamma_T}^2 = \int_{\Gamma_T} |v|^2\,\mathrm{d}S$ for the norm on the boundary. Moreover, we have

$$\frac{\Delta t}{\alpha}\int_{\Gamma_T} \widetilde{\lambda}^n \cdot (\widetilde{\lambda}^{n+1} - \widetilde{\lambda}^n)\,\mathrm{d}S = \frac{\Delta t}{2\alpha}\Big(\big\|\widetilde{\lambda}^{n+1}\big\|_{\Gamma_T}^2 - \big\|\widetilde{\lambda}^n\big\|_{\Gamma_T}^2 - \big\|\widetilde{\lambda}^{n+1} - \widetilde{\lambda}^n\big\|_{\Gamma_T}^2\Big).$$

Using again (4.52), it follows that

$$\frac{\Delta t}{\alpha}\int_{\Gamma_T} \widetilde{\lambda}^n \cdot (\widetilde{\lambda}^{n+1} - \widetilde{\lambda}^n)\,\mathrm{d}S = \frac{\Delta t}{2\alpha}\Big(\big\|\widetilde{\lambda}^{n+1}\big\|_{\Gamma_T}^2 - \big\|\widetilde{\lambda}^n\big\|_{\Gamma_T}^2\Big) - \frac{\alpha\Delta t}{2}\big\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\big\|_{\Gamma_T}^2. \quad (4.56)$$

Following [Burman et al., 2022a], to estimate the second term of (4.55) we use the algebraic identity

$$(a-b)\cdot c = \frac{1}{2}\Big(|a|^2 - |b|^2 - |c-a|^2 + |c-b|^2\Big),$$

which can be seen as a generalization of (4.21). This identity implies that

$$\alpha\Delta t \int_{\Gamma_T} (\widetilde{v}_f^{n+1} - \widetilde{v}_f^n)\cdot \widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}S = \frac{\alpha\Delta t}{2}\Big(\big\|\widetilde{v}_f^{n+1}\big\|_{\Gamma_T}^2 - \big\|\widetilde{v}_f^n\big\|_{\Gamma_T}^2 - \big\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\big\|_{\Gamma_T}^2 + \big\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^n\big\|_{\Gamma_T}^2\Big) \quad (4.57)$$

Collecting (4.55), (4.56) and (4.57), we deduce

$$\begin{aligned}\mathcal{T}_{\Gamma_T} &= \alpha\Delta t\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\|_{\Gamma_T}^2 + \frac{\Delta t}{2\alpha}\big(\|\widetilde{\lambda}^{n+1}\|_{\Gamma_T}^2 - \|\widetilde{\lambda}^{n}\|_{\Gamma_T}^2\big) - \frac{\alpha\Delta t}{2}\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\|_{\Gamma_T}^2 \\ &\quad + \frac{\alpha\Delta t}{2}\big(\|\widetilde{v}_f^{n+1}\|_{\Gamma_T}^2 - \|\widetilde{v}_f^{n}\|_{\Gamma_T}^2\big) - \frac{\alpha\Delta t}{2}\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n+1}\|_{\Gamma_T}^2 + \frac{\alpha\Delta t}{2}\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n}\|_{\Gamma_T}^2 \\ &= \frac{\Delta t}{2\alpha}\big(\|\widetilde{\lambda}^{n+1}\|_{\Gamma_T}^2 - \|\widetilde{\lambda}^{n}\|_{\Gamma_T}^2\big) + \frac{\alpha\Delta t}{2}\big(\|\widetilde{v}_f^{n+1}\|_{\Gamma_T}^2 - \|\widetilde{v}_f^{n}\|_{\Gamma_T}^2\big) + \frac{\alpha\Delta t}{2}\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n}\|_{\Gamma_T}^2.\end{aligned}$$

Coming back to (4.54), we have shown that

$$\begin{aligned}&\big(\mathcal{E}^{n+1} - \mathcal{E}^{n}\big) + \big(\mathcal{E}_{\Gamma_T}^{n+1} - \mathcal{E}_{\Gamma_T}^{n}\big) + \frac{\alpha\Delta t}{2}\|\widetilde{v}_s^{n+\frac{1}{2}\sharp} - \widetilde{v}_f^{n}\|_{\Gamma_T}^2 \\ &+ \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{v}_f^{n+1} - v_f^{n}\right|^2 \mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{v}_s^{n+1} - v_s^{n+1}\right|^2 \mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|v_f^{n+1} - \widetilde{v}_f^{n+1}\right|^2 \mathrm{d}x \\ &\qquad + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(\widetilde{v}_f^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\Big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+\frac{1}{2}\sharp}\Big)^2\,\mathrm{d}x \\ &\qquad\qquad = \Delta t\int_\Omega \phi^2 k_f^{-1}\big(v_f^{n} - \widetilde{v}_f^{n+1}\big)\cdot\widetilde{v}_s^{n+\frac{1}{2}\sharp}\,\mathrm{d}x, \quad (4.58)\end{aligned}$$

with

$$\mathcal{E}_{\Gamma_T}^{n} = \frac{\Delta t}{2\alpha}\|\widetilde{\lambda}^{n}\|_{\Gamma_T}^2 + \frac{\alpha\Delta t}{2}\|\widetilde{v}_f^{n}\|_{\Gamma_T}^2. \quad (4.59)$$

The right-hand side of (4.58) can be bounded as in the proof of Theorem 4.5, and the final estimate (4.53) follows from discrete Grönwall Lemma after summing (4.58) between 0 and $N-1$. □

Therefore, the Robin-Robin coupling designed in Scheme 7 furnishes a stable time discretization of the total stress boundary conditions (4.7). The drawback of the method is that it induces an artificial energy term on the boundary, see (4.59). To ensure that this additional energy does not pollute the scheme convergence, the Robin coefficient α must be chosen to be large enough, but not too large in view of the term $\frac{\alpha\Delta t}{2}\|\widetilde{v}_f^{N}\|_{\Gamma_T}^2$ appearing in the left-hand side of (4.58). This issue will be explored numerically in Section 4.5.

Remark 4.12. In [Burtschell et al., 2017], the authors follow the Robin coupling derived from Nitsche's method in [Astorino et al., 2010]. In our case, such an approach leads to the following scheme, in which h denotes the mesh size and $\gamma > 0$ the Nitsche's penalty coefficient.

Scheme 8 Nitsche's method for total stress boundary conditions

Step 1: `(prediction step)`

– **Step 1.1**: `(fluid prediction sub-step)`
Find $\widetilde{v}_f^{n+1} \in [\mathrm{H}^1(\Omega)]^d$ such that

$$\begin{aligned}&\int_\Omega \rho_f\phi\frac{\widetilde{v}_f^{n+1} - v_f^{n}}{\Delta t}\cdot w_f\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(w_f)\,\mathrm{d}x \\ &\quad + \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1} - v_s^{n})\cdot w_f\,\mathrm{d}x + \frac{\gamma\mu_f}{h}\int_{\Gamma_N}\phi\left(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n-\frac{1}{2}\sharp}\right)\cdot w_f \\ &\quad - \int_{\Gamma_N}\phi\,\sigma_f(\widetilde{v}_f^{n+1})n\cdot w_f = \int_\Omega \rho_f\phi\, f^{n+\frac{1}{2}}\cdot w_f,\end{aligned}$$

for all $w_f \in [\mathrm{H}^1(\Omega)]^d$.

– **Step 1.2**: `(structure prediction sub-step)`
Find $u_s^{n+1} \in [\mathrm{H}^1(\Omega)]^d$ and $\widetilde{v}_s^{n+1} \in [\mathrm{H}^1(\Omega)]^d$ such that

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{\widetilde{v}_s^{n+1}-v_s^n}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \sigma_s\Big(\frac{u_s^{n+1}-u_s^n}{\Delta t}\Big) : \varepsilon(d_s)\,\mathrm{d}x \\ - \int_\Omega \sigma_s(\widetilde{v}_s^{n+\frac{1}{2}\sharp}) : \varepsilon(d_s)\,\mathrm{d}x + \int_\Omega \sigma_s(u_s^{n+\frac{1}{2}}) : \varepsilon(w_s)\,\mathrm{d}x \\ - \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n+\frac{1}{2}\sharp})\cdot w_s\,\mathrm{d}x + \frac{\gamma\mu_f}{h}\int_{\Gamma_N}\phi\,(\widetilde{v}_s^{n+\frac{1}{2}\sharp}-\widetilde{v}_s^{n-\frac{1}{2}\sharp})\cdot w_s \\ = \int_\Omega \rho_s(1-\phi)f^{n+\frac{1}{2}}\cdot w_s\,\mathrm{d}x + \int_{\Gamma_N}\pi_\tau(b^{n+\frac{1}{2}})\cdot w_s - \mathcal{R}_f(w_s),\end{aligned}$$

for all $w_s \in [\mathrm{H}^1(\Omega)]^d$ and $d_s \in [\mathrm{H}^1(\Omega)]^d$, where $\mathcal{R}_f(w_s)$ denotes the residual coming from the fluid prediction sub-step, namely

$$\begin{aligned}\mathcal{R}_f(w_s) = \int_\Omega \rho_f\phi\frac{\widetilde{v}_f^{n+1}-v_f^n}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(\widetilde{v}_f^{n+1}) : \varepsilon(w_s)\,\mathrm{d}x \\ + \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_f^{n+1}-\widetilde{v}_s^{n-\frac{1}{2}\sharp})\cdot w_f\,\mathrm{d}x - \int_\Omega \rho_f\phi\, f^{n+\frac{1}{2}}\cdot w_s.\end{aligned}$$

Step 2: `(correction step)`
Find $v_s^{n+1} \in [\mathrm{L}^2(\Omega)]^d$, $v_f^{n+1} \in [\mathrm{L}^2(\Omega)]^d$ and $p^{n+1} \in \mathrm{H}^1(\Omega)$ with $p^{n+1}|_{\Gamma_N} = -b^{n+\frac{1}{2}}\cdot n$,

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{v_s^{n+1}-\widetilde{v}_s^{n+1}}{\Delta t}\cdot w_s\,\mathrm{d}x + \int_\Omega \rho_f\phi\frac{v_f^{n+1}-\widetilde{v}_f^{n+1}}{\Delta t}\cdot w_f\,\mathrm{d}x \\ + \int_\Omega (1-\phi)\nabla p^{n+1}\cdot w_s\,\mathrm{d}x + \int_\Omega \phi\,\nabla p^{n+1}\cdot w_f\,\mathrm{d}x \\ - \int_\Omega (1-\phi)v_s^{n+1}\cdot\nabla q\,\mathrm{d}x - \int_\Omega \phi\,v_f^{n+1}\cdot\nabla q\,\mathrm{d}x = 0, \quad (4.60)\end{aligned}$$

for all $w_s \in [\mathrm{L}^2(\Omega)]^d$, $w_f \in [\mathrm{L}^2(\Omega)]^d$ and $q \in \mathrm{H}^1(\Omega)$.

Note that in Scheme 8, the fluid is advanced before the solid, contrary to Scheme 7. One can show that Scheme 8 is stable provided that the time step restriction (4.25) is fulfilled, that the penalty coefficient γ is large enough – more precisely, $\gamma > 4C_{\mathrm{ie}}$ with C_{ie} a trace-inverse inequality constant – and that $\gamma\mu_f\Delta t = \mathcal{O}(h)$. An advantage of the Robin-Robin coupling presented before compared to Scheme 8 is that it gets rid of the latter condition.

4.4 Convergence analysis

In this section, we provide a complete error analysis for the projection scheme proposed previously. To simplify, we will restrict ourselves to the case of homogeneous Dirichlet conditions. Moreover, to avoid the time step condition (4.25) that happens to be very restrictive for the targeted biomedical applications, we will assume that the permeability is treated implicitly in the prediction step as in Schemes 3 and 4. In order to reuse some notation and computations from Chapter 3, the solid will be discretized using a backward Euler scheme. The scheme for which we are going to perform the

convergence analysis is summarized in Scheme 9. Note that Scheme 9 includes both non-incremental and incremental versions thanks to a parameter $i \in \{0, 1\}$, with the convention

$$i = \begin{cases} 0 & \text{for the non-incremental version of the scheme,} \\ 1 & \text{for the incremental version of the scheme.} \end{cases}$$

Scheme 9 is almost similar to Schemes 3 and 4, the only difference being the discretization of the solid part that induces numerical dissipation.

Scheme 9 Implicit treatment of permeability (dissipative version)

Step 1: `(prediction step)`
Find u_s^{n+1}, $\widetilde{v}_s^{n+1}$ and $\widetilde{v}_f^{n+1}$ such that $u_s^{n+1}|_{\Gamma_D} = \widetilde{v}_s^{n+1}|_{\Gamma_D} = \widetilde{v}_f^{n+1}|_{\Gamma_D} = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{\widetilde{v}_s^{n+1} - v_s^n}{\Delta t} - \operatorname{div}\big(\sigma_s(u_s^{n+1})\big) \\ \qquad\qquad - \phi^2 k_f^{-1}\big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+1}\big) + i(1-\phi)\nabla p^n = \rho_s(1-\phi) f^{n+1}, \\ \rho_f \phi \dfrac{\widetilde{v}_f^{n+1} - v_f^n}{\Delta t} - \operatorname{div}\big(\phi\, \sigma_f(\widetilde{v}_f^{n+1})\big) \\ \qquad\qquad + \phi^2 k_f^{-1}\big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+1}\big) + i\phi\, \nabla p^n = \rho_f \phi\, f^{n+1}, \\ \dfrac{u_s^{n+1} - u_s^n}{\Delta t} = \widetilde{v}_s^{n+1}. \end{cases}$$

Step 2: `(correction step)`
Find v_s^{n+1}, v_f^{n+1} and p^{n+1} such that $\int_\Omega p^{n+1}\,\mathrm{d}x = 0$ and

$$\begin{cases} \rho_s(1-\phi)\dfrac{v_s^{n+1} - \widetilde{v}_s^{n+1}}{\Delta t} + (1-\phi)\,\nabla(p^{n+1} - ip^n) = 0, \\ \rho_f \phi \dfrac{v_f^{n+1} - \widetilde{v}_f^{n+1}}{\Delta t} + \phi\,\nabla(p^{n+1} - ip^n) = 0, \\ \operatorname{div}\big((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\big) = 0, \\ \big((1-\phi)v_s^{n+1} + \phi\, v_f^{n+1}\big)\cdot n|_{\Gamma_D} = 0. \end{cases}$$

Remark 4.13. The correction step at time t^n implies that

$$\frac{\rho_s(1-\phi)}{\Delta t} v_s^n = \frac{\rho_s(1-\phi)}{\Delta t}\widetilde{v}_s^n - (1-\phi)\,\nabla(p^n - ip^{n-1}) \quad \text{and} \quad \frac{\rho_f \phi}{\Delta t} v_f^n = \frac{\rho_f \phi}{\Delta t}\widetilde{v}_f^n - \phi\,\nabla(p^n - ip^{n-1}).$$

This relation can be used to totally eliminate the end-of-step velocities v_s^n and v_f^n from the prediction step, leading to

$$\begin{cases} \rho_s(1-\phi)\dfrac{\widetilde{v}_s^{n+1} - \widetilde{v}_s^n}{\Delta t} - \operatorname{div}\big(\sigma_s(u_s^{n+1})\big) \\ \qquad\qquad - \phi^2 k_f^{-1}\big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+1}\big) + (1-\phi)\nabla\big(p^n + i(p^n - p^{n-1})\big) = \rho_s(1-\phi) f^{n+1}, \\ \rho_f \phi \dfrac{\widetilde{v}_f^{n+1} - \widetilde{v}_f^n}{\Delta t} - \operatorname{div}\big(\phi\, \sigma_f(\widetilde{v}_f^{n+1})\big) \\ \qquad\qquad + \phi^2 k_f^{-1}\big(\widetilde{v}_f^{n+1} - \widetilde{v}_s^{n+1}\big) + \phi\,\nabla\big(p^n + i(p^n - p^{n-1})\big) = \rho_f \phi\, f^{n+1}, \\ \dfrac{u_s^{n+1} - u_s^n}{\Delta t} = \widetilde{v}_s^{n+1}. \end{cases} \tag{4.61}$$

Furthermore, the mixture divergence constraint yields

$$\operatorname{div}\left((1-\phi)\widetilde{v}_s^{n+1}+\phi\,\widetilde{v}_f^{n+1}\right)-\Delta t\operatorname{div}\left(\rho_{\text{eff}}^{-1}\nabla(p^{n+1}-ip^n)\right)=0. \tag{4.62}$$

From (4.61) and (4.62), we see that Scheme 9 can be interpreted as a penalized version of the monolithic scheme, in which the pressure gradient is treated explicitly. At the continuous level, this corresponds to penalizing the incompressibility constraint of Problem (4.1) as follows:

$$\operatorname{div}\left((1-\phi)v_s+\phi\,v_f\right)-\varepsilon\operatorname{div}\left(\rho_{\text{eff}}^{-1}\nabla p\right)=0 \quad \text{if } i=0, \text{ with } \varepsilon=\Delta t,$$

and

$$\operatorname{div}\left((1-\phi)v_s+\phi\,v_f\right)-\varepsilon\operatorname{div}\left(\rho_{\text{eff}}^{-1}\nabla\partial_t p\right)=0 \quad \text{if } i=1, \text{ with } \varepsilon=\Delta t^2.$$

The value of the penalty parameter ε gives us a first intuition of the time convergence order of the scheme in non-incremental and incremental versions, which will be justified theoretically in what follows.

The section is organized as follows. First, we will give the discrete setting associated with the spatial discretization of Scheme 9. Then, we derive the error equations between the continuous solution and the fully discrete solution of the scheme. The error system is established simultaneously for the non-incremental and incremental versions. Lastly, the final error analysis will be presented separately for the case $i=0$ and $i=1$.

4.4.1 Total discretization

As mentioned in (4.13), the projection scheme is consistent with the monolithic scheme studied in the previous chapter. Consequently, we globally use the same spatial discretization than the one proposed in Chapter 3 and we adopt the same notation as in this chapter: we consider conforming approximations X_h and Q_h of the spaces $[\mathrm{H}_0^1(\Omega)]^d$ and $\mathrm{L}_0^2(\Omega)$. Moreover, we assume that the discrete inf-sup condition

$$\exists\beta>0, \forall p_h\in Q_h, \quad \sup_{v_h\in X_h}\frac{\displaystyle\int_\Omega \operatorname{div} v_h\, p_h\,\mathrm{d}x}{\|v_h\|_{[\mathrm{H}_0^1(\Omega)]^d}}\geq\beta\|p_h\|. \tag{4.63}$$

is satisfied. Note that one might be tempted to use standard finite elements that do not satisfy the inf-sup condition since Scheme 9 does not require to solve any saddle-point problem, neither in the prediction step nor in the correction step that can be formulated as a Poisson problem, see (4.11). However, we will numerically illustrate in Section 4.5 that the inf-sup condition (4.63) is fundamental to retrieve a correct approximation of the pressure field, as it has been shown in the case of projection schemes for incompressible fluids [Guermond and Quartapelle, 1998]. This condition can be dropped only if the time step is large enough or by using stabilization techniques. The stabilization approach is the one followed in [Markert et al., 2009], in which the time step is in addition restricted by a CFL condition coming from the explicit treatment of the structural displacement.

In order to solve the correction step as a Poisson problem for the pressure, we assume in addition that the discrete pressure space Q_h is a conforming approximation of the space $\mathrm{H}^1(\Omega)$. This is the case for many of the finite elements satisfying (4.63), such as the Taylor-Hood or MINI elements. Supposing to simplify that $f_h^{n+1}=f(t^{n+1})$, the fully discrete formulation of Scheme 9 then reads:

Step 1: `(prediction step)`
Find $u_{s,h}^{n+1} \in X_h$, $\widetilde{v}_{s,h}^{n+1} \in X_h$ and $\widetilde{v}_{f,h}^{n+1} \in X_h$ such that

$$\begin{aligned}
&\int_\Omega \sigma_s\Big(\frac{u_{s,h}^{n+1}-u_{s,h}^{n}}{\Delta t}\Big) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)\frac{\widetilde{v}_{s,h}^{n+1}-v_{s,h}^{n}}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x \\
&+ \int_\Omega \rho_f\phi\frac{\widetilde{v}_{f,h}^{n+1}-v_{f,h}^{n}}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x - \int_\Omega \sigma_s(\widetilde{v}_{s,h}^{n+1}) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \sigma_s(u_{s,h}^{n+1}) : \varepsilon(w_{s,h})\,\mathrm{d}x \\
&+ \int_\Omega \phi\,\sigma_f(\widetilde{v}_{f,h}^{n+1}) : \varepsilon(w_{f,h})\,\mathrm{d}x + \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_{f,h}^{n+1}-\widetilde{v}_{s,h}^{n+1})\cdot(w_{f,h}-w_{s,h})\,\mathrm{d}x \\
&+ i\int_\Omega (1-\phi)\nabla p_h^n\cdot w_{s,h}\,\mathrm{d}x + i\int_\Omega \phi\,\nabla p_h^n\cdot w_{f,h}\,\mathrm{d}x \\
&= \int_\Omega \rho_s(1-\phi)\,f(t^{n+1})\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,f(t^{n+1})\cdot w_{f,h}\,\mathrm{d}x, \quad (4.64)
\end{aligned}$$

for all $w_{s,h} \in X_h$, $d_{s,h} \in X_h$ and $w_{f,h} \in X_h$.

Step 2: `(correction step)`
Find $v_{s,h}^{n+1} \in Y_{s,h}$, $v_{f,h}^{n+1} \in Y_{f,h}$ and $p_h^{n+1} \in Q_h$ such that

$$\begin{aligned}
&\int_\Omega \rho_s(1-\phi)\frac{v_{s,h}^{n+1}-\widetilde{v}_{s,h}^{n+1}}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\frac{v_{f,h}^{n+1}-\widetilde{v}_{f,h}^{n+1}}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x \\
&+ \int_\Omega (1-\phi)\nabla(p_h^{n+1}-ip_h^n)\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \phi\,\nabla(p_h^{n+1}-ip_h^n)\cdot w_{f,h}\,\mathrm{d}x \\
&- \int_\Omega (1-\phi)v_{s,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x - \int_\Omega \phi\,v_{f,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x = 0, \quad (4.65)
\end{aligned}$$

for all $w_{s,h} \in Y_{s,h}$, $w_{f,h} \in Y_{f,h}$ and $q_h \in Q_h$, where the discrete spaces

$$Y_{s,h} = X_h + (1-\phi)\nabla Q_h \quad \text{and} \quad Y_{f,h} = X_h + \phi\nabla Q_h$$

are conforming approximations of $[\mathrm{L}^2(\Omega)]^d$ since $Q_h \subset \mathrm{H}^1(\Omega)$.

4.4.2 Error system

To derive the equations satisfied by the error between the fully discrete solution of (4.64) – (4.65) and the continuous solution of (4.1), we use the same projector P_h that has been introduced in the previous chapter, see (3.40). The advantage of this projector is that it is adapted to the bilinear form involved in the system, so that it generates few residual terms. Let us denote by $u_{s,h}(t^n)$, $v_{s,h}(t^n)$, $v_{f,h}(t^n)$ and $p_h(t^n)$ the projections on the discrete spaces of the continuous solution at time t^n using this projector, namely

$$\big(u_{s,h}(t^n), v_{s,h}(t^n), v_{f,h}(t^n), p_h(t^n)\big) = P_h\big(u_s(t^n), v_s(t^n), v_f(t^n), p(t^n)\big).$$

Provided that the continuous solution is regular enough, we know from (3.43) and (3.44) that

$$\Big\|\big(u_s(t^n), v_s(t^n), v_f(t^n), p(t^n)\big) - \big(u_{s,h}(t^n), v_{s,h}(t^n), v_{f,h}(t^n), p_h(t^n)\big)\Big\|_X \le C(h^\ell + h^r), \quad (4.66)$$

where the constant $C > 0$ depends on the continuous solution, the convergence orders ℓ and $r \le \ell$ depend on the choice of the discrete spaces, and $X = [\mathrm{H}_0^1(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times [\mathrm{L}^2(\Omega)]^d \times \mathrm{L}^2(\Omega)$.

In particular, (4.66) implies that it is sufficient to study the error between the projection of the continuous solution and the fully discrete solution since

$$\begin{aligned}\Big\|\big(u_s(t^n),v_s(t^n),v_f(t^n),p(t^n)\big)-\big(u_{s,h}^n,v_{s,h}^n,v_{f,h}^n,p_h^n\big)\Big\|_X \\ \leq \Big\|\big(u_s(t^n),v_s(t^n),v_f(t^n),p(t^n)\big)-\big(u_{s,h}(t^n),v_{s,h}(t^n),v_{f,h}(t^n),p_h(t^n)\big)\Big\|_X \\ + \Big\|\big(u_{s,h}(t^n),v_{s,h}(t^n),v_{f,h}(t^n),p_h(t^n)\big)-\big(u_{s,h}^n,v_{s,h}^n,v_{f,h}^n,p_h^n\big)\Big\|_X.\end{aligned}$$

To that end, we introduce the discrete errors

$$\begin{aligned} e_{u,h}^n &= u_{s,h}(t^n)-u_{s,h}^n, & \delta_h^n &= p_h(t^n)-p_h^n,\\ \widetilde{e}_{s,h}^n &= v_{s,h}(t^n)-\widetilde{v}_{s,h}^n, & e_{s,h}^n &= v_{s,h}(t^n)-v_{s,h}^n,\\ \widetilde{e}_{f,h}^n &= v_{f,h}(t^n)-\widetilde{v}_{f,h}^n, & e_{s,h}^n &= v_{f,h}(t^n)-v_{f,h}^n. \end{aligned} \tag{4.67}$$

As in Chapter 3 – see (3.51) – we get that the P_h-projection of the solution satisfies

$$\begin{aligned}&\int_\Omega \sigma_s\Big(\frac{u_{s,h}(t^{n+1})-u_{s,h}(t^n)}{\Delta t}\Big):\varepsilon(d_{s,h})\,\mathrm{d}x\\ &\quad+\int_\Omega \rho_s(1-\phi)\frac{v_{s,h}(t^{n+1})-v_{s,h}(t^n)}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x+\int_\Omega \rho_f\phi\frac{v_{f,h}(t^{n+1})-v_{f,h}(t^n)}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x\\ &\qquad-\int_\Omega \sigma_s\big(v_{s,h}(t^{n+1})\big):\varepsilon(d_{s,h})\,\mathrm{d}x+\int_\Omega \sigma_s\big(u_{s,h}(t^{n+1})\big):\varepsilon(w_{s,h})\,\mathrm{d}x\\ &\quad+\int_\Omega \phi\,\sigma_f\big(v_{f,h}(t^{n+1})\big):\varepsilon(w_{f,h})\,\mathrm{d}x+\int_\Omega \phi^2k_f^{-1}\big(v_{f,h}(t^{n+1})-v_{s,h}(t^{n+1})\big)\cdot(w_{f,h}-w_{s,h})\,\mathrm{d}x\\ &\qquad+\int_\Omega(1-\phi)\nabla p_h(t^{n+1})\cdot w_{s,h}\,\mathrm{d}x+\int_\Omega\phi\,\nabla p_h(t^{n+1})\cdot w_{f,h}\,\mathrm{d}x\\ &\qquad-\int_\Omega(1-\phi)v_{s,h}(t^{n+1})\cdot\nabla q_h\,\mathrm{d}x-\int_\Omega\phi\,v_{f,h}(t^{n+1})\cdot\nabla q_h\,\mathrm{d}x\\ &=\int_\Omega\sigma_s\big(R_{u,h}^{n+1}\big):\varepsilon(d_{s,h})\,\mathrm{d}x+\int_\Omega\rho_s(1-\phi)R_{s,h}^{n+1}\cdot w_{s,h}\,\mathrm{d}x+\int_\Omega\rho_f\phi\,R_{f,h}^{n+1}\cdot w_{f,h}\,\mathrm{d}x\\ &\qquad+\int_\Omega\rho_s(1-\phi)\,f(t^{n+1})\cdot w_{s,h}\,\mathrm{d}x+\int_\Omega\rho_f\phi\,f(t^{n+1})\cdot w_{f,h}\,\mathrm{d}x,\end{aligned} \tag{4.68}$$

for all $w_{s,h}\in X_h$, $d_{s,h}\in X_h$, $w_{f,h}\in X_h$ and $q_h\in Q_h$, with the residual term

$$\big(R_{u,h}^{n+1},R_{s,h}^{n+1},R_{f,h}^{n+1}\big)=\mathcal{R}^{n+1}+\lambda_0\mathcal{S}_h^{n+1}+\mathcal{T}_h^{n+1}$$

that is computed from the space and time consistency terms $\mathcal{R}^{n+1}$, $\mathcal{S}_h^{n+1}$ and $\mathcal{T}_h^{n+1}$ defined in Chapter 3. Note that this residual term depends on a parameter λ_0. In Chapter 3, this parameter needs to satisfy the condition $\lambda_0>(\rho_f\phi_{\min})^{-1}\|\theta\|_{L^\infty(\Omega)}$. Here, since $\theta=0$, we can choose any positive parameter for λ_0. In the sequel, we will therefore assume for instance that $\lambda_0=1$. Moreover, in view of the consistency terms estimates (3.55), (3.56) and (3.57), we have

$$\int_\Omega\sigma_s\big(R_{u,h}^{n+1}\big):\varepsilon\big(R_{u,h}^{n+1}\big)\,\mathrm{d}x+\int_\Omega\rho_s(1-\phi)\left|R_{s,h}^{n+1}\right|^2\mathrm{d}x+\int_\Omega\rho_f\phi\left|R_{f,h}^{n+1}\right|^2\mathrm{d}x\leq C(\Delta t+h^\ell+h^r)^2. \tag{4.69}$$

We are now ready to derive the equations satisfied by the errors defined in (4.67). Substracting

the discrete prediction step (4.64) from (4.68) with $q_h = 0$, we obtain

$$\begin{aligned}
&\int_\Omega \sigma_s\Big(\frac{e_{u,h}^{n+1} - e_{u,h}^n}{\Delta t}\Big) : \varepsilon(d_{s,h})\,\mathrm{d}x \\
&\qquad + \int_\Omega \rho_s(1-\phi)\frac{\widetilde{e}_{s,h}^{n+1} - e_{s,h}^n}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\frac{\widetilde{e}_{f,h}^{n+1} - e_{f,h}^n}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad\quad - \int_\Omega \sigma_s\big(\widetilde{e}_{s,h}^{n+1}\big) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \sigma_s\big(e_{u,h}^{n+1}\big) : \varepsilon(w_{s,h})\,\mathrm{d}x \\
&\qquad + \int_\Omega \phi\,\sigma_f\big(\widetilde{e}_{f,h}^{n+1}\big) : \varepsilon(w_{f,h})\,\mathrm{d}x + \int_\Omega \phi^2 k_f^{-1}\big(\widetilde{e}_{f,h}^{n+1} - \widetilde{e}_{s,h}^{n+1}\big)\cdot(w_{f,h} - w_{s,h})\,\mathrm{d}x \\
&\qquad + \int_\Omega (1-\phi)\nabla\big(p_h(t^{n+1}) - ip_h^n\big)\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \phi\,\nabla\big(p_h(t^{n+1}) - ip_h^n\big)\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad = \int_\Omega \sigma_s\big(R_{u,h}^{n+1}\big) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\,R_{f,h}^{n+1}\cdot w_{f,h}\,\mathrm{d}x, \quad (4.70)
\end{aligned}$$

for all $w_{s,h} \in X_h$, $d_{s,h} \in X_h$ and $w_{f,h} \in X_h$. For the correction step, we write

$$v_{s,h}^{n+1} - \widetilde{v}_{s,h}^{n+1} = v_{s,h}^{n+1} - v_{s,h}(t^{n+1}) + v_{s,h}(t^{n+1}) - \widetilde{v}_{s,h}^{n+1} = \widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n+1},$$

and likewise $v_{f,h}^{n+1} - \widetilde{v}_{f,h}^{n+1} = \widetilde{e}_{f,h}^{n+1} - e_{f,h}^{n+1}$. Moreover, since

$$-\int_\Omega (1-\phi)v_{s,h}(t^{n+1})\cdot\nabla q_h\,\mathrm{d}x - \int_\Omega \phi\,v_{f,h}(t^{n+1})\cdot\nabla q_h\,\mathrm{d}x = 0, \quad \forall q_h \in Q_h,$$

and

$$-\int_\Omega (1-\phi)v_{s,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x - \int_\Omega \phi\,v_{f,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x = 0, \quad \forall q_h \in Q_h,$$

we also have

$$-\int_\Omega (1-\phi)e_{s,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x - \int_\Omega \phi\,e_{f,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x = 0, \quad \forall q_h \in Q_h.$$

Incorporating these results into (4.65), we get

$$\begin{aligned}
&\int_\Omega \rho_s(1-\phi)\frac{e_{s,h}^{n+1} - \widetilde{e}_{s,h}^{n+1}}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \rho_f\phi\frac{e_{f,h}^{n+1} - \widetilde{e}_{f,h}^{n+1}}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad + \int_\Omega (1-\phi)\nabla(ip_h^n - p_h^{n+1})\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \phi\,\nabla(ip_h^n - p_h^{n+1})\cdot w_{f,h}\,\mathrm{d}x \\
&\qquad\quad - \int_\Omega (1-\phi)e_{s,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x - \int_\Omega \phi\,e_{f,h}^{n+1}\cdot\nabla q_h\,\mathrm{d}x = 0, \quad (4.71)
\end{aligned}$$

for all $w_{s,h} \in Y_{s,h}$, $w_{f,h} \in Y_{f,h}$ and $q_h \in Q_h$. The pressure residuals multiplying the test functions $(1-\phi)w_{s,h}$ and $\phi\,w_{f,h}$ in (4.70) and (4.71), namely the terms $p_h(t^{n+1}) - ip_h^n$ and $ip_h^n - p_h^{n+1}$, play a key role in the time convergence of the scheme, as we are now going to see it in the error analysis.

4.4.3 Error analysis

The error analysis hinges on the discrete energy balances derived previously. Let us define the energy associated with the errors (4.67) by

$$E_h^n = \frac{1}{2}\int_\Omega \sigma_s(e_{u,h}^n) : \varepsilon(e_{u,h}^n)\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\,\big|e_{s,h}^n\big|^2\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\,\big|e_{f,h}^n\big|^2\,\mathrm{d}x.$$

The convergence of Scheme 9 in the non-incremental case is stated in Theorem 4.14 below, which is the main result of this section.

Theorem 4.14. *Assume that $i = 0$ and that the solution of Problem (4.1) is regular enough, in particular that $p \in \mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$. If $\Delta t < 1$ and if the initialization of Scheme 9 is such that*

$$E_h^0 \leq C(h^\ell + h^r)^2, \tag{4.72}$$

then for all $0 \leq N \leq n_T$, it holds

$$\begin{aligned} E_h^N + \frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \sigma_s(e_{u,h}^{n+1} - e_{u,h}^n) : \varepsilon(e_{u,h}^{n+1} - e_{u,h}^n)\,\mathrm{d}x + \frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1} - e_{s,h}^n\right|^2 \mathrm{d}x \\ + \frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1} - e_{f,h}^n\right|^2 \mathrm{d}x + \frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|e_{s,h}^{n+1} - \widetilde{e}_{s,h}^{n+1}\right|^2 \mathrm{d}x \\ + \frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|e_{f,h}^{n+1} - \widetilde{e}_{f,h}^{n+1}\right|^2 \mathrm{d}x + \frac{\Delta t}{2}\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}) : \varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x \\ + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1} - \widetilde{e}_{s,h}^{n+1}\right)^2 \mathrm{d}x \leq C(h^\ell + h^r)^2 + C\Delta t, \end{aligned} \tag{4.73}$$

with $C > 0$ a constant independent of h and Δt.

Proof. As in the stability analysis – see (4.42) – we consider the test functions $w_{s,h} = \Delta t\,\widetilde{e}_{s,h}^{n+1}$, $d_{s,h} = \Delta t\,e_{u,h}^{n+1}$, $w_{f,h} = \Delta t\,\widetilde{e}_{f,h}^{n+1}$ in the error prediction step (4.70), and $w_{s,h} = \Delta t\,e_{s,h}^{n+1}$, $w_{f,h} = \Delta t\,e_{f,h}^{n+1}$, $q_h = \Delta t\,\delta_h^{n+1}$ in the error correction step (4.71), leading to

$$\begin{aligned} (E_h^{n+1} - E_h^n) + \frac{1}{2}\int_\Omega \sigma_s(e_{u,h}^{n+1} - e_{u,h}^n) : \varepsilon(e_{u,h}^{n+1} - e_{u,h}^n)\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1} - e_{s,h}^n\right|^2 \mathrm{d}x \\ + \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1} - e_{f,h}^n\right|^2 \mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|e_{s,h}^{n+1} - \widetilde{e}_{s,h}^{n+1}\right|^2 \mathrm{d}x + \frac{1}{2}\int_\Omega \rho_f\phi\left|e_{f,h}^{n+1} - \widetilde{e}_{f,h}^{n+1}\right|^2 \mathrm{d}x \\ + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}) : \varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1} - \widetilde{e}_{s,h}^{n+1}\right)^2 \mathrm{d}x \\ = \mathcal{T}_p^{n+1} + \mathcal{T}_u^{n+1} + \mathcal{T}_s^{n+1} + \mathcal{T}_f^{n+1}, \end{aligned} \tag{4.74}$$

with the residual terms

$$\begin{aligned} \mathcal{T}_p^{n+1} = -\Delta t\int_\Omega (1-\phi)\nabla p_h(t^{n+1})\cdot\widetilde{e}_{s,h}^{n+1}\,\mathrm{d}x - \Delta t\int_\Omega \phi\,\nabla p_h(t^{n+1})\cdot\widetilde{e}_{f,h}^{n+1}\,\mathrm{d}x \\ + \Delta t\int_\Omega (1-\phi)\nabla p_h^{n+1}\cdot e_{s,h}^{n+1}\,\mathrm{d}x + \Delta t\int_\Omega \phi\,\nabla p_h^{n+1}\cdot e_{f,h}^{n+1}\,\mathrm{d}x \\ + \Delta t\int_\Omega (1-\phi)\nabla\delta_h^{n+1}\cdot e_{s,h}^{n+1}\,\mathrm{d}x + \Delta t\int_\Omega \phi\,\nabla\delta_h^{n+1}\cdot e_{f,h}^{n+1}\,\mathrm{d}x, \end{aligned}$$

and

$$\begin{aligned} \mathcal{T}_u^{n+1} &= \Delta t\int_\Omega \sigma_s\left(R_{u,h}^{n+1}\right) : \varepsilon(e_{u,h}^{n+1})\,\mathrm{d}x, \\ \mathcal{T}_s^{n+1} &= \Delta t\int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot\widetilde{e}_{s,h}^{n+1}\,\mathrm{d}x, \\ \mathcal{T}_f^{n+1} &= \Delta t\int_\Omega \rho_f\phi\,R_{f,h}^{n+1}\cdot\widetilde{e}_{f,h}^{n+1}\,\mathrm{d}x. \end{aligned}$$

Let us start by estimating the pressure residual term $\mathcal{T}_p^{n+1}$. Since $p_h^{n+1} = p_h(t^{n+1}) - \delta_h^{n+1}$, the terms of the form $(1-\phi)\nabla\delta_h^{n+1}\cdot e_{s,h}^{n+1}$ and $\phi\,\nabla\delta_h^{n+1}\cdot e_{f,h}^{n+1}$ cancel out, so that

$$\mathcal{T}_p^{n+1} = -\Delta t\int_\Omega (1-\phi)\nabla p_h(t^{n+1})\cdot(\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n+1})\,\mathrm{d}x - \Delta t\int_\Omega \phi\,\nabla p_h(t^{n+1})\cdot(\widetilde{e}_{f,h}^{n+1} - e_{f,h}^{n+1})\,\mathrm{d}x.$$

These two terms can then be controled using the numerical dissipation coming from the correction step. Indeed, Young inequality implies that

$$\begin{aligned}-\Delta t\int_\Omega (1-\phi)\nabla p_h(t^{n+1})\cdot(\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n+1})\,\mathrm{d}x \le \frac{\gamma_1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n+1}\right|^2\mathrm{d}x \\ + \frac{\Delta t^2}{2\gamma_1\rho_s}\int_\Omega (1-\phi)\left|\nabla p_h(t^{n+1})\right|^2\mathrm{d}x,\end{aligned}$$

for all $\gamma_1 > 0$. Handling the fluid term in the same way, it follows that

$$\mathcal{T}_p^{n+1} \le \frac{\gamma_1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n+1}\right|^2\mathrm{d}x + \frac{\gamma_2}{2}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1} - e_{f,h}^{n+1}\right|^2\mathrm{d}x + C\Delta t^2\left\|\nabla p_h(t^{n+1})\right\|^2, \tag{4.75}$$

where $\gamma_2 > 0$ and C is a positive constant depending only on γ_1, γ_2, ρ_s, ρ_f and ϕ.

For the displacement residual term $\mathcal{T}_u^{n+1}$, in virtue of the consistency estimate (4.69), we have

$$\begin{aligned}\mathcal{T}_u^{n+1} &\le \frac{\Delta t}{2}\int_\Omega \sigma_s\big(R_{u,h}^{n+1}\big) : \varepsilon(R_{u,h}^{n+1})\,\mathrm{d}x + \frac{\Delta t}{2}\int_\Omega \sigma_s\big(e_{u,h}^{n+1}\big) : \varepsilon(e_{u,h}^{n+1})\,\mathrm{d}x \\ &\le C\Delta t(\Delta t + h^\ell + h^r)^2 + \Delta t E_h^{n+1}.\end{aligned} \tag{4.76}$$

Then, the fluid residual term $\mathcal{T}_f^{n+1}$ is controled thanks to the fluid viscous dissipation by using Young inequality together with Korn inequality as follows:

$$\begin{aligned}\mathcal{T}_f^{n+1} &\le \gamma_3\Delta t\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}) : \varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x + C\Delta t\int_\Omega \rho_f\phi\left|R_{f,h}^{n+1}\right|^2\mathrm{d}x \\ &\le \gamma_3\Delta t\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}) : \varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x + C\Delta t(\Delta t + h^\ell + h^r)^2,\end{aligned} \tag{4.77}$$

with $\gamma_3 > 0$ and C a positive constant depending on γ_3, μ_f, ϕ and Ω. Finally, let us consider the solid residual term $\mathcal{T}_s^{n+1}$, which we decompose as

$$\mathcal{T}_s^{n+1} = \Delta t\int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot(\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n})\,\mathrm{d}x + \Delta t\int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot e_{s,h}^{n}\,\mathrm{d}x.$$

The first term is controled with the help of the solid numerical dissipation by writing

$$\begin{aligned}\Delta t\int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot(\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n})\,\mathrm{d}x \le \frac{\gamma_4}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n}\right|^2\mathrm{d}x \\ + \frac{\Delta t^2}{2\gamma_4}\int_\Omega \rho_s(1-\phi)\left|R_{s,h}^{n+1}\right|^2\mathrm{d}x\end{aligned}$$

and the second term is controled as in (4.76), which results in

$$\Delta t\int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot e_{s,h}^{n}\,\mathrm{d}x \le \frac{\Delta t}{2}\int_\Omega \rho_s(1-\phi)\left|R_{s,h}^{n+1}\right|^2\mathrm{d}x + \Delta t E_h^n.$$

Recalling (4.69), we deduce

$$\mathcal{T}_s^{n+1} \le \frac{\gamma_4}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1} - e_{s,h}^{n}\right|^2\mathrm{d}x + \Delta t E_h^n + C\Delta t(\Delta t + h^\ell + h^r)^2. \tag{4.78}$$

Now, gathering (4.75), (4.76), (4.77) and (4.78) into (4.74), it follows that

$$
\begin{aligned}
(E_h^{n+1}-E_h^n)+\frac{1}{2}\int_\Omega \sigma_s(e_{u,h}^{n+1}-e_{u,h}^n):\varepsilon(e_{u,h}^{n+1}-e_{u,h}^n)\,\mathrm{d}x+\frac{1-\gamma_4}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1}-e_{s,h}^n\right|^2\mathrm{d}x\\
+\frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1}-e_{f,h}^n\right|^2\mathrm{d}x+\frac{1-\gamma_1}{2}\int_\Omega \rho_s(1-\phi)\left|e_{s,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right|^2\mathrm{d}x\\
+\frac{1-\gamma_2}{2}\int_\Omega \rho_f\phi\left|e_{f,h}^{n+1}-\widetilde{e}_{f,h}^{n+1}\right|^2\mathrm{d}x+(1-\gamma_3)\Delta t\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}):\varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x\\
+\Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right)^2\mathrm{d}x\le C\Delta t(\Delta t+h^\ell+h^r)^2\\
+\Delta t(E_h^{n+1}+E_h^n)+C\Delta t^2\left\|\nabla p_h(t^{n+1})\right\|^2.
\end{aligned}
$$

Setting $\gamma_1=\gamma_2=\gamma_3=\gamma_4=\frac{1}{2}$ and summing between 0 and $N-1$ yields

$$
\begin{aligned}
E_h^N+\frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \sigma_s(e_{u,h}^{n+1}-e_{u,h}^n):\varepsilon(e_{u,h}^{n+1}-e_{u,h}^n)\,\mathrm{d}x+\frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1}-e_{s,h}^n\right|^2\mathrm{d}x\\
+\frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1}-e_{f,h}^n\right|^2\mathrm{d}x+\frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|e_{s,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right|^2\mathrm{d}x\\
+\frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|e_{f,h}^{n+1}-\widetilde{e}_{f,h}^{n+1}\right|^2\mathrm{d}x+\frac{\Delta t}{2}\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}):\varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x\\
+\Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right)^2\mathrm{d}x\le E_h^0+C(\Delta t+h^\ell+h^r)^2\\
+\Delta t\sum_{n=0}^{N-1}(E_h^{n+1}+E_h^n)+C\Delta t^2\sum_{n=0}^{N-1}\left\|\nabla p_h(t^{n+1})\right\|^2. \qquad (4.79)
\end{aligned}
$$

To conclude, let us focus on the pressure term remaining in the right-hand side of the above estimate, which is responsible for the loss of the scheme's first-order convergence in time. Using Poincaré-Wirtinger inequality together with the regularity assumption made on the solution, we observe that

$$
\begin{aligned}
C\Delta t^2\sum_{n=0}^{N-1}\left\|\nabla p_h(t^{n+1})\right\|^2&\le C\Delta t^2\sum_{n=0}^{N-1}\left\|\nabla p_h(t^{n+1})-\nabla p(t^{n+1})\right\|^2+C\Delta t^2\sum_{n=0}^{N-1}\left\|\nabla p(t^{n+1})\right\|^2\\
&\le C\Delta t^2\sum_{n=0}^{N-1}\left\|p_h(t^{n+1})-p(t^{n+1})\right\|^2+C\Delta t\cdot\Delta t\sum_{n=0}^{N-1}\left\|\nabla p(t^{n+1})\right\|^2\\
&\le C\Delta t(h^\ell+h^r)^2+C\Delta t\|p\|^2_{\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))}.
\end{aligned}
$$

Since $\Delta t<1$, estimate (4.73) then follows from (4.79) and discrete Grönwall Lemma. □

Remark 4.15. Another possible strategy to control the solid residual term $\mathcal{T}_s^{n+1}$ is to use the decomposition

$$
\mathcal{T}_s^{n+1}=\Delta t\int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot(\widetilde{e}_{s,h}^{n+1}-\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x+\Delta t\int_\Omega \rho_s(1-\phi)R_{s,h}^{n+1}\cdot\widetilde{e}_{f,h}^{n+1}\,\mathrm{d}x.
$$

The first term can be estimated using the friction dissipation term, while the second can be bounded thanks to the viscous fluid dissipation. Nevertheless, this leads to an estimate with a constant C that depends on $k_{\max}^{-1}$, which is often large in practice. Note moreover that in the parabolic/parabolic case $\eta>0$ this difficulty does not appear since $\mathcal{T}_s^{n+1}$ can then be directly controled using the viscous solid dissipation.

Therefore, we have proved that the non-incremental version of Scheme 9 shows a $\mathcal{O}(\sqrt{\Delta t})$ convergence in time. The next theorem establishes that the incremental version of the scheme enhances it by recovering first-order convergence.

Theorem 4.16. *Assume that $i = 1$ and that the solution of Problem (4.1) is regular enough, in particular that $\partial_t p \in \mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ and $p(0) \in \mathrm{H}^1(\Omega)$. If $\Delta t < 1$ and if the initialization of Scheme 9 satisfies (4.72), then for all $0 \le N \le n_T$, it holds*

$$
\begin{aligned}
E_h^N + \frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla \delta_h^N\right|^2 \mathrm{d}x + \frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \sigma_s(e_{u,h}^{n+1}-e_{u,h}^n):\varepsilon(e_{u,h}^{n+1}-e_{u,h}^n)\,\mathrm{d}x \\
+ \frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1}-e_{s,h}^n\right|^2 \mathrm{d}x + \frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1}-e_{f,h}^n\right|^2 \mathrm{d}x \\
+ \frac{\Delta t}{2}\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}):\varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x + \Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right)^2 \mathrm{d}x \\
\le C(h^\ell + h^r)^2 + C\Delta t^2, \quad (4.80)
\end{aligned}
$$

with $C > 0$ a constant independent of h and Δt.

Proof. The proof is very close to that of Theorem 4.14. Let us denote by

$$\psi_h^n = p_h(t^{n+1}) - p_h^n$$

the pressure term involved in the prediction step (4.70). Then, the pressure term involved in the correction step (4.71) can be rewritten as

$$p_h^n - p_h^{n+1} = p_h^n - p_h(t^{n+1}) + p_h(t^{n+1}) - p_h^{n+1} = \delta_h^{n+1} - \psi_h^n.$$

Therefore, testing (4.70) by $(w_{s,h}, d_{s,h}, w_{f,h}) = \Delta t\,(\widetilde{e}_{s,h}^{n+1}, e_{u,h}^{n+1}, \widetilde{e}_{f,h}^{n+1})$ and (4.71) successively by $(w_{s,h}, w_{f,h}, q_h) = \Delta t\,(e_{s,h}^{n+1}, e_{f,h}^{n+1}, \delta_h^{n+1})$ and $(w_{s,h}, w_{f,h}, q_h) = \left(\rho_s^{-1}\Delta t^2\,\nabla\psi_h^n, \rho_f^{-1}\Delta t^2\,\nabla\psi_h^n, 0\right)$, we obtain as in (4.43) that

$$
\begin{aligned}
\left(E_h^{n+1} - E_h^n\right) + \frac{1}{2}\int_\Omega \sigma_s(e_{u,h}^{n+1}-e_{u,h}^n):\varepsilon(e_{u,h}^{n+1}-e_{u,h}^n)\,\mathrm{d}x + \frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1}-e_{s,h}^n\right|^2 \mathrm{d}x \\
+ \frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1}-e_{f,h}^n\right|^2 \mathrm{d}x + \Delta t\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}):\varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x + \Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right)^2 \mathrm{d}x \\
+ \frac{\Delta t^2}{2}\left(\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^{n+1}\right|^2 \mathrm{d}x - \int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\psi_h^n\right|^2 \mathrm{d}x\right) = \mathcal{T}_u^{n+1} + \mathcal{T}_s^{n+1} + \mathcal{T}_f^{n+1}. \quad (4.81)
\end{aligned}
$$

The displacement and velocities residual terms $\mathcal{T}_u^{n+1}$, $\mathcal{T}_s^{n+1}$ and $\mathcal{T}_f^{n+1}$ are estimated exactly as in the proof of Theorem 4.14. The only terms requiring a special attention are the new pressure terms appearing in the left-hand side, that will be treated as in [Guermond and Quartapelle, 1998].

Since

$$\psi_h^n = p_h(t^{n+1}) - p_h(t^n) + p_h(t^n) - p_h^n = p_h(t^{n+1}) - p_h(t^n) + \delta_h^n,$$

the inequality $(a+b)^2 \le (1+\gamma)a^2 + \left(1+\frac{1}{\gamma}\right)b^2$ with $\gamma = \Delta t$ yields

$$\Delta t^2\left|\nabla\psi_h^n\right|^2 \le \Delta t^2(1+\Delta t)\left|\delta_h^n\right|^2 + \Delta t(1+\Delta t)\left|\nabla\left(p_h(t^{n+1}) - p_h(t^n)\right)\right|^2.$$

Hence

$$
\begin{aligned}
\frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\psi_h^n\right|^2 \mathrm{d}x \le \frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^n\right|^2 \mathrm{d}x + \frac{\Delta t^3}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^n\right|^2 \mathrm{d}x \\
+ C\Delta t(1+\Delta t)\left\|\nabla\left(p_h(t^{n+1}) - p_h(t^n)\right)\right\|^2,
\end{aligned}
$$

so that (4.81) becomes

$$\begin{aligned}
&\left(E_h^{n+1}-E_h^n\right)+\frac{1}{2}\int_\Omega \sigma_s(e_{u,h}^{n+1}-e_{u,h}^n):\varepsilon(e_{u,h}^{n+1}-e_{u,h}^n)\,\mathrm{d}x+\frac{1}{2}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1}-e_{s,h}^n\right|^2\,\mathrm{d}x\\
&+\frac{1}{2}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1}-e_{f,h}^n\right|^2\,\mathrm{d}x+\Delta t\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}):\varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x+\Delta t\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right)^2\mathrm{d}x\\
&\qquad+\frac{\Delta t^2}{2}\left(\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^{n+1}\right|^2\,\mathrm{d}x-\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^{n}\right|^2\,\mathrm{d}x\right)\le \mathcal{T}_u^{n+1}+\mathcal{T}_s^{n+1}+\mathcal{T}_f^{n+1}\\
&\qquad\qquad+\frac{\Delta t^3}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^{n}\right|^2\,\mathrm{d}x+C\Delta t(1+\Delta t)\left\|\nabla\left(p_h(t^{n+1})-p_h(t^n)\right)\right\|^2.
\end{aligned}$$

Controling $\mathcal{T}_u^{n+1}$, $\mathcal{T}_s^{n+1}$ and $\mathcal{T}_f^{n+1}$ as in Theorem 4.14 and summing the above estimates for n between 0 and $N-1$, we infer

$$\begin{aligned}
&E_h^N+\frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^{N}\right|^2\,\mathrm{d}x+\frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \sigma_s(e_{u,h}^{n+1}-e_{u,h}^n):\varepsilon(e_{u,h}^{n+1}-e_{u,h}^n)\,\mathrm{d}x\\
&\qquad+\frac{1}{4}\sum_{n=0}^{N-1}\int_\Omega \rho_s(1-\phi)\left|\widetilde{e}_{s,h}^{n+1}-e_{s,h}^n\right|^2\,\mathrm{d}x+\frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega \rho_f\phi\left|\widetilde{e}_{f,h}^{n+1}-e_{f,h}^n\right|^2\,\mathrm{d}x\\
&\qquad+\frac{\Delta t}{2}\sum_{n=0}^{N-1}\int_\Omega \phi\,\sigma_f(\widetilde{e}_{f,h}^{n+1}):\varepsilon(\widetilde{e}_{f,h}^{n+1})\,\mathrm{d}x+\Delta t\sum_{n=0}^{N-1}\int_\Omega \phi^2 k_f^{-1}\left(\widetilde{e}_{f,h}^{n+1}-\widetilde{e}_{s,h}^{n+1}\right)^2\mathrm{d}x\\
&\qquad\le E_h^0+\frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^{0}\right|^2\,\mathrm{d}x+C(\Delta t+h^\ell+h^r)^2+\Delta t\sum_{n=0}^{N-1}(E_h^{n+1}+E_h^n)\\
&\qquad\qquad+\Delta t\sum_{n=0}^{N-1}\frac{\Delta t^2}{2}\int_\Omega \rho_{\mathrm{eff}}^{-1}\left|\nabla\delta_h^{n}\right|^2\,\mathrm{d}x+C\Delta t(1+\Delta t)\sum_{n=0}^{N-1}\left\|\nabla\left(p_h(t^{n+1})-p_h(t^n)\right)\right\|^2.
\end{aligned}$$

To retrieve (4.80), we note that

$$\begin{aligned}
\sum_{n=0}^{N-1}\left\|\nabla\left(p_h(t^{n+1})-p_h(t^n)\right)\right\|^2&=\sum_{n=0}^{N-1}\left\|\nabla\left(p_h(t^{n+1})-p(t^{n+1})+p(t^{n+1})-p(t^n)+p(t^n)-p_h(t^n)\right)\right\|^2\\
&\le CN(h^\ell+h^r)^2+C\Delta t\|\partial_t p\|^2_{\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))}
\end{aligned}$$

in view of (4.66), and thus

$$C\Delta t(1+\Delta t)\sum_{n=0}^{N-1}\left\|\nabla\left(p_h(t^{n+1})-p_h(t^n)\right)\right\|^2\le C(h^\ell+h^r)^2+C\Delta t^2.$$

The final conclusion once again follows from discrete Grönwall Lemma. □

4.5 Numerical results

In this section, we present simulations to validate numerically the schemes previously analyzed. The simulations were performed using the FEniCS software [Logg et al., 2012; Alnæs et al., 2015]. All types of boundary conditions are investigated, starting with Dirichlet boundary conditions.

4.5.1 Dirichlet boundary conditions

The implementation of the projection scheme is validated thanks to the manufactured solution method in the unit square domain $\Omega = (0,1)^2$. To build an analytical solution, we proceed as in Chapter 3. We consider the function

$$v^{\mathrm{ref}}(x,y) = \Big(\sin(2\pi y)\big(\cos(2\pi x)-1\big), \sin(2\pi x)\big(1-\cos(2\pi y)\big)\Big),$$

which verifies $\operatorname{div} v^{\mathrm{ref}} = 0$ in Ω and $v^{\mathrm{ref}} = 0$ on $\partial\Omega$. Assuming that the porosity is constant, we set

$$v_s^{\mathrm{ref}}(x,y,t) = t\phi\, v^{\mathrm{ref}}(x,y) \quad \text{and} \quad v_f^{\mathrm{ref}}(x,y,t) = t(1-\phi)\, v^{\mathrm{ref}}(x,y),$$

so that

$$\operatorname{div}\big((1-\phi)v_s^{\mathrm{ref}} + \phi\, v_f^{\mathrm{ref}}\big) = t\phi(1-\phi)\operatorname{div} v^{\mathrm{ref}} = 0.$$

For the pressure analytical solution, we take $p^{\mathrm{ref}}(x,y,t) = t\sin(2\pi x)\sin(2\pi y)$. In order to satisfy the inf-sup condition (4.63), the solid and fluid parts are discretized using $[\mathbb{P}^2]^d$ elements, and the pressure with $\mathbb{P}^1$ elements. The projection scheme is implemented with the Poisson-formulation of the correction step. Hence, denoting by f_s^{ref} and f_f^{ref} the right-hand side computed from the analytical solution, the weak formulation of Scheme 1 associated with the manufactured solution above reads:

Step 1: `(prediction step)`

- **Step 1.1**: `(structure prediction sub-step)`
  Find $u_{s,h}^{n+1} \in X_h$ and $\widetilde{v}_{s,h}^{n+1} \in X_h$, $\forall w_{s,h} \in X_h, \forall d_{s,h} \in X_h$,

$$\begin{aligned}\int_\Omega \rho_s(1-\phi)\frac{\widetilde{v}_{s,h}^{n+1} - v_{s,h}^n}{\Delta t}\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \sigma_s\Big(\frac{u_{s,h}^{n+1}-u_{s,h}^n}{\Delta t}\Big) : \varepsilon(d_{s,h})\,\mathrm{d}x \\ - \int_\Omega \sigma_s(\widetilde{v}_{s,h}^{n+\frac{1}{2}\sharp}) : \varepsilon(d_{s,h})\,\mathrm{d}x + \int_\Omega \sigma_s(u_{s,h}^{n+\frac{1}{2}}) : \varepsilon(w_{s,h})\,\mathrm{d}x \\ - \int_\Omega \phi^2 k_f^{-1}(v_{f,h}^n - \widetilde{v}_{s,h}^{n+\frac{1}{2}\sharp})\cdot w_{s,h}\,\mathrm{d}x = \int_\Omega f_s^{\mathrm{ref}}(t^{n+1})\cdot w_{s,h}\,\mathrm{d}x.\end{aligned}$$

- **Step 1.2**: `(fluid prediction sub-step)`
  Find $\widetilde{v}_{f,h}^{n+1} \in X_h$, $\forall w_{f,h} \in X_h$,

$$\begin{aligned}\int_\Omega \rho_f\phi\frac{\widetilde{v}_{f,h}^{n+1} - v_{f,h}^n}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x + \int_\Omega \phi\,\sigma_f(\widetilde{v}_{f,h}^{n+1}) : \varepsilon(w_{f,h})\,\mathrm{d}x \\ + \int_\Omega \phi^2 k_f^{-1}(\widetilde{v}_{f,h}^{n+1} - \widetilde{v}_{s,h}^{n+\frac{1}{2}\sharp})\cdot w_{f,h}\,\mathrm{d}x = \int_\Omega f_f^{\mathrm{ref}}(t^{n+1})\cdot w_{f,h}\,\mathrm{d}x.\end{aligned}$$

Step 2: `(pressure step)`
Find $p_h^{n+1} \in Q_h$, $\forall q_h \in Q_h$,

$$\int_\Omega \rho_{\mathrm{eff}}^{-1}\nabla p_h^{n+1}\cdot\nabla q_h\,\mathrm{d}x = -(\Delta t)^{-1}\int_\Omega \operatorname{div}\big((1-\phi)\widetilde{v}_{s,h}^{n+1} + \phi\,\widetilde{v}_{f,h}^{n+1}\big)q_h\,\mathrm{d}x.$$

Step 3: `(correction step)`

– **Step 3.1**: `(solid correction sub-step)`
Find $v_{s,h}^{n+1} \in Y_{s,h}$ such that, $\forall w_{s,h} \in Y_{s,h}$,

$$\int_\Omega \rho_s(1-\phi) v_{s,h}^{n+1} \cdot w_{s,h}\,\mathrm{d}x = \int_\Omega \rho_s(1-\phi)\widetilde{v}_{s,h}^{n+1} \cdot w_{s,h}\,\mathrm{d}x - \Delta t \int_\Omega (1-\phi)\nabla p_h^{n+1} \cdot w_{s,h}.$$

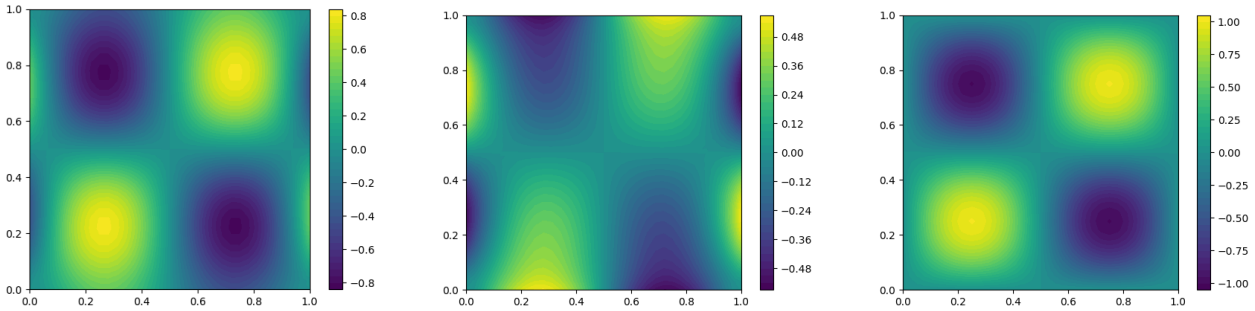
– **Step 3.2**: `(fluid correction sub-step)`
Find $v_{f,h}^{n+1} \in Y_{f,h}$ such that, $\forall w_{f,h} \in Y_{f,h}$,

$$\int_\Omega \rho_f \phi\, v_{f,h}^{n+1} \cdot w_{f,h}\,\mathrm{d}x = \int_\Omega \rho_f \phi\, \widetilde{v}_{f,h}^{n+1} \cdot w_{f,h}\,\mathrm{d}x - \Delta t \int_\Omega \phi\, \nabla p_h^{n+1} \cdot w_{f,h}.$$

The simulation is run with the following parameters: $\phi = 0.5$, $\rho_f = \rho_s = 10^3$, $\mu_f = \lambda = \mu = 1$ and $k_f = 10^{-6} I$. From Theorem 4.5, the stability of the scheme is guaranteed as long as the time step condition (4.25) is fulfilled, namely

$$\Delta t < \left(\frac{\rho_f \rho_s (1-\phi_{\max})}{2\phi_{\max}^3 (k_{\max}^{-1})^2} \right)^{1/2} = 2 \times 10^{-3}.$$

The importance of the time step condition is highlighted by Figure 4.1. As a matter of fact, if the time step condition is not satisfied, Figure 4.1b shows that the pressure profile is far from the reference solution.

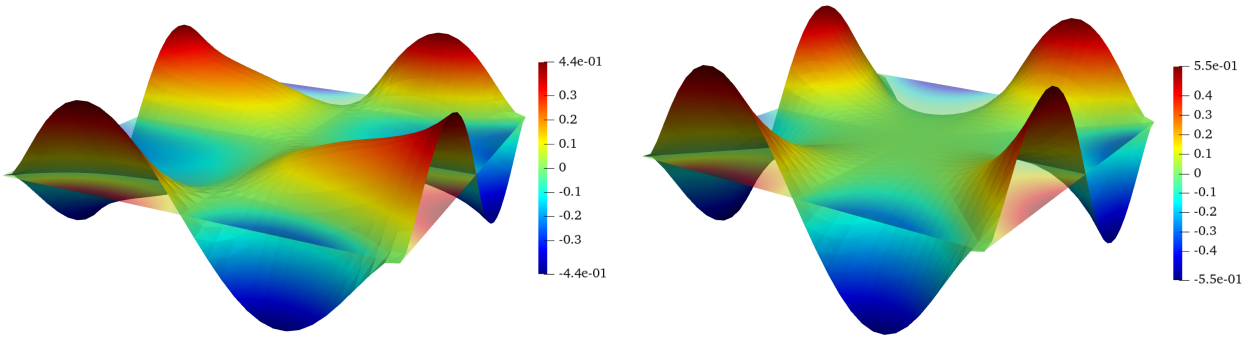


(a) $\Delta t = 10^{-3}$ (b) $\Delta t = 10^{-2}$ (c) Reference solution

Figure 4.1 – Pressure profile at $T = 1$ for different values of time step compared to the analytical solution.

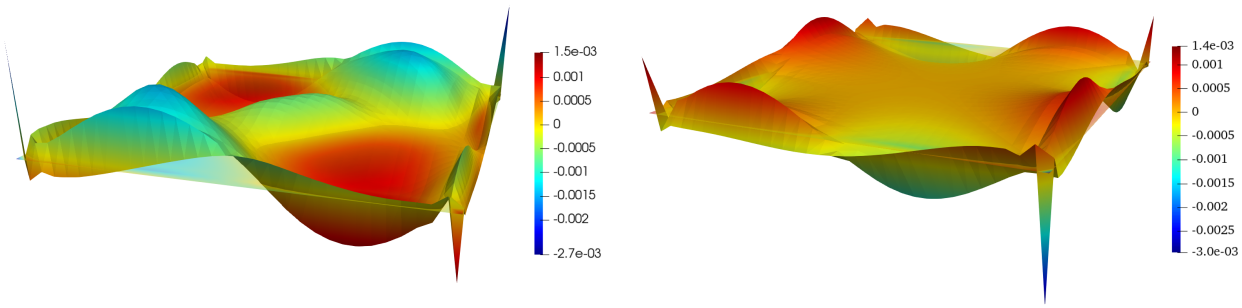
In Figure 4.1a, the pressure profile is closer to the reference solution, but we note that the pressure is badly approximated on the boundaries. This is confirmed by Figure 4.2, which plots the difference between the numerical and analytical solutions. The error is smaller for the incremental version of the scheme, see Figure 4.2b, but still large. In fact, this large error comes from the explicit treatment of the permeability term in the solid prediction sub-step. Indeed, if it is treated implicitly, Figure 4.3 shows that the error is considerably reduced. Moreover, we observe on Figure 4.3 that the error is mainly located in the corners of the domain and in the boundaries. The error occuring at the corners is probably due to the lack of regularity of the domain. As for the error on the boundaries, it is a consequence of the non-physical pressure boundary condition $\rho_{\text{eff}}^{-1} \nabla p_h^{n+1} \cdot n = 0$, which is known to induce a numerical boundary layer in fluid problems [Rannacher, 1992].

Finally, Figure 4.4 emphasizes the importance of the discrete inf-sup condition. Even if none of the steps of the projection method in Poisson form requires solving a saddle-point problem, we have seen in (4.13) that the proposed splitting schemes are consistent with the monolithic algorithm, for which the inf-sup condition is essential. As a result, if the inf-sup condition (4.63) is not satisfied, numerical oscillations appear on the pressure profile of Figure 4.4a.



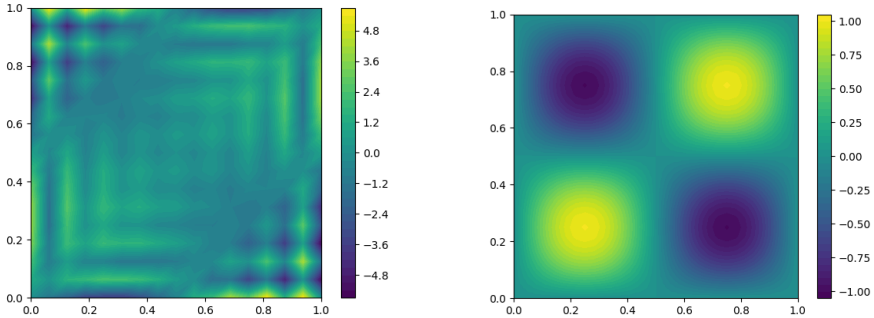
(a) Non-incremental version (b) Incremental version

Figure 4.2 – Pressure error field $p(t^n) - p_h^n$ at $T = 1$ obtained for Schemes 1 and 2.



(a) Non-incremental version (b) Incremental version

Figure 4.3 – Pressure error field $p(t^n) - p_h^n$ at $T = 1$ obtained for Schemes 3 and 4 (scale factor 100).



(a) Solid and fluid quantities discretized with $[\mathbb{P}^1]^d$ elements

(b) Solid and fluid quantities discretized with $[\mathbb{P}^2]^d$ elements

Figure 4.4 – Pressure field at $T = 1$ for different choices of discretization for the solid and fluid quantities.

4.5.2 Neumann boundary conditions

Here, we use the physical parameters from the swelling test case of [Burtschell et al., 2019], namely $\Omega = (0, 0.01)^2$, $\phi = 0.1$, $\rho_f = \rho_s = 10^3$, $\mu_f = 0.035$, $\lambda = 710$, $\mu = 4066$ and $k_f = 10^{-4} I$, with $\Delta t = 10^{-2}$. On the left side of the domain, we impose a growing external pressure, while the pressure is maintained at zero on the right side of the domain and the top and bottom sides are fixed. More precisely, we set

$$u_s = v_s = v_f = 0, \qquad \text{on } (0,1) \times \{0\} \cup (0,1) \times \{1\}, \tag{4.82}$$

and

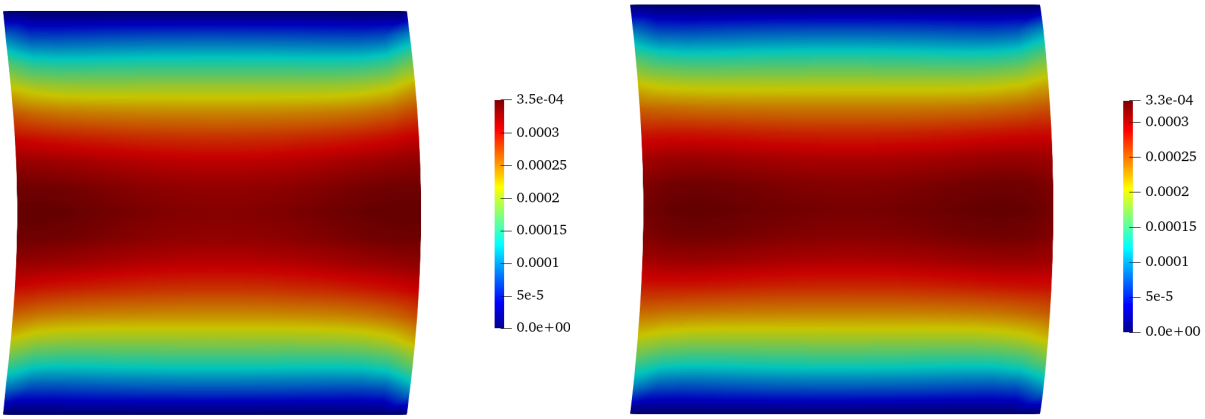
$$\begin{aligned} \sigma_s(u_s)n - (1-\phi)\, pn &= 0, && \text{on } \{1\} \times (0,1), \\ \phi\, \sigma_f(v_f)n - \phi\, pn &= 0, && \text{on } \{1\} \times (0,1), \\ \sigma_s(u_s)n - (1-\phi)\, pn &= -(1-\phi) p_{\text{ext}} n, && \text{on } \{0\} \times (0,1), \\ \phi\, \sigma_f(v_f)n - \phi\, pn &= -\phi\, p_{\text{ext}} n, && \text{on } \{0\} \times (0,1), \end{aligned} \tag{4.83}$$

with

$$p_{\text{ext}}(t) = 1000(1 - e^{-4t^2}).$$

The Dirichlet boundary condition (4.82) is enforced strongly during the prediction step, and the Neumann boundary conditions (4.83) are enforced strongly by imposing $p = 0$ and $p = p_{\text{ext}}$ during the pressure step.

Because of the pressure gradient between the left and right sides, we expect the porous medium to bend to the right. This behaviour is exactly the one observed in Figure 4.5, which also illustrates the accuracy of the projection scheme with respect to the monolithic scheme studied in the previous chapter.



(a) Projection scheme (b) Monolithic scheme

Figure 4.5 – Solid displacement at $T = 1$ resulting from the set of boundary conditions (4.82) – (4.83), computed with Scheme 6 or Scheme (3.16).

4.5.3 Total stress boundary condition

To test the implementation of total stress boundary conditions, we come back to the manufactured solution method of Section 4.5.1, with $\Omega = (0,1)^2$, $\phi = 0.5$, $\rho_f = \rho_s = \mu_f = \lambda = \mu = 1$, $k_f = I$, $\Delta t = 10^{-3}$ and $T = 1$. On the left side of the domain, instead of imposing homogeneous Dirichlet boundary conditions, we set

$$\begin{aligned} v_f &= v_s, && \text{on } \{0\} \times (0,1), \\ \sigma^{\text{tot}} n &= b^{\text{ref}}, && \text{on } \{0\} \times (0,1), \end{aligned}$$

where b^{ref} corresponds to the total stress associated with the reference solution. This total stress boundary condition is imposed with the Robin-Robin coupling method proposed in Scheme 7. On the other sides of the domain, we enforce homogeneous Dirichlet conditions as before.

In Figure 4.6, we see that the Robin coefficient α has a strong influence on the accuracy of the method. In the discrete energy balance (4.53), the Robin-Robin coupling induces artificial energy on the boundary, scaling as $\Delta t(\alpha^{-1} + \alpha)$. Therefore, α must be chosen neither too large nor too small, as reported in [Burman et al., 2022a]. This heuristic reasoning is corroborated by Figure 4.6, which indicates to choose α between 100 and 200 for the considered test case.

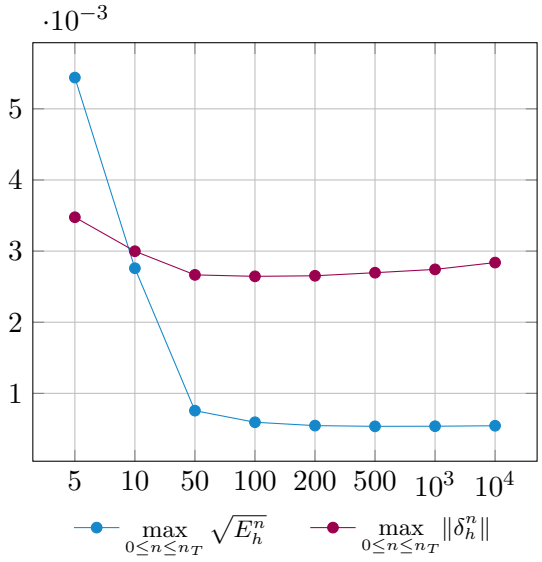


Figure 4.6 – Energy and pressure errors with respect to the Robin coefficient α.

Conclusion

In this chapter, we proposed and analyzed a projection scheme for a linearized incompressible poromechanics problem. Different variants of the scheme were given depending on the treatment of the pressure and permeability terms in the prediction step and on the boundary conditions considered. Stability analysis were performed for all these variants, in particular in the case of total stress boundary conditions for which a Robin-Robin coupling approach allowed us to obtain stability irrespectively of possible added-mass effects. For Dirichlet boundary conditions, a complete convergence analysis was provided and confirmed that the major difference between the incremental and non-incremental versions of the scheme lies in the convergence rate in time of the pressure field. The proposed schemes were implemented and validated on simple test cases. Further perspectives include the use of these splitting schemes in more complex scenarios and a more detailed comparison of their computational efficiency with respect to the monolothic approach from Chapter 3.

CHAPTER 5

Modeling and simulation of artificial microvessel perfusion

This chapter is based on results obtained in collaboration with Claire A. Dessalles (University of Geneva), Céline Grandmont and Philippe Moireau. In this perspective chapter, we explore the relevance of the poromechanis model (13) for simulating experiments with artificial microvessels carried out by Claire A. Dessalles as part of her PhD research at the LadHyX laboratory (École polytechnique, CNRS, Institut Polytechnique de Paris) under the supervision of A. Babataheri and A. Barakat. Several numerical results are provided and compared with experimental data.

Contents

Introduction

This chapter results from a collaboration with Claire A. Dessalles, who was a PhD student at the Hydrodynamics Laboratory of École Polytechnique (LadHyX) and is now a post-doctoral researcher at the University of Geneva. During her PhD thesis [Dessalles, 2021], Claire A. Dessalles developed microfluidics experiments to study the mechanical properties of human vessels. This kind of experiment is called an organ-on-chip platform [Huh et al., 2010; J. Polacheck et al., 2013] since it aims at devicing a microfluidics chip to mimick the behaviour of a human organ, here a microvessel. In [Dessalles et al., 2021], Claire A. Dessalles and co-authors designed a microvessel-on-chip system based on a porous hydrogel. This device sheds light on the key role played by endothelial cells – the cells covering blood vessels, see Figure 5.1 – in microvessels mechanics, whose dysfunction can lead to pathologies such as atherosclerosis [Cunningham and Gotlieb, 2005; Chatzizisis et al., 2007; Hahn and Schwartz, 2009] or brain and lung diseases [O'Rourke and Safar, 2005; Stone et al., 2015; De Montgolfier et al., 2019; Wu and Birukov, 2019]. Moreover, it shows that the mechanical forces involved in microvessels are highly dynamic and very sensitive to the poroelastic behavior of their environment.

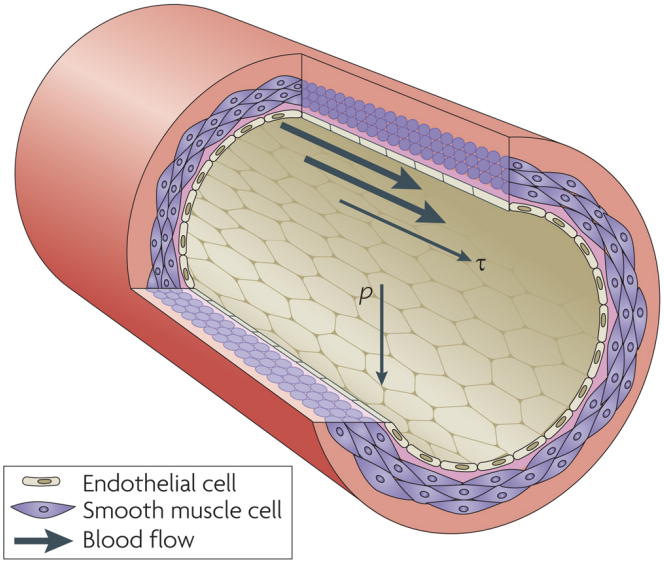


Figure 5.1 – Mechanical forces in blood vessel walls [Hahn and Schwartz, 2009].

The main goal of this chapter is to reproduce and characterize this poroelastic behavior by using the poromechanics model studied in this thesis. A second objective is to go towards the coupling of the model with the blood flow in the microvessel, which requires to consider the interaction between the fluid inside the vessel and the porous structure surrounding it and hence can be seen as a fluid-porous structure interaction problem.

The chapter is organized as follows. In Section 5.1, we briefly recall the microfluidics experiment leading to the microvessel-on-chip platform set up in [Dessalles et al., 2021], and explain how to represent it by the poromechanics model analyzed in this thesis. Then, in Section 5.2, numerical results are presented and compared to experimental data from [Dessalles, 2021]. Finally, Section 5.3 is a perspective section to go towards a fluid-porous structure interaction modeling of the experiment.

5.1 Experimental setup and modeling

5.1.1 Experiments description

The microfluidics experiment carried out by Claire A. Dessalles involves a collagen hydrogel pierced by a microchannel into which water is injected. The hydrogel is contained into a rectangular box with an open top and the water channel has a cylindrical shape, as shown in Figure 5.2. The experimentalist controls the water flow rate imposed at the channel inlet and the pressure at the top of the box as well as at the channel outlet. Then, the velocity profile in the water and the channel deformation can be observed using respectively particle tracking velocimetry and optical coherence tomography tools. In addition, endothelial cells can be deposited on the surface of the channel.

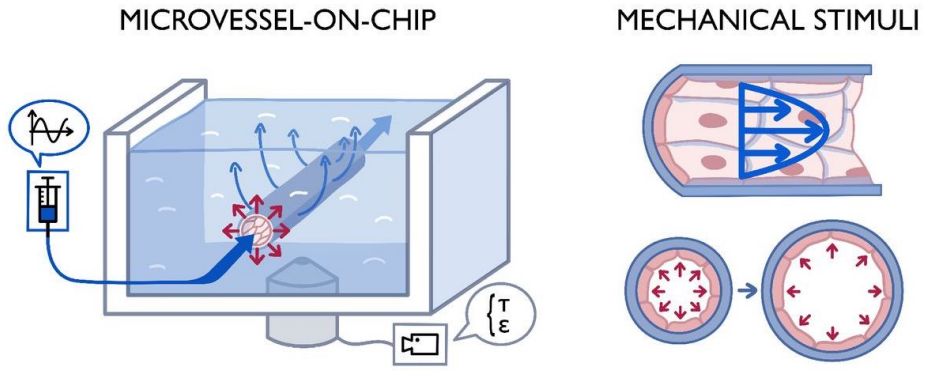


Figure 5.2 – Diagram of the microfluidic chip experimental setup described in [Dessalles et al., 2021].

As depicted in Figure 5.3, the dimensions of the setup are small. In particular, the channel radius is 60 µm. We will see in the next section that this requires to use a proper system of units and a refined mesh close to the channel for the simulation.

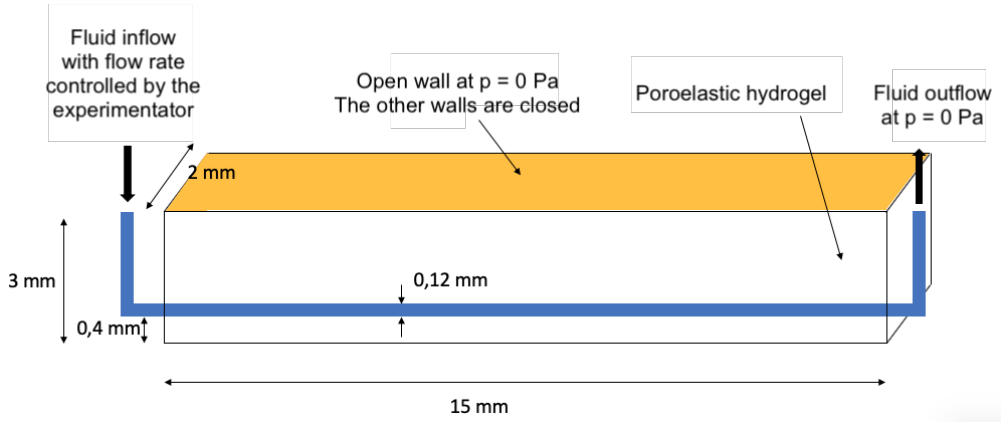


Figure 5.3 – Dimensions of the microfluidics experiment. Courtesy from Claire A. Dessalles.

The hydrogel at stake is made of 99.4% water and a few elastic collagen fibers. It is fabricated by diluting collagen fibers in water. Therefore, when taken out of its box, it is almost liquid, as depicted in Figure 5.4. For more details concerning the hydrogel fabrication, the cells seeding and the imagery techniques employed, we refer the reader to [Dessalles et al., 2021].

Since human vessel membranes are partly composed of collagen, this experiment models a microvessel (the channel) surrounded by a poroelastic biomaterial (the collagen hydrogel), possibly

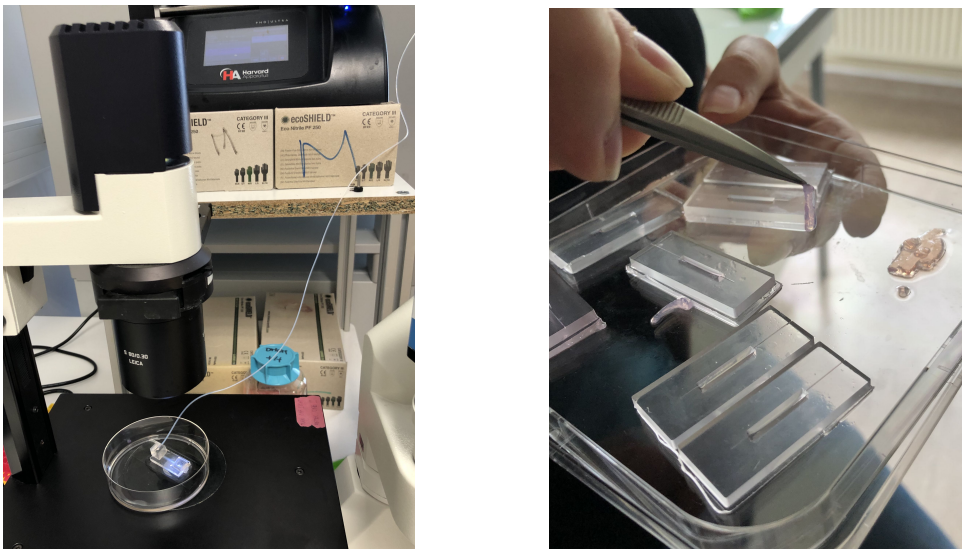


Figure 5.4 – Photograph of the experimental setup (left) and of a hydrogel outside its rectangular box (right).

with a layer of endothelium (cells). The aim of the setup is to observe the influence of water flow rate and cells on the channel deformation and the water circulation.

Parameter	Value	Physical meaning
ϕ	0.994	Porosity
ρ_f	$1.0 \times 10^3\,\mathrm{kg\,m^{-3}}$	Water density
ρ_s	$1.0 \times 10^5\,\mathrm{kg\,m^{-3}}$	Collagen fibers density
μ_f	$1.0 \times 10^{-3}\,\mathrm{Pa\,s}$	Water viscosity
k_f	$2.0 \times 10^{-11}\,I\ \mathrm{m^2\,Pa^{-1}\,s^{-1}}$	Hydraulic conductivity tensor
E	$3900\,\mathrm{Pa}$	Young modulus
ν	0.3	Poisson ratio
α	0.99	Biot-Willis coefficient
κ	$2.0 \times 10^9\,\mathrm{Pa}$	Bulk modulus

Table 5.1 – Physical parameters for the microfluidics experiment.

Table 5.1 reports the typical physical parameters associated with the microvessel-on-chip platform. The porosity value is directly infered from the concentration property of the hydrogel. Since the hydrogel is mostly composed of water, it is nearly-incompressible and the value of the bulk modulus has been chosen to be slightly smaller than the bulk modulus of pure water. The hydraulic conductivity value was calculated in [Dessalles et al., 2021] following an experimental formula from [Huxley et al., 1987] but also thanks to an analytical model. However, the hydrogel mechanical parameters were estimated in [Dessalles, 2021] without a high degree of accuracy, so that one goal of Section 5.2 will be to calibrate these parameters using the porous model studied in the previous chapters together with experimental data. Let us now show how to use this model for the considered experiment.

5.1.2 Porous modeling of the hydrogel

Our strategy is to model only the porous hydrogel involved in the microfluidics experiment. The water inside the channel is taken into account as an external pressure Neumann boundary condition, and the presence of cells on the channel wall is not considered.

Let us denote by Ω the volume occupied by the hydrogel in the microfluidics experiment. It is a rectangular box pierced by a cylinder, as depicted in Figure 5.5. To model the poroelastic behavior of the hydrogel, we consider the poromechanics model studied in this thesis in absence of external body forces and additional fluid mass input, namely

$$\begin{cases} \rho_s(1-\phi)\,\partial_t v_s - \operatorname{div}\big(\sigma_s(u_s)\big) - \phi^2\, k_f^{-1}(v_f - v_s) + (\alpha-\phi)\,\nabla p = 0, & \text{in } \Omega, \\ \rho_f \phi\,\partial_t v_f - \operatorname{div}\big(\phi\,\sigma_f(v_f)\big) + \phi^2\, k_f^{-1}(v_f - v_s) + \phi\,\nabla p = 0, & \text{in } \Omega, \\ \dfrac{\alpha-\phi}{\kappa}\partial_t p + \operatorname{div}\big((\alpha-\phi)\,v_s + \phi\,v_f\big) = 0, & \text{in } \Omega, \end{cases} \tag{5.1}$$

with the usual notation. The problem is complemented with the following boundary conditions. The top of the box is opened and a zero pressure is applied by the experimentalist, so that it can be considered as a free surface. Therefore, denoting by Γ_{top} this surface – see Figure 5.5, we impose

$$\begin{aligned} \sigma_s(u_s)n - (1-\phi)\,pn = 0, \quad & \text{on } \Gamma_{\text{top}}, \\ \phi\,\sigma_f(v_f)n - \phi\,pn = 0, \quad & \text{on } \Gamma_{\text{top}}. \end{aligned} \tag{5.2}$$

On the surface in contact with the water channel, denoted by Γ_{channel}, the hydrogel is subject to a pressure coming from the fluid injected into the channel. The corresponding Neumann boundary condition reads:

$$\begin{aligned} \sigma_s(u_s)n - (1-\phi)\,pn = -(1-\phi)p_{\text{channel}}n, \quad & \text{on } \Gamma_{\text{channel}}, \\ \phi\,\sigma_f(v_f)n - \phi\,pn = -\phi\,p_{\text{channel}}n, \quad & \text{on } \Gamma_{\text{channel}}. \end{aligned} \tag{5.3}$$

Determining exactly the pressure p_{channel} requires simulating the flow inside the channel and will be explored in Section 5.3. Indeed, this pressure depends on the flow rate at the inlet of the channel, on the pressure at its outlet, but also on time and on the channel deformation. Here, we consider a linear pressure along the channel axis that is independent of time, namely

$$p_{\text{channel}}(x,y,z) = \frac{p_{\text{in}} - p_{\text{out}}}{L}(L-z) + p_{\text{out}}, \tag{5.4}$$

where z denotes the direction along the channel axis (with the convention $z=0$ at the inlet), L the channel length, p_{in} the pressure at the channel inlet and $p_{\text{out}} < p_{\text{in}}$ the pressure at its outlet.

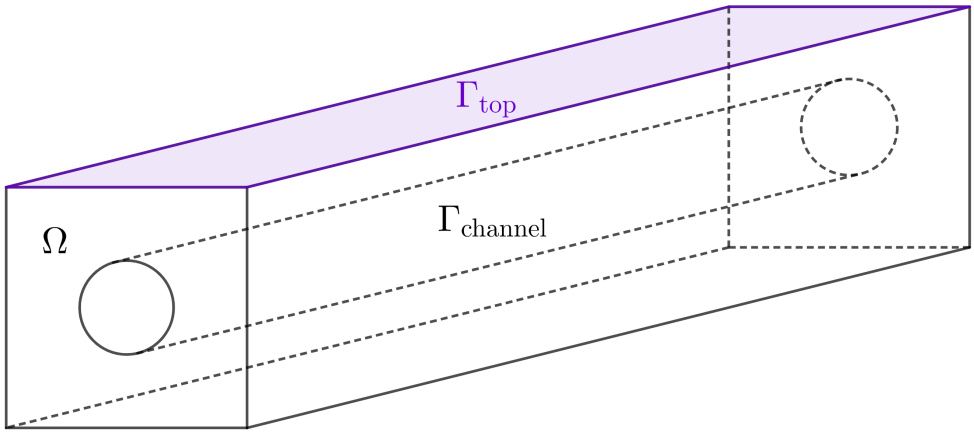


Figure 5.5 – Porous domain and boundaries of the collagen hydrogel.

Finally, since the rest of the box is closed, we set

$$u_s = v_s = v_f = 0, \quad \text{on } \partial\Omega \setminus \Gamma_{\text{top}} \cup \Gamma_{\text{channel}}. \tag{5.5}$$

To solve numerically Problem (5.1) – (5.5), we use the backward Euler scheme presented in Chapter 3, see (3.16). More precisely, at each time step, we solve the variational formulation: find $(u_{s,h}^{n+1}, v_{f,h}^{n+1}, p_h^{n+1}) \in X_h \times X_h \times Q_h$ such that for all $(w_{s,h}, w_{f,h}, q_h) \in X_h \times X_h \times Q_h$,

$$\begin{aligned}
&\int_\Omega \frac{\alpha-\phi}{\kappa}\frac{p_h^{n+1}-p_h^n}{\Delta t} q_h \,\mathrm{d}x + \int_\Omega \rho_f \phi \,\frac{v_{f,h}^{n+1}-v_{f,h}^n}{\Delta t}\cdot w_{f,h}\,\mathrm{d}x \\
&\quad + \int_\Omega \rho_s(1-\phi)\,\frac{2}{\Delta t^2}\left(u_{s,h}^{n+1}-u_{s,h}^n-\Delta t v_{s,h}^n\right)\cdot w_{s,h}\,\mathrm{d}x + \int_\Omega \sigma_s(u_{s,h}^{n+\frac{1}{2}}) : \varepsilon(w_{s,h})\,\mathrm{d}x \\
&\quad + \int_\Omega \phi\,\sigma_f(v_{f,h}^{n+1}) : \varepsilon(w_{f,h})\,\mathrm{d}x + \int_\Omega \phi^2\, k_f^{-1}\left(v_{f,h}^{n+1} - \frac{u_{s,h}^{n+1}-u_{s,h}^n}{\Delta t}\right)\cdot(w_{f,h}-w_{s,h})\,\mathrm{d}x \\
&\quad - \int_\Omega p_h^{n+1}\,\mathrm{div}\left((1-\phi)\,w_{s,h}+\phi\,w_{f,h}\right)\mathrm{d}x + \int_\Omega \mathrm{div}\left((1-\phi)\,\frac{u_{s,h}^{n+1}-u_{s,h}^n}{\Delta t}+\phi\,v_{f,h}^{n+1}\right) q_h\,\mathrm{d}x \\
&\qquad = -\int_{\Gamma_{\text{channel}}} (1-\phi) p_{\text{channel}} n\cdot w_{s,h}\,\mathrm{d}S - \int_{\Gamma_{\text{channel}}} \phi\, p_{\text{channel}} n\cdot w_{f,h}\,\mathrm{d}S, \quad (5.6)
\end{aligned}$$

where X_h and Q_h correspond to the discrete spaces associated with $[\mathbb{P}^2]^d$ and $\mathbb{P}^1$ Lagrange continuous finite elements respectively. The solid velocity is then post-processed node by node thanks to the formula

$$v_{s,h}^{n+1} = 2\frac{u_{s,h}^{n+1}-u_{s,h}^n}{\Delta t} - v_{s,h}^n.$$

The numerical simulations resulting from formulation (5.6) are the object of the next section.

5.2 Numerical results

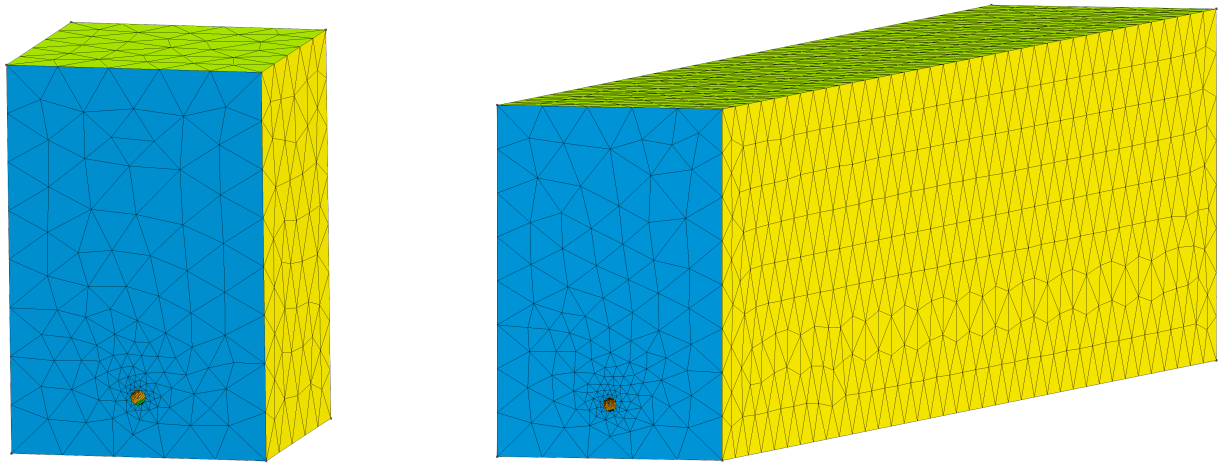
To validate our model, let us first see if the solution computed numerically behaves qualitatively as in the microfluidics experiments.

5.2.1 Validation of the model

The simulation is run with the finite element software FEniCS [Logg et al., 2012; Alnæs et al., 2015] using MUMPS (MUltifrontal Massively Parallel Sparse direct Solver). The mesh is generated using pygmsh [Geuzaine and Remacle, 2009; Schlömer, 2022]. Since the physical phenomenon is mainly happening around the channel, the geometry is meshed more finely close to the channel, see Figure 5.6. Moreover, in order to limit the calculation time, we first run the simulation for a shorter channel of length 2 mm before simulating the experiments channel that is 15 mm long.

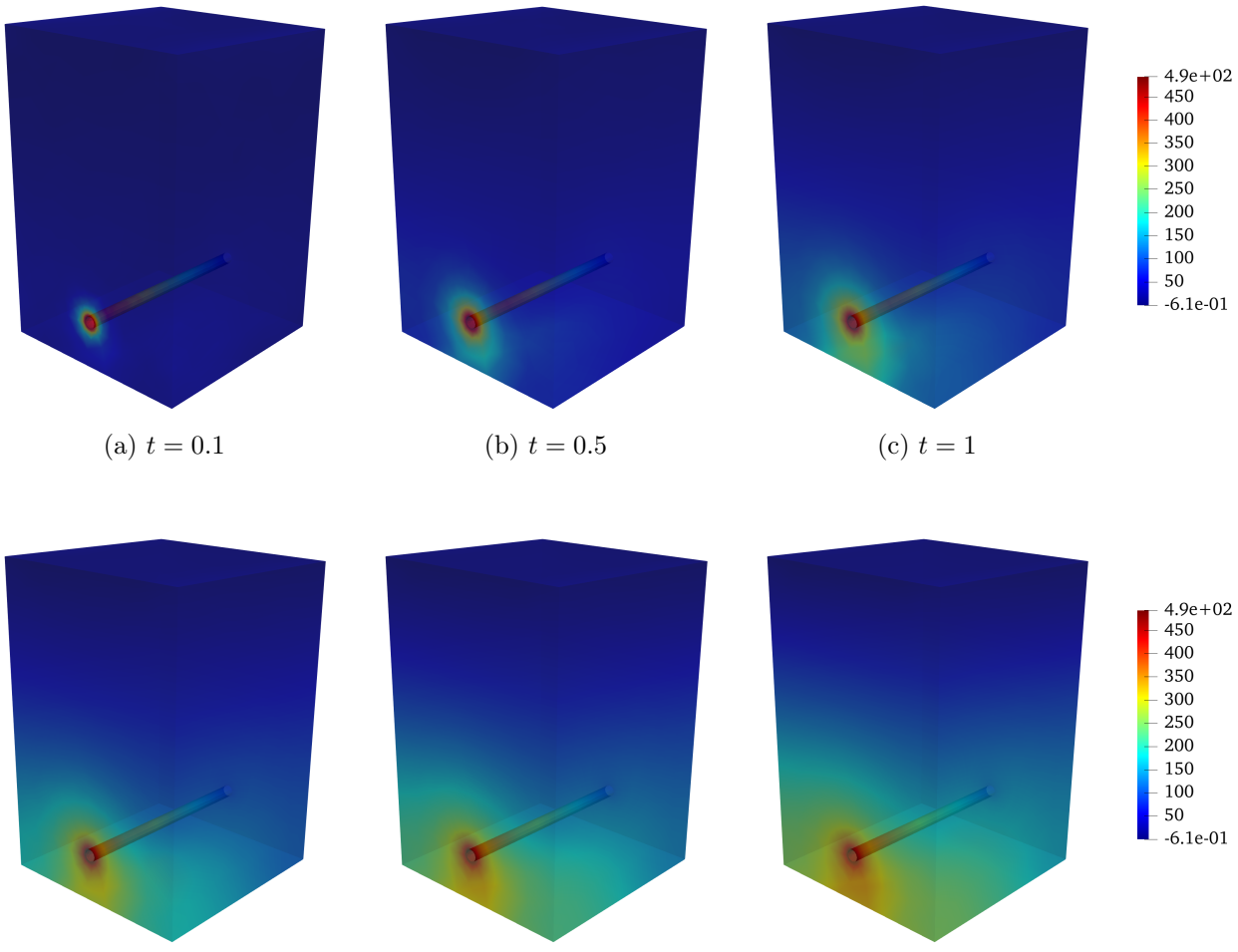
Figures 5.7, 5.8 and 5.9 show the results of the simulation of (5.6) for the short channel ($L = 2\,\text{mm}$) with $p_{\text{in}} = 500\,\text{Pa}$, $p_{\text{out}} = 50\,\text{Pa}$, $\Delta t = 0.1$ and $T = 5$. In Figure 5.7, we see that the pressure progressively increases in the hydrogel, starting around the channel and then spreading into the rest of the domain. Moreover, we note that the pressure is higher under the channel than above. As a consequence, the channel bends upwards and the hydrogel swells to the top, as illustrated in Figure 5.8. In addition, Figure 5.9 shows that the water in the hydrogel flows towards the free surface and that some water escapes at the top of the gel, see Figure 5.9a. This behavior corresponds exactly to the one observed by Claire A. Dessalles during the microfluidics experiments. Furthermore, the order of magnitude of the solid displacement obtained numerically – see the colorbar of Figure 5.8 – is consistent with the average displacement value of $8.1 \times 10^{-6}\,\text{m}$ measured experimentally.

Then, we run the simulation for the experiments channel ($L = 15\,\text{mm}$) with $p_{\text{in}} = 300\,\text{Pa}$, $p_{\text{out}} = 100\,\text{Pa}$, $\Delta t = 0.1$ and $T = 5$. The mesh and all the physical parameters are expressed in CGS units.



(a) $L = 2\,\mathrm{mm}$ (b) $L = 15\,\mathrm{mm}$

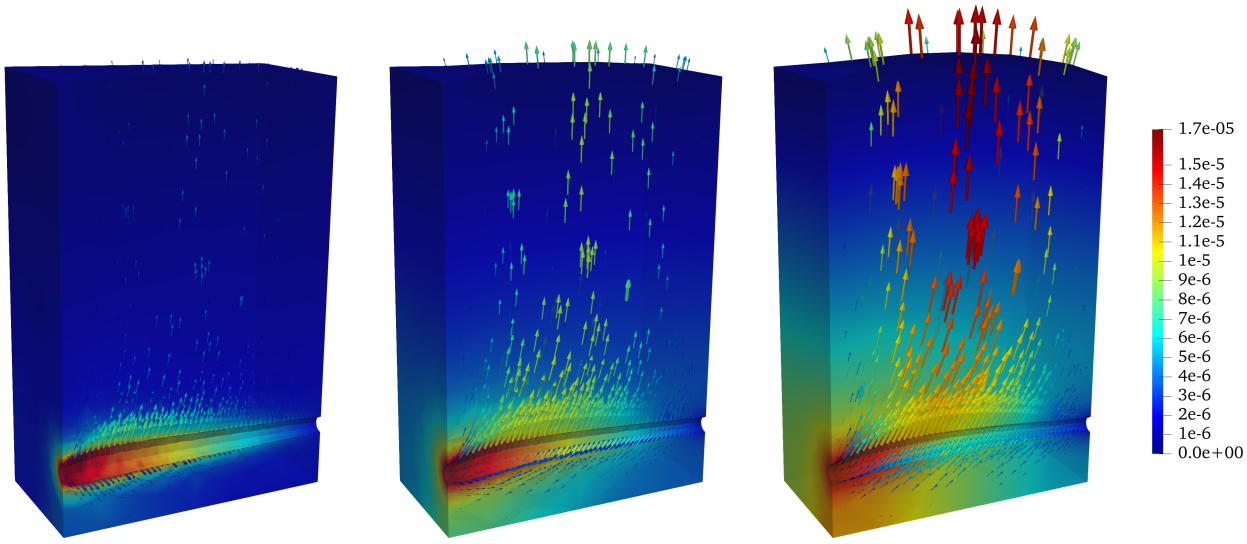
Figure 5.6 – Meshes of the porous hydrogel domain generated for the simulation.



(a) $t = 0.1$ (b) $t = 0.5$ (c) $t = 1$

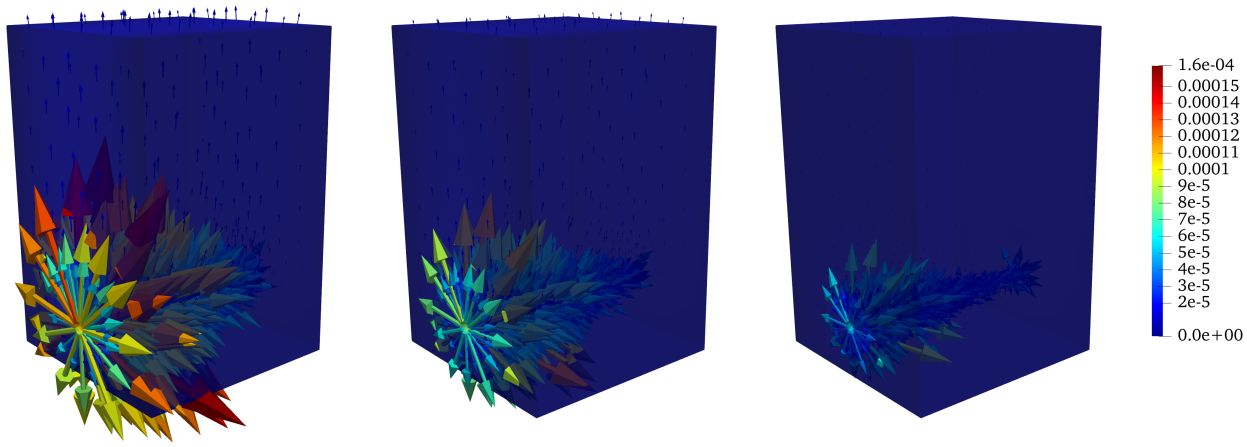
(d) $t = 2$ (e) $t = 3$ (f) $t = 5$

Figure 5.7 – Interstitial pressure in the hydrogel (Pa) over time.



(a) $t = 0.1$ (b) $t = 0.5$ (c) $t = 4$

Figure 5.8 – Deformed channel (scale factor 10) with pressure coloration and solid displacement vector (m).



(a) $t = 0.2$ (b) $t = 0.6$ (c) $t = 5$

Figure 5.9 – Fluid velocity in the hydrogel ($\mathrm{m\,s^{-1}}$) over time.

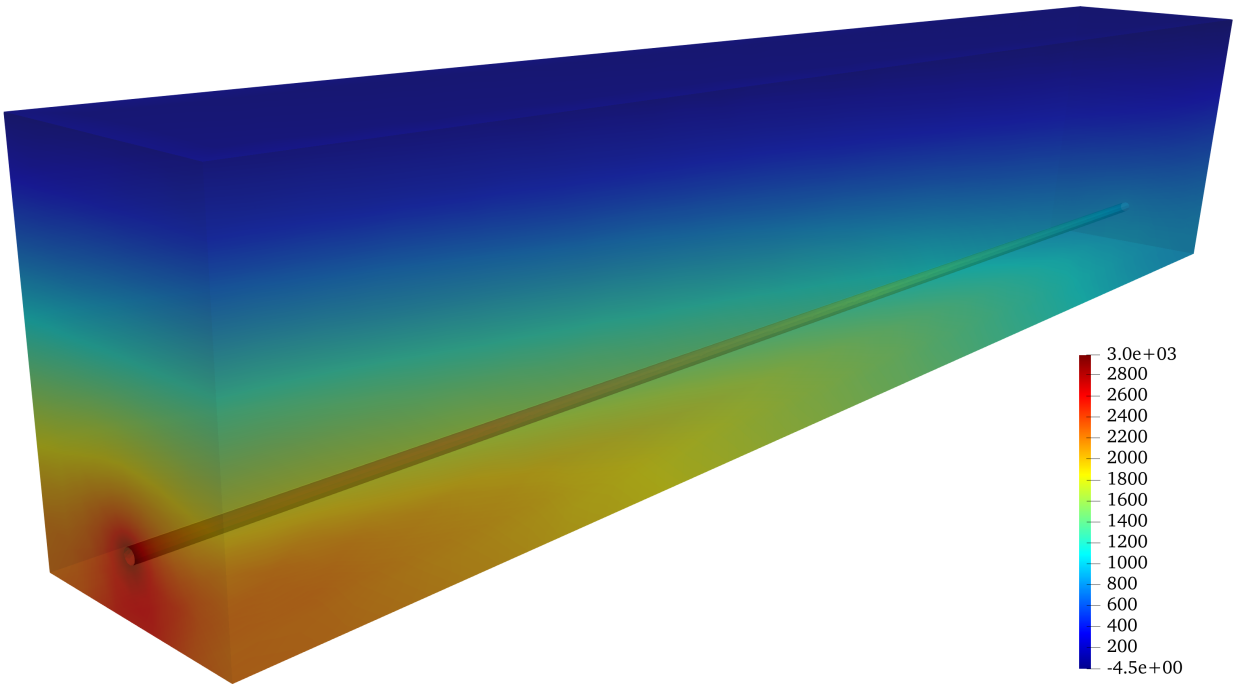
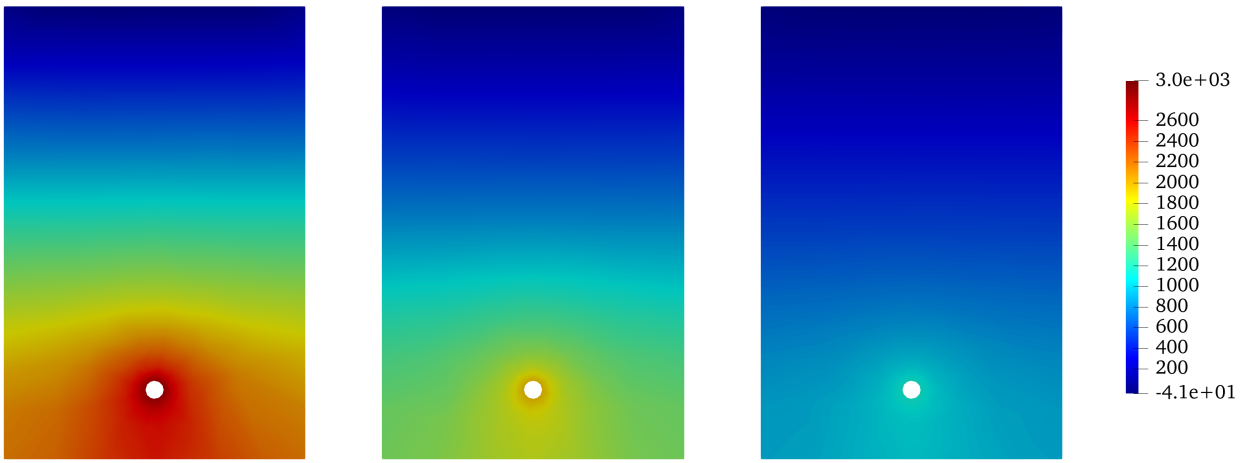


Figure 5.10 – Interstitial pressure in the experiments hydrogel (10^{-1}Pa) at $T = 5$.



(a) $z = 1\,\text{mm}$ (b) $z = 7.5\,\text{mm}$ (c) $z = 14\,\text{mm}$

Figure 5.11 – Slice of pressure (10^{-1}Pa) near the inlet, in the middle of the channel and near the outlet.

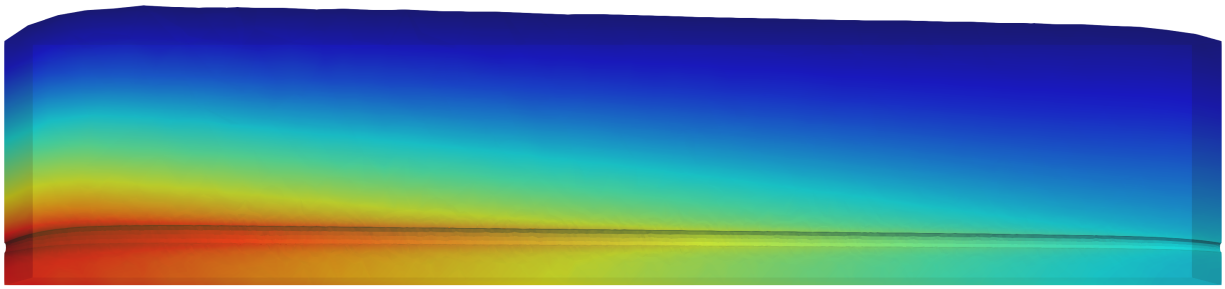


Figure 5.12 – Channel deformation (scale factor 20) with pressure coloration at $T = 5$.

The results of the long channel simulation are shown in Figures 5.10, 5.11 and 5.12. The pressure growth observed in Figure 5.7 for the short channel is confirmed by Figure 5.10. Furthermore, Figure 5.11 highlights that the pressure gradient is not homogeneous along the channel axis: the smaller the pressure value imposed on the channel, the more the pressure in the gel is concentrated around the channel. Finally, Figure 5.12 illustrates the channel deformation and the swelling of the gel towards the free surface on the top. Once again, the channel deformation profile obtained in Figure 5.12 is in accordance with the experimental results, which validates the use of the poromechanics model (5.1) to simulate the microfluidics experiments from [Dessalles et al., 2021].

5.2.2 Pressure ramps

Next, we apply the porous hydrogel modeling (5.1) – (5.3) to the simulation of pressure ramps. Instead of imposing a spatially dependent pressure that is independent of time as in the previous subsection, we study the gel response to a gradual pressure. More precisely, we reproduce a part of the experiment detailed in Figure 5.13, in which the microvessel-on-chip is subject to various loading flow rates at the channel entrance.

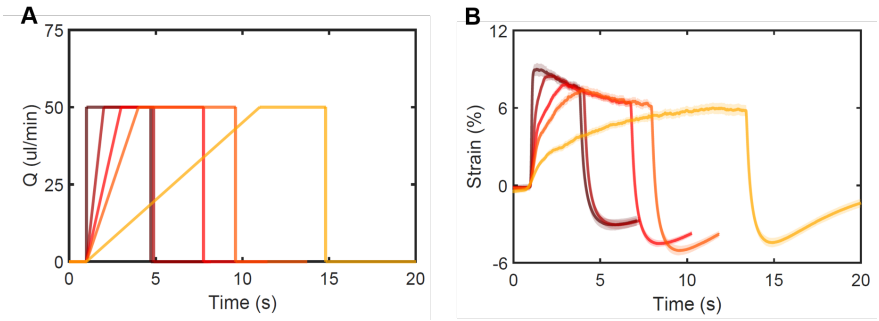


Figure 5.13 – Deformation of the channel as a function of time for different loading rates (color coded). The imposed flow rate is shown in panel A. The loading is progressively slower while the unloading is kept instantaneous. The response of the channel is shown in panel B, with a disappearing overshoot and an increasing undershoot. Courtesy from [Dessalles, 2021].

The associated boundary conditions are similar to the ones proposed in Section 5.1.2. The only difference is the expression of the pressure term p_{channel} appearing in the right-hand side of (5.6), for which we take

$$p_{\text{channel}}(t) = \begin{cases} p_0 t / t_{\text{load}} & \text{if } t \leq t_{\text{load}}, \\ p_0 & \text{if } t > t_{\text{load}}, \end{cases}$$

where t_{load} represents the loading time before reaching the constant value p_0, see Figure 5.14 – left. Compared to Figure 5.13A, this expression focuses on the first part of the imposed flow rate curve, and not on the part where the flow rate falls back to zero. Therefore, we expect to reproduce the first parts of the curve shown in Figure 5.13B. Note however that the pressure load is modeled in a rather crude way. Moreover, note that Figure 5.13B represents the deformation of the channel as a percentage of its initial radius (equal to 60 µm). This quantity is directly proportional to the radial displacement of the channel, which will be the plotted quantity in the numerical simulations that follows.

These curves are useful from an experimental point of view, but also to calibrate mechanical parameters. Indeed, the response curve hides different physical parameters, as we are now going to see it thanks to a parametric study. To analyze the influence of the various parameters on the channel response, we start from the parameters listed in Table 5.1, with $p_0 = 300\,\text{Pa}$, $t_{\text{load}} = 1\,\text{s}$, $\Delta t = 0.1$ and $T = 5$ or $T = 10$. Then, we vary each parameter separately and plot the corresponding radial displacement response curves. These curves are generated at three points of the hydrogel that

are close to the inlet. The first one is located on the channel surface (point A), the second is at the same altitude than the first one but near the right wall (point B) and the third one is located above the first one on the top of the box (point C), see Figure 5.14 – right.

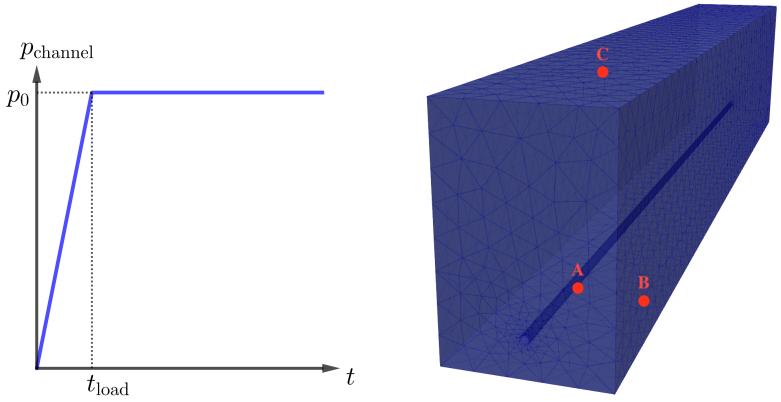


Figure 5.14 – Pressure imposed on the channel (left) and points used for the parametric study (right).

Figure 5.15 sheds light on the influence of the different parameters on the response curve at point A, which is located on the channel and which corresponds to the point where the measures of Figure 5.13B were done. Note that Figure 5.15 represents the x component of the displacement since it corresponds to its radial value. From Figure 5.15, we deduce the following sensitivity analysis.

The Young modulus E mainly plays a role on the amplitude of the response. The stiffer the structure, the lower the gel deforms. In Figure 5.13B, the curve associated with $t_{\text{load}} = 1$ is the second sharpest one. This curve shows a maximum value of 8% of strain, which corresponds to a maximum radial displacement of 4.8 µm. Hence, the best Young modulus fitting this value seems to be $E = 2.5 \times 10^3$ Pa (green curve).

The Poisson ratio of the solid skeleton ν impacts the maximum value of the response, the speed at which the curve goes down, and most importantly the final value of the radial displacement. Note that the Poisson ratio corresponds here to the solid skeleton of the hydrogel, which is a structure full of holes since the hydrogel is mostly composed of water. Therefore, the porous hydrogel quasi-incompressibility does not mean that ν must be close to 0.5. Indeed, the curve for $\nu = 0.49$ (red curve) shows a radial displacement final value that is very close to the peak value, which does not correspond to the observations of Figure 5.13B.

The bulk modulus κ does not affect the response curve significantly. As a matter of fact, the incompressible regime is reached from $\kappa = 100$ Pa, value above which this parameter has no influence on the response.

The intrinsic permeability k influences the response peak, but above all controls the speed at which the curve goes down. The greater the permeability, the faster the response curve descends. Comparing the permeability curves of Figure 5.15 with experimental data shows that the permeability value must be between 10^{-14} m^2 and 5×10^{-14} m^2 (orange and green curves), in accordance with the value $k = 2 \times 10^{-14}$ m^2 found in [Dessalles et al., 2021].

The fluid viscosity has the opposite effect than the permeability. This comes from the relation between the hydraulic conductivity k_f, the fluid viscosity μ_f and the intrinsic permability k, that reads $k_f = k/\mu_f$.

It appears that the porosity ϕ does not modify the response curve. This is probably because E and ν should depend on ϕ since they represent the homogenized skeleton parameters but this dependency is not taken into account in (5.1).

The response curves are not related to the solid and fluid densities ρ_s and ρ_f. For this particular test case, this means that the inertial effects are negligible in comparison to the other terms of the model, in particular the friction effect between the two phases.

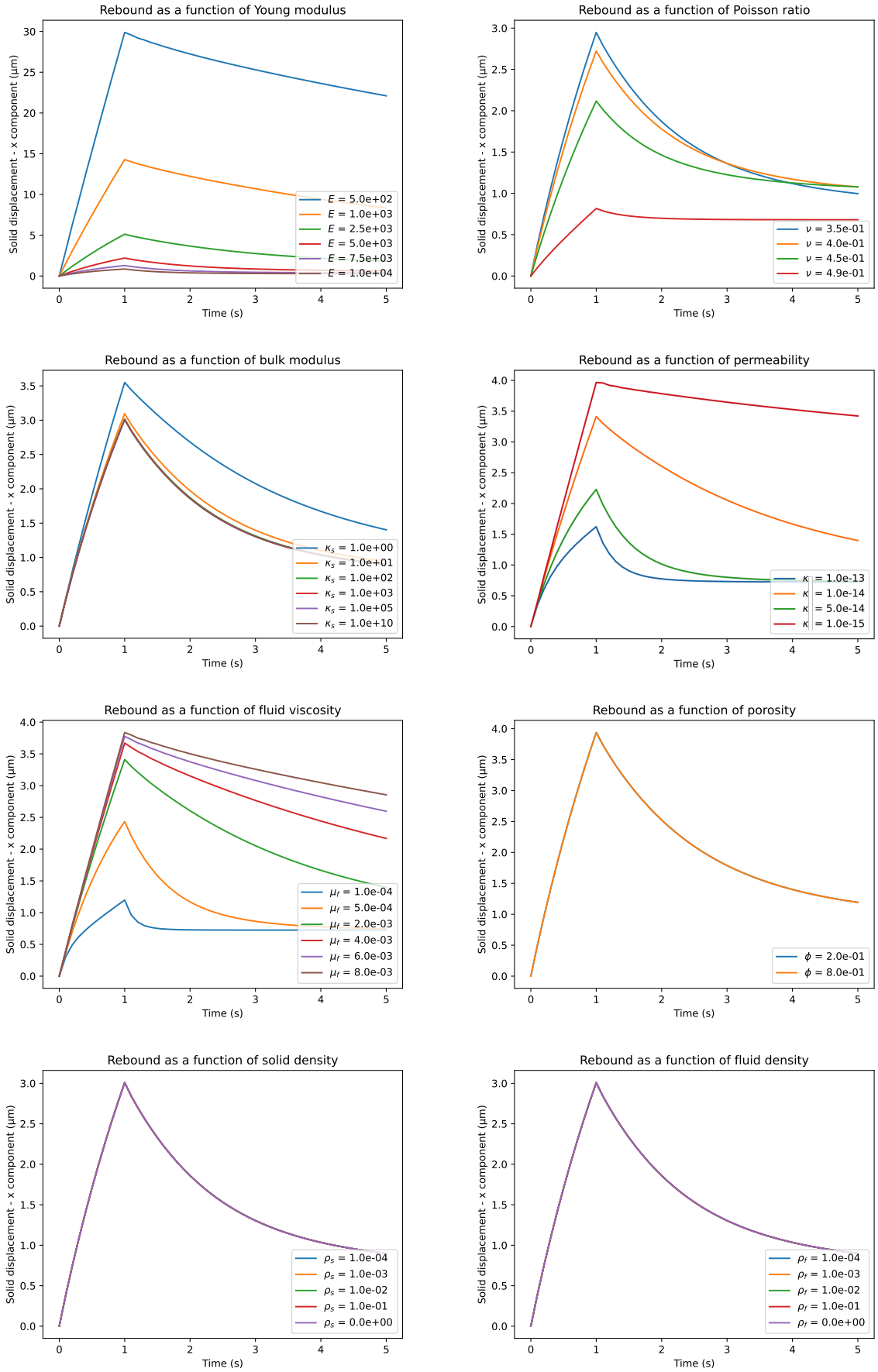


Figure 5.15 – Parametric study of the response curve at point A with respect to the physical parameters.

This sensitivity analysis is confirmed by Figures 5.16 and 5.17, in which the same parametric study is carried out at points B and C, which are respectively located close to the right wall and on the top free surface, see Figure 5.14 – right. Moreover, Figure 5.16 shows that the response peak are shifted in time when the Young modulus and the fluid viscosity (or the permeability) vary, thus highlighting the propagation effects occuring between the channel and the box wall. Independently of the parametric study, let us mention that Figures 5.16 and 5.17 illustrates the ability of numerical simulation to compute quantities for which experimental data are not available.

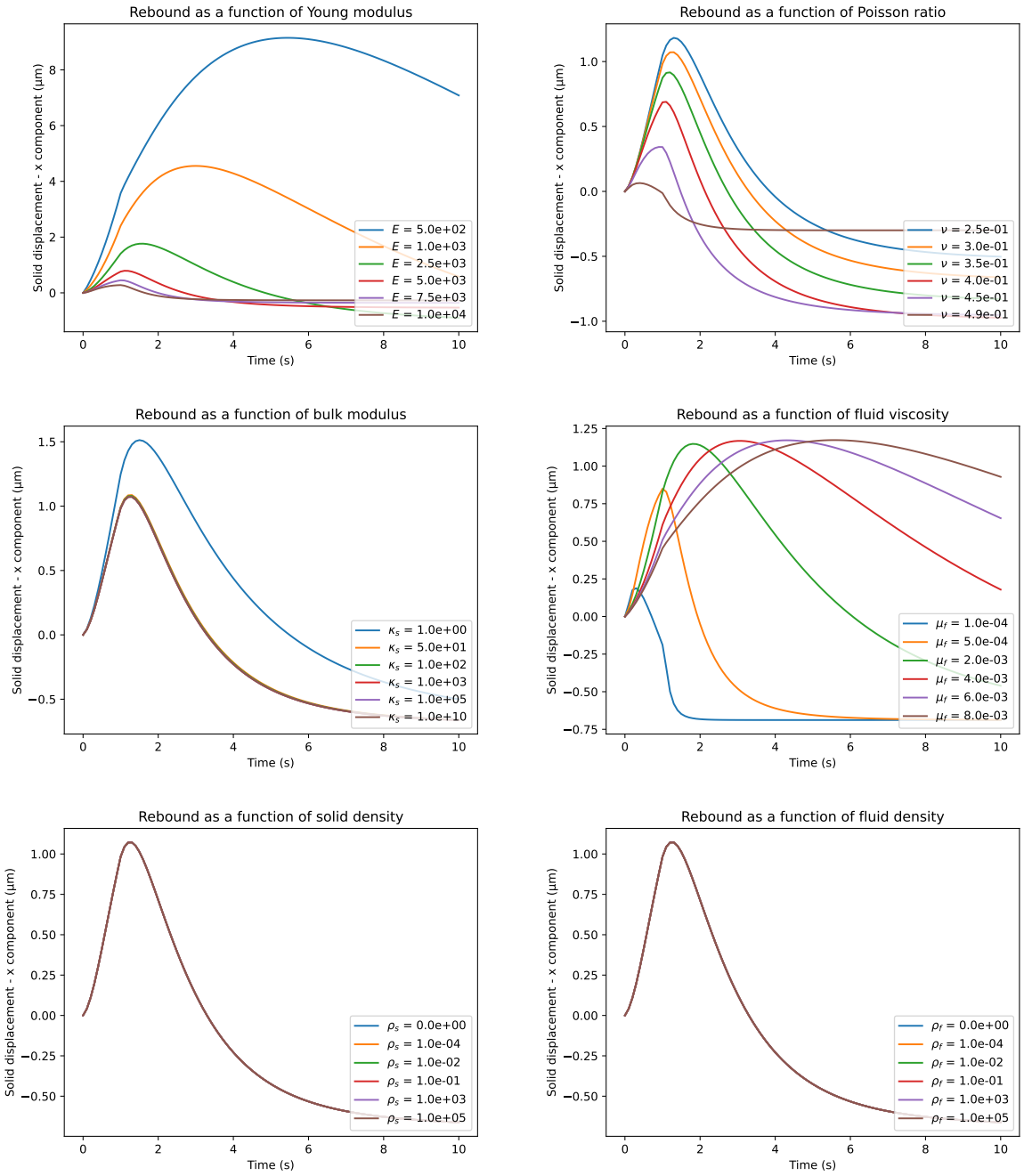


Figure 5.16 – Parametric study of the response curve at point B with respect to the physical parameters.

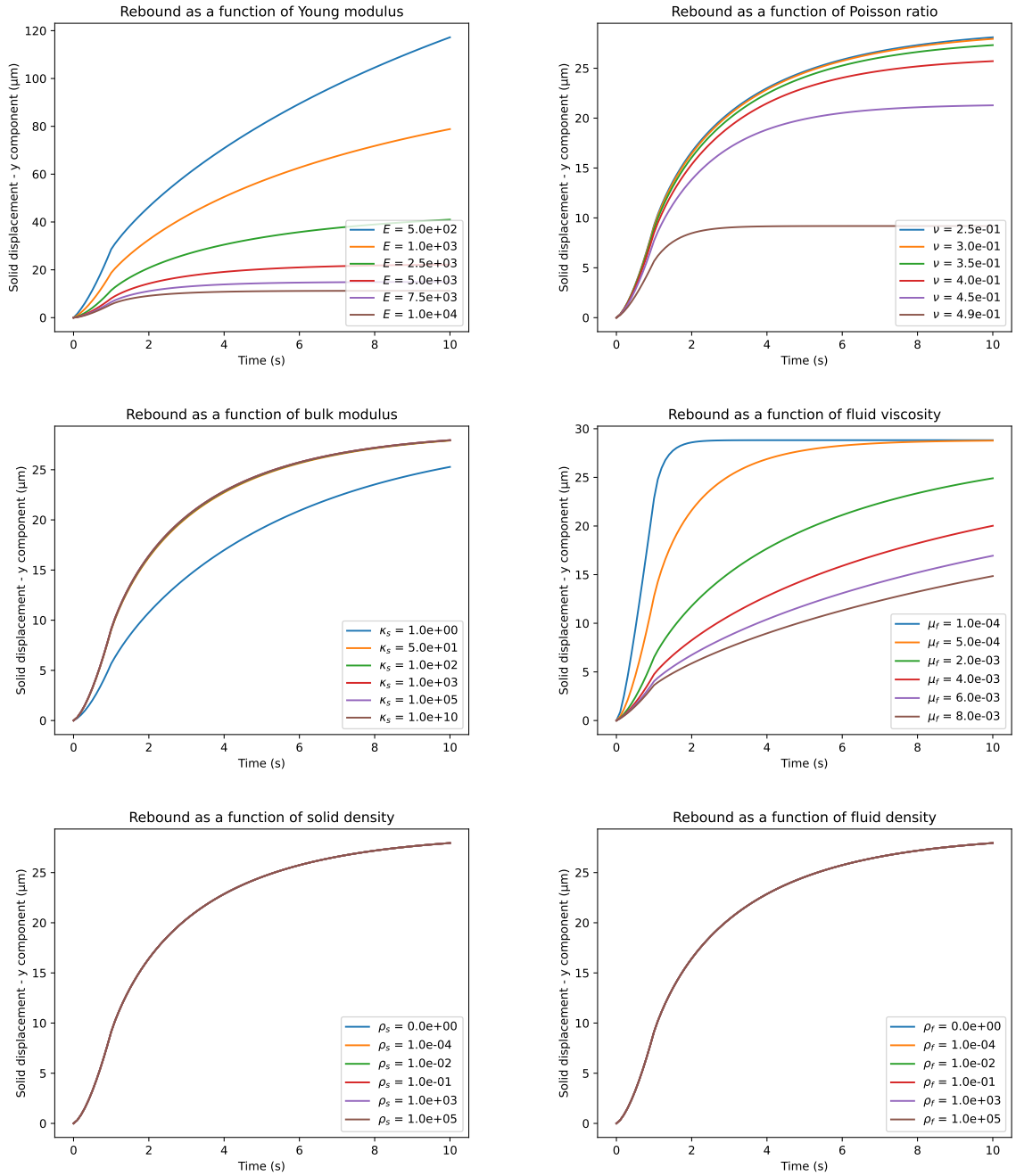


Figure 5.17 – Parametric study of the response curve at point C with respect to the physical parameters.

Coming back to simulations on the channel (at point A), Figure 5.18 illustrates the importance of the pressure ramps parameters on the poromechanical response. Unsurprisingly, Figure 5.18 – left shows that the final value p_0 imposed on the channel is directly correlated with the radial displacement amplitude. In Figure 5.18 – right, we infer that the final value of the radial displacement does not depend on the ramp sharpness, which reproduces the behavior observed in Figure 5.13B. Further work needs to be done to extract the permeability value from this curve, which can be done by fitting exponentially the different peaks of Figure 5.18 – right, see [Dessalles, 2021, Figure 3.9].

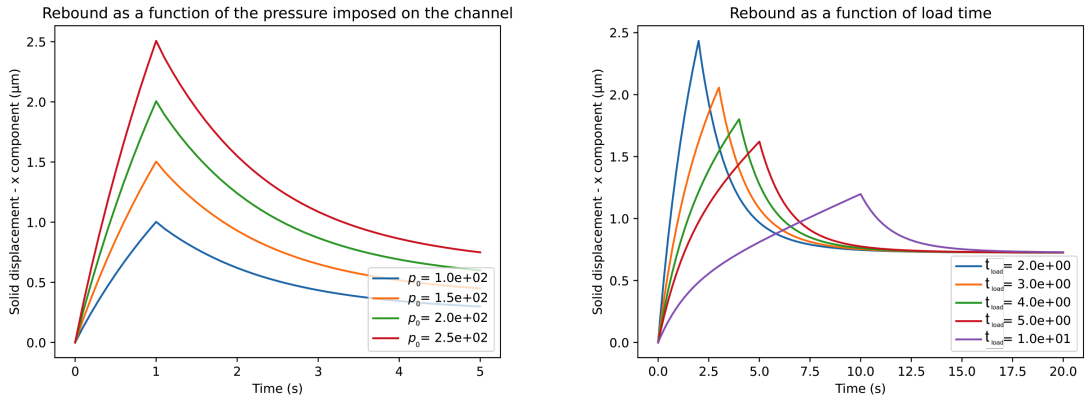


Figure 5.18 – Parametric study of the response curve at point A with respect to the pressure ramp parameters.

Finally, we investigate the influence of the box. From Figure 5.19, we conclude that the domain geometry strongly impacts the solid displacement. Indeed, the narrower the box containing the hydrogel, the less space the channel has to deform without being constrained by the lateral walls.

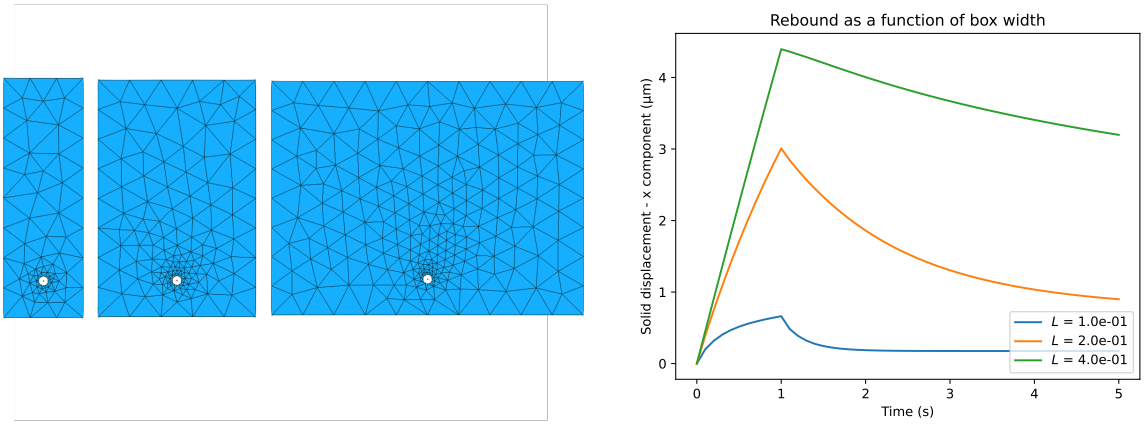


Figure 5.19 – Parametric study of the response curve at point A with respect to the domain width.

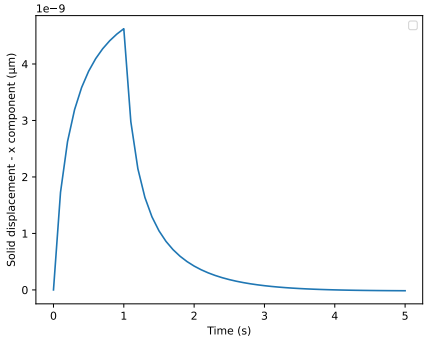


Figure 5.20 – Solid displacement at point A when closing the top surface of the box.

In Figure 5.20, Neumann boundary conditions (5.2) are replaced by homogeneous Dirichlet boundary conditions, which is equivalent to closing the top surface of the box containing the hydrogel. We find that the displacement value is very small and rapidly goes to zero. This corroborates the importance of the free surface on the top, as commented in [Dessalles, 2021].

5.3 Perspectives

A first perspective of this chapter is to take into account the presence of cells on the channel walls. This may be done by using a membrane model or by modeling the cells as an additional porous layer following [Bociu et al., 2021].

Another perspective is to take into account the water flow inside the channel in order to remove the approximation made in (5.4). To do so, we need to couple the poromechanics model (5.1) with a Stokes flow inside the channel, which requires to formulate proper transmission conditions on the interface.

Note that such transmission conditions have already been studied in the literature of fluid-porous structure interaction problems. For the coupling of Darcy and Stokes equations, these conditions were derived by Beavers, Joseph and Saffman [Beavers and Joseph, 1967; Saffman, 1971] and justified *a posteriori* by homogenization [Mikelic and Jäger, 2000]. For the coupling of Biot and Stokes equations, they were among others formulated in [Murad et al., 2001]. However, to the best of our knowledge, coupling the linearized poromechanics model derived in [Burtschell et al., 2019] with a Stokes flow has never been considered.

Let us denote by Ω_c the volume occupied by the water channel, Γ_{in} and Γ_{out} the inlet and outlet channel boundaries, and Γ the interface between the channel and the hydrogel as depicted in Figure 5.21.

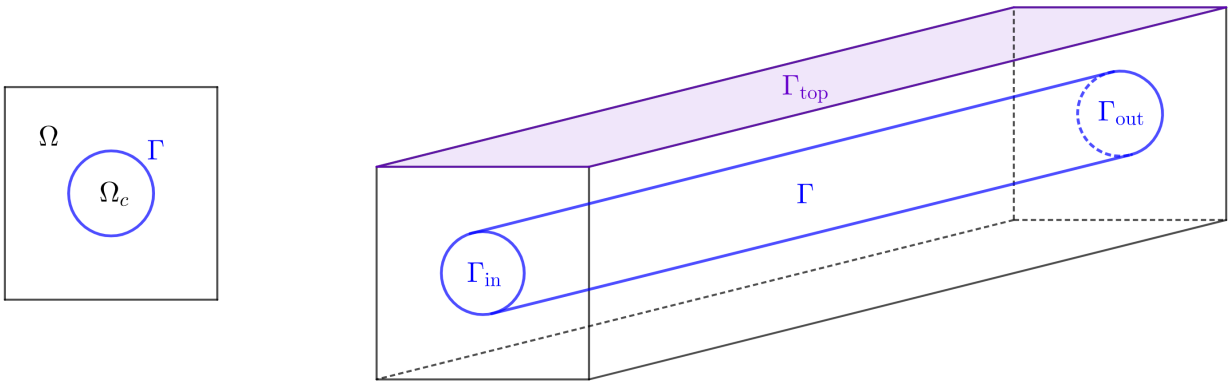


Figure 5.21 – Porous and fluid domains and boundaries in the fluid-porous structure interaction setting.

The water flow in the channel is slow, so that we can neglect convection phenomena. Therefore, we use Stokes equation to model the water channel, namely

$$\begin{cases} \partial_t v_c - \operatorname{div}\left(\sigma_c(v_c)\right) + \nabla p_c = 0, & \text{in } \Omega_c, \\ \operatorname{div} v = 0, & \text{in } \Omega_c, \end{cases} \tag{5.7}$$

where v_c and p_c denote respectively the channel's velocity and pressure, and $\sigma_c(v_c) = 2\mu_f \varepsilon(v_c)$. System (5.7) is complemented with Neumann boundary conditions on the inlet and outlet boundaries.

To couple (5.1) with (5.7), we propose the following transmission conditions on the interface:

$$\begin{aligned} v_c &= (1-\phi)\, v_s + \phi\, v_f, && \text{on } \Gamma, \\ \sigma_s^{\text{tot}} n &= (1-\phi)\, \sigma_c^{\text{tot}} n, && \text{on } \Gamma, \\ \phi\, \sigma_f^{\text{tot}} n &= \phi\, \sigma_c^{\text{tot}} n, && \text{on } \Gamma, \end{aligned}$$

where $\sigma_c^{\text{tot}} = \sigma_c(v_c) - p_c I$, $\sigma_s^{\text{tot}} = \sigma_s(u_s) - (1-\phi)pI$ and $\sigma_f^{\text{tot}} = \sigma_f(v_f) - pI$. One can show that these transmission conditions are compatible with an energy balance, paving the way to the well-posedness of the coupled system.

From a numerical point of view, transmission conditions in fluid-porous structure interaction problems can be imposed using various mixed formulations [Arbogast and Brunson, 2007; Gatica et al., 2011; Li and Yotov, 2022], Lagrange multipliers [Ambartsumyan et al., 2017], discontinuous Galerkin methods [Rivière and Yotov, 2005; Girault and Rivière, 2009], mortar elements [Girault et al., 2014], domain decomposition techniques [Discacciati et al., 2007; Badia et al., 2009] or again Nitsche's method [Bukač et al., 2015a]. Here, we employ a Lagrange multiplier living on the interface, which requires to use the mixed-dimensional branch of FEniCS for the implementation of the coupling [Daversin-Catty et al., 2021].

The resulting numerical results are shown in Figures 5.22 and 5.23 for parameters similar to the ones used in Section 5.2.1 for validating the porous model. In Figure 5.22, we see that the channel deformation profile is quite similar to the one obtained in Figure 5.12 without the fluid-porous structure interaction coupling. Nevertheless, Figure 5.23 confirms that the evolution pressure on the interface cannot be modeled by (5.4).

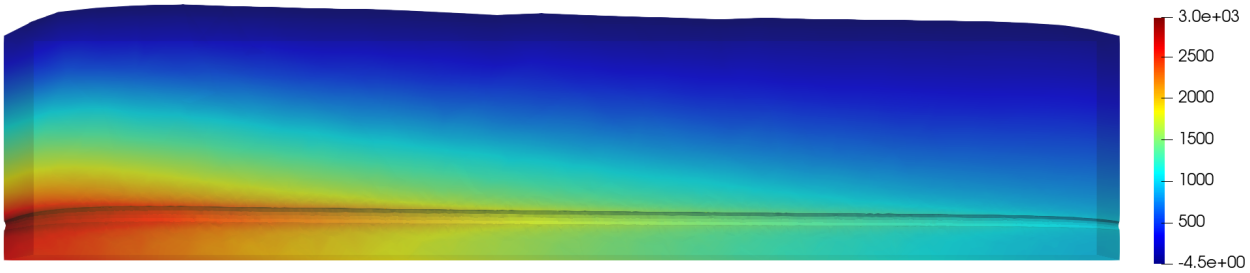


Figure 5.22 – Channel deformation (scale factor 20) with pressure coloration at $T = 5$ simulated with the fluid-porous structure interaction coupling.

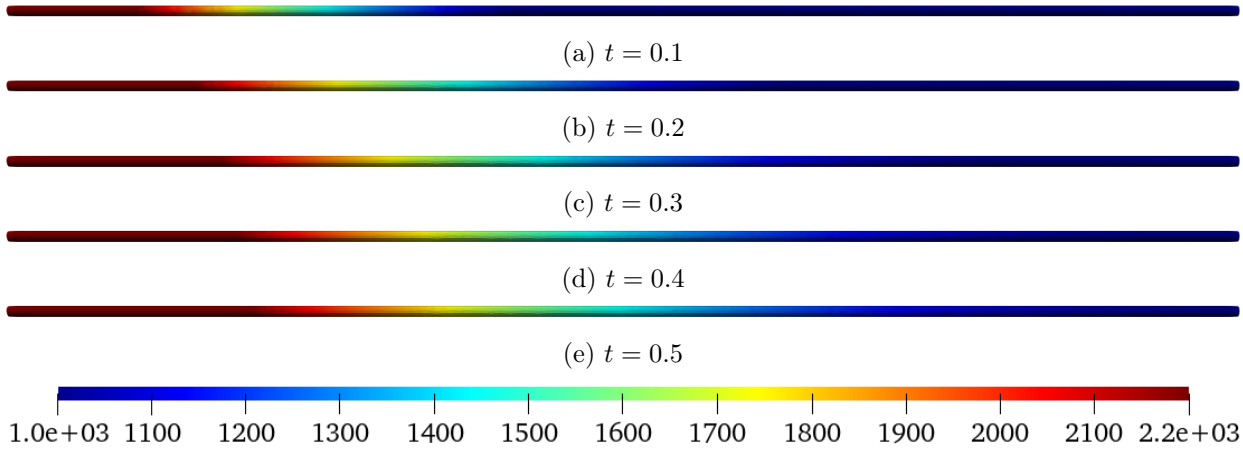


Figure 5.23 – Pressure in the water channel (10^{-1}Pa) over time.

In [Dessalles, 2021], the longitudinal wave propagation in the microfluidic experiment was studied. In particular, it was found in [Dessalles, 2021, Figure 3.17] that the corresponding pulse wave velocity depends on the amplitude of the inlet flow rate, hence contradicting elastic wave propagation theory. In the future, we hope that the fluid-porous structure interaction coupling will enable us to retrieve the pulse wave velocity numerically and to analyze how it is affected by the poroelasticity parameters.

Résumé substantiel

De nombreux tissus biologiques peuvent être modélisés comme des milieux poreux, c'est-à-dire des milieux continus composés d'une structure solide irriguée par un fluide. Dans les tissus biologiques, le fluide peut désigner le sang, les flux d'air dans les poumons ou encore le liquide céphalo-rachidien, fluides qui peuvent tous être considérés comme incompressibles. On appelle perfusion cette interaction entre les solides et les fluides dans les tissus biologiques. De plus, pour de telles applications, le milieu poreux en tant que tel est quasi-incompressible.

Dans cette thèse, nous analysons un modèle d'équations aux dérivées partielles récent qui décrit le mouvement d'un milieu poreux quasi-incompressible ou incompressible. Ce modèle provient de la linéarisation d'un modèle de poromécanique non linéaire adapté au contexte des tissus mous perfusés, mais nous montrons qu'il est également fortement relié aux équations de Biot plus classiques en poroélasticité. Dans ce modèle, les équations du solide et du fluide ont un comportement respectivement hyperbolique et parabolique, et sont couplées par la pression interstitielle associée à la contrainte d'incompressibilité. D'un point de vue théorique comme numérique, les principales difficultés du modèle résident dans ce couplage hyperbolique – parabolique et dans sa structure de type point-selle en régime incompressible.

La première contribution de cette thèse est de démontrer l'existence et l'unicité des solutions fortes ou faibles dans les cas quasi-incompressible et incompressible. La preuve repose sur une combinaison de théorie des semi-groupes et d'estimations d'énergie. Dans le cas non visqueux, la forme bilinéaire sous-jacente n'est pas coercive. Afin de résoudre cette difficulté technique, on fait appel à la notion de `T`-coercivité pour obtenir l'existence et l'unicité des solutions fortes. Par ailleurs, on met en lumière l'influence de la viscosité sur la régularité des solutions faibles.

La notion de `T`-coercivité, développée initialement pour des problèmes non contraints, est ici étendue aux problèmes de type point-selle avec ou sans pénalisation. Au niveau discret, cette méthode permet de concevoir simplement des approximations numériques adaptées car la dérivation de la condition inf-sup discrète uniforme découle en général directement de l'étude du problème continu. Ainsi, la preuve par `T`-coercivité utilisée au niveau continu donne un cadre pour guider la conception d'éléments finis stables dans la limite incompressible et pour effectuer l'analyse numérique du système au cœur de cette thèse. Ce faisant, l'outil de `T`-coercivité permet de gérer de façon compacte à la fois le couplage hyperbolique – parabolique et la contrainte d'incompressibilité.

Deux types de schémas numériques sont considérés pour discrétiser le problème. Tout d'abord, un schéma monolithique, pour lequel on démontre la convergence spatiale et temporelle avec des estimations d'erreur robustes par rapport à l'incompressibilité, la porosité et la perméabilité. Afin d'accélérer le temps de calcul, un schéma à pas fractionnaires est également proposé et analysé. L'intérêt de ce schéma est de découpler les degrés de liberté du solide, du fluide et de la pression à chaque pas de temps. En revanche, sa stabilité est particulièrement sensible à la perméabilité, ce qui nous amène à formuler plusieurs variantes offrant une stabilité inconditionnelle. En outre, des conditions aux limites générales couplant le fluide et le solide sur le bord du domaine sont envisagées et imposées grâce à une méthode de type Robin-Robin. Dans le cas des conditions de Dirichlet, on prouve la convergence espace – temps du schéma, que ce soit dans sa version incrémentale ou non

incrémentale.

Enfin, la pertinence de ce modèle pour les applications biomédicales est illustrée en comparant des simulations de microvaisseaux sur puce à des données expérimentales. Plus précisément, le modèle de poromécanique étudié dans ce travail est utilisé pour simuler la réponse mécanique d'un hydrogel percé par un microvaisseau dans lequel est injecté un fluide. Après calibration des différents paramètres physiques, on retrouve le comportement qualitatif de déformation du gel. À l'avenir, un tel modèle pourrait alors être couplé à une distribution de fluide dans le canal, ce qui permettrait en particulier de prendre en compte l'influence de cellules sur la paroi et fournirait ainsi un outil numérique pour appréhender le rôle joué par l'endothélium dans le fonctionnement mécanique des vaisseaux humains.

Bibliography

Ager, C., Schott, B., Winter, M., and Wall, W. A. (2019). A Nitsche-based cut finite element method for the coupling of incompressible fluid flow with poroelasticity. *Computer Methods in Applied Mechanics and Engineering*, 351:253–280.

Alnæs, M. S., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M. E., and Wells, G. N. (2015). The FEniCS project version 1.5. *Archive of Numerical Software*, 3(100):9–23.

Ambartsumyan, I., Khattatov, E., Yotov, I., and Zunino, P. (2017). A Lagrange multiplier method for a Stokes-Biot fluid-poroelastic structure interaction model. *arXiv:1710.06750 [math]*. arXiv: 1710.06750.

Ambartsumyan, I., Khattatov, E., Yotov, I., and Zunino, P. (2018). A Lagrange multiplier method for a Stokes–Biot fluid–poroelastic structure interaction model. *Numerische Mathematik*, 140(2):513–553.

Ambrosi, D. and Preziosi, L. (2002). On the closure of mass balance models for tumor growth. *Mathematical Models and Methods in Applied Sciences*, 12(05):737–754.

Angot, P. (2018). Well-posed Stokes/Brinkman and Stokes/Darcy coupling revisited with new jump interface conditions. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52(5):1875–1911.

Angot, P., Boyer, F., and Hubert, F. (2009). Asymptotic and numerical modelling of flows in fractured porous media. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(2):239–275.

Arbogast, T. and Brunson, D. S. (2007). A computational method for approximating a Darcy–Stokes system governing a vuggy porous medium. *Computational Geosciences*, 11(3):207–218.

Armstrong, M. H., Buganza Tepole, A., Kuhl, E., Simon, B. R., and Vande Geest, J. P. (2016). A Finite Element Model for Mixed Porohyperelasticity with Transport, Swelling, and Growth. *PLOS ONE*, 11(4):e0152806.

Assous, F., Ciarlet Jr, P., and Labrunie, S. (2018). *Mathematical foundations of computational electromagnetism*. Springer.

Astorino, M., Chouly, F., and Fernández, M. A. (2010). Robin based semi-implicit coupling in fluid-structure interaction: Stability analysis and numerics. *SIAM Journal on Scientific Computing*, 31(6):4041–4065.

Astorino, M. and Grandmont, C. (2010). Convergence analysis of a projection semi-implicit coupling scheme for fluid–structure interaction problems. *Numerische Mathematik*, 116(4):721–767.

Auriault, J. L. (1980). Dynamic behaviour of a porous medium saturated by a newtonian fluid. *International Journal of Engineering Science*, 18(6):775–785.

Auriault, J.-L. (1997). Poroelastic media. In *Homogenization and porous media*, pages 163–182. Springer.

Avalos, G. and Triggiani, R. (2009). Semigroup well-posedness in the energy space of a parabolic-hyperbolic coupled Stokes-Lamé PDE system of fluid-structure interaction. *Discrete & Continuous Dynamical Systems - S*, 2(3):417–447.

Babuška, I. (1973). The finite element method with lagrangian multipliers. *Numerische Mathematik*, 20(3):179–192.

Badia, S. and Codina, R. (2007). Convergence analysis of the FEM approximation of the first order projection method for incompressible flows with and without the inf-sup condition. *Numerische Mathematik*, 107(4):533–557.

Badia, S., Quaini, A., and Quarteroni, A. (2008). Splitting Methods Based on Algebraic Factorization for Fluid-Structure Interaction. *SIAM Journal on Scientific Computing*, 30(4):1778–1805.

Badia, S., Quaini, A., and Quarteroni, A. (2009). Coupling Biot and Navier–Stokes equations for modelling fluid–poroelastic media interaction. *Journal of Computational Physics*, 228(21):7986–8014.

Baffico, L., Grandmont, C., and Maury, B. (2010). Multiscale modeling of the respiratory tract. *Mathematical Models and Methods in Applied Sciences*, 20(01):59–93.

Baker, G. A. (1976). Error estimates for finite element methods for second order hyperbolic equations. *SIAM journal on numerical analysis*, 13(4):564–576.

Barnafi, N., Zunino, P., Dedè, L., and Quarteroni, A. (2021). Mathematical analysis and numerical approximation of a general linearized poro-hyperelastic model. *Computers & Mathematics with Applications*, 91:202–228.

Barnafi Wittwer, N. A., Gregorio, S. D., Dede', L., Zunino, P., Vergara, C., and Quarteroni, A. (2022). A Multiscale Poromechanics Model Integrating Myocardial Perfusion and the Epicardial Coronary Vessels. *SIAM Journal on Applied Mathematics*, 82(4):1167–1193.

Barré, M. and Ciarlet Jr, P. (2022). The T-coercivity approach for mixed problems. Submitted.

Barré, M., Grandmont, C., and Moireau, P. (2023). Analysis of a linearized poromechanics model for incompressible and nearly incompressible materials. *Evolution Equations and Control Theory*, 12(3):846–906.

Barucq, H., Madaune-Tort, M., and Saint-Macary, P. (2004). Theoretical aspects of wave propagation for Biot's consolidation problem. *Monografías del Seminario Mathemático García de Galdeano*, 31:449–458.

Barucq, H., Madaune-Tort, M., and Saint-Macary, P. (2005). On nonlinear Biot's consolidation models. *Nonlinear Analysis*, 63:e985–e995.

Basser, P. J. (1992). Interstitial pressure, volume, and flow during infusion into brain tissue. *Microvascular Research*, 44(2):143–165.

Beavers, G. S. and Joseph, D. D. (1967). Boundary conditions at a naturally permeable wall. *Journal of Fluid Mechanics*, 30(1):197–207. Publisher: Cambridge University Press.

Benešová, B., Kampschulte, M., and Schwarzacher, S. (2023). Variational methods for fluid–structure interaction and porous media. *Nonlinear Analysis: Real World Applications*, 71:103819.

Bensoussan, A., Da Prato, G., Delfour, M. C., and Mitter, S. K. (2007). *Representation and control of infinite dimensional systems*. Springer Science & Business Media.

Berger, L., Bordas, R., Burrowes, K., Grau, V., Tavener, S., and Kay, D. (2016). A poroelastic model coupled to a fluid network with applications in lung modelling. *International Journal for Numerical Methods in Biomedical Engineering*, 32(1).

Bertrand, F., Brodbeck, M., and Ricken, T. (2022). On robust discretization methods for poroelastic problems: Numerical examples and counter-examples. *Examples and Counterexamples*, 2:100087.

Biot, M. A. (1941). General theory of three-dimensional consolidation. *Journal of applied physics*, 12(2):155–164.

Biot, M. A. (1955). Theory of elasticity and consolidation for a porous anisotropic solid. *Journal of applied physics*, 26(2):182–185.

Biot, M. A. (1956a). Theory of propagation of elastic Waves in a fluid-saturated porous solid. I. Low-frequency range. *The Journal of the Acoustical Society of America*, 28(2):168–178.

Biot, M. A. (1956b). Theory of propagation of elastic waves in a fluid-saturated porous solid. ii. higher frequency range. *The Journal of the acoustical Society of america*, 28(2):179–191.

Biot, M. A. and Temple, G. (1972). Theory of finite deformations of porous solids. *Indiana University Mathematics Journal*, 21(7):597–620.

Bociu, L., Canic, S., Muha, B., and Webster, J. T. (2021). Multilayered poroelasticity interacting with Stokes flow. *SIAM Journal on Mathematical Analysis*, 53(6):6243–6279.

Bociu, L., Guidoboni, G., Sacco, R., and Verri, M. (2019). On the role of compressibility in poroviscoelastic models. *Mathematical Biosciences and Engineering*, 16(5):6167–6028.

Bociu, L., Guidoboni, G., Sacco, R., and Webster, J. T. (2016). Analysis of nonlinear poro-elastic and poro-visco-elastic models. *Archive for Rational Mechanics and Analysis*, 222(3):1445–1519.

Bociu, L., Muha, B., and Webster, J. T. (2022). Weak solutions in nonlinear poroelasticity with incompressible constituents. *Nonlinear Analysis: Real World Applications*, 67:103563.

Bociu, L. and Webster, J. T. (2021). Nonlinear quasi-static poroelasticity. *Journal of Differential Equations*, 296:242–278.

Boffi, D., Brezzi, F., Fortin, M., et al. (2013). *Mixed finite element methods and applications*, volume 44. Springer.

Bogovskii, M. E. (1979). Solution of the first boundary value problem for the equation of continuity of an incompressible medium. In *Doklady Akademii Nauk*, volume 248, pages 1037–1040. Russian Academy of Sciences.

Bonaldi, F., Brenner, K., Droniou, J., and Masson, R. (2021). Gradient discretization of two-phase flows coupled with mechanical deformation in fractured porous media. *Computers & Mathematics with Applications*, 98:40–68.

Bonnet-Ben Dhia, A.-S., Carvalho, C., and Ciarlet, Jr, P. (2018). Mesh requirements for the finite element approximation of problems with sign-changing coefficients. *Numer. Math.*, 138:801–838.

Bonnet-Ben Dhia, A.-S., Chesnel, L., and Ciarlet, Jr, P. (2012). T-coercivity for scalar interface problems between dielectrics and metamaterials. *Math. Mod. Num. Anal.*, 46:1363–1387.

Bonnet-Ben Dhia, A.-S., Chesnel, L., and Ciarlet Jr, P. (2014). T-coercivity for the maxwell problem with sign-changing coefficients. *Communications in Partial Differential Equations*, 39(6):1007–1031.

Bonnet-Ben Dhia, A.-S., Chesnel, L., and Ciarlet, Jr, P. (2014a). T-coercivity for the Maxwell problem with sign-changing coefficients. *Communications in Partial Differential Equations*, 39:1007–1031.

Bonnet-Ben Dhia, A.-S., Chesnel, L., and Ciarlet, Jr, P. (2014b). Two-dimensional Maxwell's equations with sign-changing coefficients. *Appl. Numer. Math.*, 79:29–41.

Bonnet-Ben Dhia, A.-S., Chesnel, L., and Claeys, X. (2013). Radiation condition for a non-smooth interface between a dielectric and a metamaterial. *Mathematical Models and Methods in Applied Sciences*, 23(09):1629–1662.

Bonnet-Ben Dhia, A.-S., Ciarlet Jr, P., and Zwölf, C. M. (2010a). Time harmonic wave diffraction problems in materials with sign-shifting coefficients. *Journal of Computational and Applied Mathematics*, 234(6):1912–1919.

Bonnet-Ben Dhia, A.-S., Ciarlet Jr, P., and Zwölf, C. M. (2010b). Time harmonic wave diffraction problems in materials with sign-shifting coefficients. *Journal of Computational and Applied Mathematics*, 234(6):1912–1919.

Boon, W. M., Hornkjøl, M., Kuchta, M., Mardal, K.-A., and Ruiz-Baier, R. (2022). Parameter-robust methods for the biot–stokes interfacial coupling without lagrange multipliers. *Journal of Computational Physics*, 467:111464.

Both, J. W., Barnafi, N. A., Radu, F. A., Zunino, P., and Quarteroni, A. (2022). Iterative splitting schemes for a soft material poromechanics model. *Computer Methods in Applied Mechanics and Engineering*, page 29.

Both, J. W., Borregales, M., Nordbotten, J. M., Kumar, K., and Radu, F. A. (2017). Robust fixed stress splitting for Biot's equations in heterogeneous media. *Applied Mathematics Letters*, 68:101–108.

Both, J. W., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2019a). Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media. *Computers & Mathematics with Applications*, 77(6):1479–1502.

Both, J. W., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2019b). The gradient flow structures of thermo-poro-visco-elastic processes in porous media.

Both, J. W., Pop, I. S., and Yotov, I. (2021). Global existence of weak solutions to unsaturated poroelasticity. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(6):2849–2897.

Bourgeois, E. (1997). Mécanique des milieux poreux en transformation finie: formulation des problèmes et méthodes de résolution. *Thèse de doctorat, École Nationale des Ponts et Chaussées*.

Bowen, R. M. (1980). Incompressible porous media models by use of the theory of mixtures. *International Journal of Engineering Science*, 18(9):1129–1148.

Boyer, F. and Fabrie, P. (2012). *Mathematical Tools for the Study of the Incompressible Navier-Stokes Equations and Related Models*, volume 183. Springer Science & Business Media.

Brezzi, F. (1974). On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers. *Publications mathématiques et informatique de Rennes*, 8(S4):1–26.

Brun, M. K., Ahmed, E., Nordbotten, J. M., and Radu, F. A. (2019). Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport. *Journal of Mathematical Analysis and Applications*, 471(1):239–266.

Buffa, A. (2005). Remarks on the discretization of some noncoercive operator with applications to heterogeneous Maxwell equations. *SIAM J. Numer. Anal.*, 43:1–18.

Buffa, A. and Christiansen, S. H. (2003). The electric field integral equation on Lipschitz screens: definitions and numerical approximation. *Numerische Mathematik*, 94(2):229–267.

Buffa, A., Costabel, M., and Schwab, C. (2002). Boundary element methods for Maxwell's equations on non-smooth domains. *Numerische Mathematik*, 92(4):679–710.

Bukač, M., Yotov, I., Zakerzadeh, R., and Zunino, P. (2015a). Partitioning strategies for the interaction of a fluid with a poroelastic material based on a Nitsche's coupling approach. *Comput. Methods Appl. Mech. Engrg.*, page 33.

Bukač, M., Yotov, I., and Zunino, P. (2015b). An operator splitting approach for the interaction between a fluid and a multilayered poroelastic structure. *Numerical Methods for Partial Differential Equations*, 31(4):1054–1100.

Bukač, M., Yotov, I., and Zunino, P. (2016). Dimensional model reduction for flow through fractures in poroelastic media. *ESAIM: Mathematical Modelling and Numerical Analysis*.

Bukač, M., Čanić, S., Glowinski, R., Muha, B., and Quaini, A. (2014). A modular, operator-splitting scheme for fluid–structure interaction problems with thick structures. *International Journal for Numerical Methods in Fluids*, 74(8):577–604. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/fld.3863.

Bunoiu, R., Chesnel, L., Ramdani, K., and Rihani, M. (2020). Homogenization of maxwell's equations and related scalar problems with sign-changing coefficients. In *Annales de la Faculté des Sciences de Toulouse. Mathématiques*.

Bunoiu, R. and Ramdani, K. (2016). Homogenization of materials with sign changing coefficients. *Communications in Mathematical Sciences*, 14(4):1137–1154.

Bunoiu, R., Ramdani, K., and Timofte, C. (2021). T-coercivity for the asymptotic analysis of scalar problems with sign-changing coefficients in thin periodic domains. *Electronic Journal of Differential Equations*, pages 1–22.

Burman, E., Claus, S., Hansbo, P., Larson, M. G., and Massing, A. (2015). Cutfem: discretizing geometry and partial differential equations. *International Journal for Numerical Methods in Engineering*, 104(7):472–501.

Burman, E., Durst, R., Fernández, M., and Guzmán, J. (2021). Loosely coupled, non-iterative time-splitting scheme based on Robin-Robin coupling: Unified analysis for parabolic/parabolic and parabolic/hyperbolic problems.

Burman, E., Durst, R., Fernández, M. A., and Guzmán, J. (2022a). Fully discrete loosely coupled Robin-Robin scheme for incompressible fluid–structure interaction: Stability and error analysis. *Numerische Mathematik*, 151(4):807–840.

Burman, E., Durst, R., and Guzmán, J. (2022b). Stability and error analysis of a splitting method using robin–robin coupling applied to a fluid–structure interaction problem. *Numerical Methods for Partial Differential Equations*, 38(5):1396–1406.

Burman, E. and Fernández, M. A. (2009). Stabilization of explicit coupling in fluid–structure interaction involving fluid incompressibility. *Computer Methods in Applied Mechanics and Engineering*, 198(5-8):766–784.

Burman, E. and Fernández, M. A. (2014). Explicit strategies for incompressible fluid-structure interaction problems: Nitsche type mortaring versus Robin–Robin coupling. *International Journal for Numerical Methods in Engineering*, 97(10):739–758. Publisher: Wiley Online Library.

Burq, N. and Gérard, P. (2002). *Contrôle optimal des equations aux derivées partielles*. École Polytechnique, Département de mathématiques.

Burtschell, B. (2016). *Modélisation mécanique et méthodes numériques pour la poromécanique-Application à la perfusion du myocarde.(Mechanical modeling and numerical methods for poromechanics-Application to myocardium perfusion)*. PhD thesis, École polytechnique.

Burtschell, B., Chapelle, D., and Moireau, P. (2017). Effective and energy-preserving time discretization for a general nonlinear poromechanical formulation. *Computers & Structures*, 182:313–324.

Burtschell, B., Moireau, P., and Chapelle, D. (2019). Numerical analysis for an energy-stable total discretization of a poromechanics model with inf-sup stability. *Acta Mathematicae Applicatae Sinica, English Series*, 35(1):28–53.

Caiazzo, A., Fernández, M. A., Gerbeau, J.-F., and Martin, V. (2011). Projection Schemes for Fluid Flows through a Porous Interface. *SIAM Journal on Scientific Computing*, 33(2):541–564.

Calo, V., Brasher, N., Bazilevs, Y., and Hughes, T. (2008). Multiphysics model for blood flow and drug transport with application to patient-specific coronary artery flow. *Computational Mechanics*, 43(1):161–177.

Čanić, S., Wang, Y., and Bukač, M. (2021). A next-generation mathematical model for drug-eluting stents. *SIAM Journal on Applied Mathematics*, 81(4):1503–1529.

Cao, Y., Chen, S., and Meir, A. (2013). Analysis and numerical approximations of equations of nonlinear poroelasticity. *Discrete & Continuous Dynamical Systems-B*, 18(5):1253.

Causin, P., Gerbeau, J., and Nobile, F. (2005). Added-mass effect in the design of partitioned algorithms for fluid–structure problems. *Computer Methods in Applied Mechanics and Engineering*, 194(42-44):4506–4527.

Causin, P., Guidoboni, G., Harris, A., Prada, D., Sacco, R., and Terragni, S. (2014). A poroelastic model for the perfusion of the lamina cribrosa in the optic nerve head. *Mathematical Biosciences*, 257:33–41.

Cesmelioglu, A. (2017). Analysis of the coupled Navier–Stokes/Biot problem. *Journal of Mathematical Analysis and Applications*, 456(2):970–991.

Chabiniok, R., Burtschell, B., Chapelle, D., and Moireau, P. (2022). Dimensional reduction of a poromechanical cardiac model for myocardial perfusion studies. *Applications in Engineering Science*, 12:100121.

Chabiniok, R., Wang, V. Y., Hadjicharalambous, M., Asner, L., Lee, J., Sermesant, M., Kuhl, E., Young, A. A., Moireau, P., Nash, M. P., et al. (2016). Multiphysics and multiscale modelling, data–model fusion and integration of organ physiology in the clinic: ventricular cardiac mechanics. *Interface focus*, 6(2):20150083.

Chapelle, D., Gerbeau, J.-F., Sainte-Marie, J., and Vignon-Clementel, I. (2010). A poroelastic model valid in large strains with applications to perfusion in cardiac modeling. *Computational Mechanics*, 46(1):91–101.

Chapelle, D. and Moireau, P. (2014). General coupling of porous flows and hyperelastic formulations—from thermodynamics principles to energy balance and compatible time schemes. *European Journal of Mechanics-B/Fluids*, 46:82–96.

Chatzizisis, Y. S., Coskun, A. U., Jonas, M., Edelman, E. R., Feldman, C. L., and Stone, P. H. (2007). Role of endothelial shear stress in the natural history of coronary atherosclerosis and vascular remodeling: molecular, cellular, and vascular behavior. *Journal of the American College of Cardiology*, 49(25):2379–2393.

Chen, Y., Luo, Y., and Feng, M. (2013). Analysis of a discontinuous Galerkin method for the Biot's consolidation problem. *Applied Mathematics and Computation*, 219(17):9043–9056.

Chesnel, L. (2016). Bilaplacian problems with a sign-changing coefficient. *Mathematical Methods in the Applied Sciences*, 39(17):4964–4979.

Chesnel, L. and Ciarlet, P. (2013). T-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients. *Numerische Mathematik*, 124(1):1–29.

Chorin, A. J. (1969). On the convergence of discrete approximations to the navier-stokes equations. *Mathematics of computation*, 23(106):341–353.

Chou, D., Vardakis, J. C., Guo, L., Tully, B. J., and Ventikos, Y. (2016). A fully dynamic multi-compartmental poroelastic system: Application to aqueductal stenosis. *Journal of Biomechanics*, 49(11):2306–2312.

Ciarlet, P. G. (1988). *Mathematical Elasticity: Volume I: three-dimensional elasticity*. North-Holland.

Ciarlet Jr, P. (2012). T-coercivity: Application to the discretization of helmholtz-like problems. *Computers & Mathematics with Applications*, 64(1):22–34.

Ciarlet Jr, P. (2020). Mathematical and numerical analyses for the div-curl and div-curlcurl problems with a sign-changing coefficient. Technical Report HAL.

Ciarlet Jr, P. (2021). Lecture notes on maxwell's equations and their approximation (in french). Master's degree Analysis, Modelling and Simulation from Paris-Saclay University and Institut Polytechnique de Paris.

Ciarlet Jr, P. (2022). On the approximation of electromagnetic fields by edge finite elements – Part 4: analysis of the model with one sign-changing coefficient. *Numer. Math.*, 152:223–257.

Ciarlet Jr, P., Jamelot, E., and Kpadonou, F. D. (2017). Domain decomposition methods for the diffusion equation with low-regularity solution. *Computers Math. Applic.*, 74:2369–2384.

Cioncolini, A. and Boffi, D. (2019). The MINI mixed finite element for the Stokes problem: An experimental investigation. *Computers & Mathematics with Applications*, 77(9):2432–2446.

Clarelli, F., Di Russo, C., Natalini, R., and Ribot, M. (2013). A fluid dynamics model of the growth of phototrophic biofilms. *Journal of Mathematical Biology*, 66(7):1387–1408.

Coussy, O. (2004). *Poromechanics*. John Wiley & Sons.

Crank, J. and Nicolson, P. (1947). A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. In *Mathematical proceedings of the Cambridge philosophical society*, volume 43, pages 50–67. Cambridge University Press.

Cunningham, K. S. and Gotlieb, A. I. (2005). The role of shear stress in the pathogenesis of atherosclerosis. *Laboratory investigation*, 85(1):9–23.

Dafermos, C. M. (1968). On the existence and the asymptotic stability of solutions to the equations of linear thermoelasticity. *Archive For Rational Mechanics And Analysis*, 29(4):241–271.

D'Angelo, C. and Zunino, P. (2011). Robust numerical approximation of coupled Stokes' and Darcy's flows applied to vascular hemodynamics and biochemical transport. *ESAIM: Mathematical Modelling and Numerical Analysis*, 45(3):447–476.

Dautray, R. and Lions, J.-L. (1992). *Mathematical analysis and numerical methods for science and technology: Evolution problems I*, volume 6. Springer Science & Business Media.

Daversin-Catty, C., Richardson, C. N., Ellingsrud, A. J., and Rognes, M. E. (2021). Abstractions and automated algorithms for mixed domain finite element methods. *ACM Transactions on Mathematical Software (TOMS)*, 47(4):1–36.

De Boer, R. (2005). *Trends in continuum mechanics of porous media*, volume 18. Springer Science & Business Media.

De Montgolfier, O., Pinçon, A., Pouliot, P., Gillis, M.-A., Bishop, J., Sled, J. G., Villeneuve, L., Ferland, G., Lévy, B. I., Lesage, F., et al. (2019). High systolic blood pressure induces cerebral microvascular endothelial dysfunction, neurovascular unit damage, and cognitive decline in mice. *Hypertension*, 73(1):217–228.

Demeulenaere, O., Sandoval, Z., Mateo, P., Dizeux, A., Villemain, O., Gallet, R., Ghaleh, B., Deffieux, T., Deméné, C., Tanter, M., et al. (2022). Coronary flow assessment using 3-dimensional ultrafast ultrasound localization microscopy. *Cardiovascular Imaging*, 15(7):1193–1208.

Deparis, S., Fernández, M. A., and Formaggia, L. (2003). Acceleration of a fixed point algorithm for fluid-structure interaction using transpiration conditions. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(4):601–616.

Dessalles, C. (2021). *Forces in a microvessel-on-chip: system development, poroelasticity mechanics and cellular response*. PhD thesis, Institut polytechnique de Paris.

Dessalles, C. A., Ramón-Lozano, C., Babataheri, A., and Barakat, A. I. (2021). Luminal flow actuation generates coupled shear and strain in a microvessel-on-chip. *Biofabrication*, 14(1):015003.

Deville, M., Natalini, R., and Poignard, C. (2018). A continuum mechanics model of enzyme-based tissue degradation in cancer therapies. *Bulletin of mathematical biology*, 80(12):3184–3226.

Di Gregorio, S., Fedele, M., Pontone, G., Corno, A. F., Zunino, P., Vergara, C., and Quarteroni, A. (2021). A computational model applied to myocardial perfusion in the human heart: From large coronaries to microvasculature. *Journal of Computational Physics*, 424:109836.

Discacciati, M., Quarteroni, A., and Valli, A. (2007). Robin–Robin Domain Decomposition Methods for the Stokes–Darcy Coupling. *SIAM Journal on Numerical Analysis*, 45(3):1246–1268.

Dormieux, L., Kondo, D., and Ulm, F.-J. (2006). *Microporomechanics*. John Wiley & Sons.

Duvaut, G. and Lions, J. L. (1972). *Les inéquations en mécanique et en physique*. Dunod.

Eichel, H., Tobiska, L., and Xie, H. (2011). Supercloseness and superconvergence of stabilized low-order finite element discretizations of the stokes problem. *Mathematics of computation*, 80(274):697–722.

Ern, A. and Guermond, J.-L. (2021a). *Finite Elements II: Galerkin approximation, elliptic and mixed PDEs*, volume 73. Springer Nature.

Ern, A. and Guermond, J.-L. (2021b). *Finite Elements III: first-order and time-dependent PDEs*, volume 74. Springer Nature.

Ezziani, A. (2005). *Modélisation mathématique et numérique de la propagation d'ondes dans les milieux viscoélastiques et poroélastiques*. PhD thesis, ENSTA ParisTech.

Fernández, M. A., Gerbeau, J.-F., and Grandmont, C. (2007). A projection semi-implicit scheme for the coupling of an elastic structure with an incompressible fluid. *International Journal for Numerical Methods in Engineering*, 69(4):794–821.

Ferronato, M., Castelletto, N., and Gambolati, G. (2010). A fully coupled 3-D mixed finite element model of Biot consolidation. *Journal of Computational Physics*, 229(12):4813–4830.

Fichera, G. (1974). Uniqueness, existence and estimate of the solution in the dynamical problem of thermo-diffusion in an elastic solid. *Archives of Mechanics*, 26(5):903–920.

Fortin, M. (1977). An analysis of the convergence of mixed finite element methods. *RAIRO. Analyse numérique*, 11(4):341–354.

Frantz, C., Stewart, K. M., and Weaver, V. M. (2010). The extracellular matrix at a glance. *Journal of cell science*, 123(24):4195–4200.

Gajo, A. and Denzer, R. (2011). Finite element modelling of saturated porous media at finite strains under dynamic conditions with compressible constituents. *International journal for numerical methods in engineering*, 85(13):1705–1736. Publisher: Wiley Online Library.

Gaspar, F. J., Lisbona, F. J., and Vabishchevich, P. N. (2003). A finite difference analysis of Biot's consolidation model. *Applied Numerical Mathematics*, 44(4):487–506.

Gatica, G., Oyarzúa, R., and Sayas, F.-J. (2011). Analysis of fully-mixed finite element methods for the Stokes-Darcy coupled problem. *Mathematics of Computation*, 80(276):1911–1948.

Genet, M., Patte, C., Fetita, C., Brillet, P.-Y., and Chapelle, D. (2020). Personalized pulmonary poromechanics. *Computer Methods in Biomechanics and Biomedical Engineering*, 23(sup1):S119–S120.

Gerbeau, J.-F. and Vidrascu, M. (2003). A Quasi-Newton Algorithm Based on a Reduced Model for Fluid-Structure Interaction Problems in Blood Flows. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(4):631–647.

Geuzaine, C. and Remacle, J.-F. (2009). Gmsh: A 3-d finite element mesh generator with built-in pre-and post-processing facilities. *International journal for numerical methods in engineering*, 79(11):1309–1331.

Gil, L. (2020). *A general continuum theory of finite strain chemoporomechanics with application to subcutaneous injections*. PhD thesis, Institut polytechnique de Paris.

Gil, L., Jabbour, M., and Triantafyllidis, N. (2022). The role of the relative fluid velocity in an objective continuum theory of finite strain poroelasticity. *Journal of Elasticity*, pages 1–46.

Girault, V. and Raviart, P.-A. (1986). *Finite element methods for Navier-Stokes equations: theory and algorithms*, volume 5. Springer Science & Business Media.

Girault, V. and Rivière, B. (2009). DG Approximation of Coupled Navier–Stokes and Darcy Equations by Beaver–Joseph–Saffman Interface Condition. *SIAM Journal on Numerical Analysis*, 47(3):2052–2089.

Girault, V., Vassilev, D., and Yotov, I. (2014). Mortar multiscale finite element methods for Stokes–Darcy flows. *Numerische Mathematik*, 127(1):93–165.

Girault, V., Wheeler, M. F., Almani, T., and Dana, S. (2019). *A Priori* error estimates for a discretized poro-elastic–elastic system solved by a fixed-stress algorithm. *Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles*, 74:24.

Giret, L. (2018). *Numerical analysis of a non-conforming Domain Decomposition for the multigroup SPN equations*. PhD thesis, Paris-Saclay University.

Glowinski, R. (2003). Finite element methods for incompressible viscous flow. In *Handbook of Numerical Analysis*, volume 9 of *Numerical Methods for Fluids (Part 3)*, pages 3–1176. Elsevier.

Goda, K. (1979). A multistep technique with implicit difference schemes for calculating two- or three-dimensional cavity flows. *Journal of Computational Physics*, 30(1):76–95.

Gonzalez, O. (2000). Exact energy and momentum conserving algorithms for general models in nonlinear elasticity. *Computer Methods in Applied Mechanics and Engineering*, 190(13-14):1763–1783.

Guermond, J., Minev, P., and Shen, J. (2006). An overview of projection methods for incompressible flows. *Computer Methods in Applied Mechanics and Engineering*, 195(44-47):6011–6045.

Guermond, J.-L. (1996). Some implementations of projection methods for Navier-Stokes equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 30(5):637–667.

Guermond, J. L., Minev, P., and Shen, J. (2005). Error Analysis of Pressure-Correction Schemes for the Time-Dependent Stokes Equations with Open Boundary Conditions. *SIAM Journal on Numerical Analysis*, 43(1):239–258.

Guermond, J.-L. and Quartapelle, L. (1998). On stability and convergence of projection methods based on pressure Poisson equation. *International Journal for Numerical Methods in Fluids*, 26(9):1039–1053.

Guidoboni, G., Glowinski, R., Cavallini, N., and Canic, S. (2009). Stable loosely-coupled-type algorithm for fluid–structure interaction in blood flow. *Journal of Computational Physics*, 228(18):6916–6937.

Guo, L., Vardakis, J. C., Lassila, T., Mitolo, M., Ravikumar, N., Chou, D., Lange, M., Sarrami-Foroushani, A., Tully, B. J., Taylor, Z. A., Varma, S., Venneri, A., Frangi, A. F., and Ventikos, Y. (2018). Subject-specific multi-poroelastic model for exploring the risk factors associated with the early stages of Alzheimer's disease. *Interface Focus*, 8(1):20170019.

Haga, J. B., Osnes, H., and Langtangen, H. P. (2012). On the causes of pressure oscillations in low-permeable and low-compressible porous media. *International Journal for Numerical and Analytical Methods in Geomechanics*, 36(12):1507–1522.

Hahn, C. and Schwartz, M. A. (2009). Mechanotransduction in vascular physiology and atherogenesis. *Nature reviews Molecular cell biology*, 10(1):53–62.

Halla, M. (2021). Galerkin approximation of holomorphic eigenvalue problems: weak t-coercivity and t-compatibility. *Numerische Mathematik*, 148(2):387–407.

Hansbo, P., Hermansson, J., and Svedberg, T. (2004). Nitsche's method combined with space–time finite elements for ALE fluid–structure interaction problems. *Computer Methods in Applied Mechanics and Engineering*, 193(39-41):4195–4206.

Harlow, F. H. and Welch, J. E. (1965). Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *The physics of fluids*, 8(12):2182–2189.

Hauret, P. and Le Tallec, P. (2006). Energy-controlling time integration methods for nonlinear elastodynamics and low-velocity impact. *Computer methods in applied mechanics and engineering*, 195(37-40):4890–4916.

Heywood, J. G. and Rannacher, R. (1990). Finite-element approximation of the nonstationary navier–stokes problem. part iv: Error analysis for second-order time discretization. *SIAM Journal on Numerical Analysis*, 27(2):353–384.

Hiptmair, R. (2002). Finite elements in computational electromagnetics. *Acta Numerica*, pages 237–339.

Hong, Q., Kraus, J., Lymbery, M., and Philo, F. (2023). A new practical framework for the stability analysis of perturbed saddle-point problems and applications. *Math. Comp.*, 92:607–634.

Hong, Q., Kraus, J., Lymbery, M., and Wheeler, M. F. (2020). Parameter-Robust Convergence Analysis of Fixed-Stress Split Iterative Method for Multiple-Permeability Poroelasticity Systems. *Multiscale Modeling & Simulation*, 18(2):916–941.

Hornung, U., Kadanoff, L., Marsden, J. E., Sirovich, L., Wiggins, S., and John, F., editors (1997). *Homogenization and Porous Media*, volume 6 of *Interdisciplinary Applied Mathematics*. Springer New York, New York, NY.

Hu, X., Rodrigo, C., Gaspar, F. J., and Zikatanov, L. T. (2017). A nonconforming finite element method for the Biot's consolidation model in poroelasticity. *Journal of Computational and Applied Mathematics*, 310:143–154.

Huang, M., Wu, S., and Zienkiewicz, O. C. (2001). Incompressible or nearly incompressible soil dynamic behaviour—a new staggered algorithm to circumvent restrictions of mixed formulation. *Soil Dynamics and Earthquake Engineering*, 21(2):169–179.

Huh, D., Matthews, B. D., Mammoto, A., Montoya-Zavala, M., Yuan Hsin, H., and Ingber, D. E. (2010). Reconstituting Organ-Level Lung Functions on a Chip. *Science (New York, N.Y.)*, 328(5986):1662–1668.

Huxley, V., Curry, F., and Adamson, R. (1987). Quantitative fluorescence microscopy on single capillaries: alpha-lactalbumin transport. *American Journal of Physiology-Heart and Circulatory Physiology*, 252(1):H188–H197.

Huyghe, J. M., Arts, T., van Campen, D. H., and Reneman, R. S. (1992). Porous medium finite element model of the beating left ventricle. *American Journal of Physiology-Heart and Circulatory Physiology*, 262(4):H1256–H1267.

Iliff, J. J., Wang, M., Zeppenfeld, D. M., Venkataraman, A., Plog, B. A., Liao, Y., Deane, R., and Nedergaard, M. (2013). Cerebral arterial pulsation drives paravascular csf–interstitial fluid exchange in the murine brain. *Journal of Neuroscience*, 33(46):18190–18199.

Jamelot, E. (2023). Improved stability estimates for solving stokes problem with fortin-soulie finite elements. Technical Report HAL.

Jamelot, E. and Ciarlet Jr, P. (2013). Fast non-overlapping schwarz domain decomposition methods for solving the neutron diffusion equation. *J. Comput. Phys.*, 241:445–463.

J. Polacheck, W., Li, R., M. Uzel, S. G., and D. Kamm, R. (2013). Microfluidic platforms for mechanobiology. *Lab on a Chip*, 13(12):2252–2267. Publisher: Royal Society of Chemistry.

Kedarasetti, R. T., Drew, P. J., and Costanzo, F. (2022). Arterial vasodilation drives convective fluid flow in the brain: a poroelastic model. *Fluids and Barriers of the CNS*, 19(1):1–24.

Khaled, A.-R. and Vafai, K. (2003). The role of porous media in modeling flow and heat transfer in biological tissues. *International Journal of Heat and Mass Transfer*, 46(26):4989–5003.

Khan, A. and Zanotti, P. (2020). A nonsymmetric approach and a quasi-optimal and robust discretization for the Biot's model. Part I – Theoretical aspects. arXiv:2008.05307 [cs, math].

Koshiba, N., Ando, J., Chen, X., and Hisada, T. (2007). Multiphysics simulation of blood flow and LDL transport in a porohyperelastic arterial wall model. *Journal of biomechanical engineering*, 129(3):374–385.

Köppel, M., Martin, V., Jaffré, J., and Roberts, J. E. (2018). A Lagrange multiplier method for a discrete fracture model for flow in porous media. *Computational Geosciences*. Publisher: Springer Verlag.

Ladyženskaja, O. A., Solonnikov, V. A., and Ural'ceva, N. N. (1968). *Linear and quasi-linear equations of parabolic type*, volume 23. American Mathematical Soc.

Ladyzhenskaya, O. A. (1969). *The mathematical theory of viscous incompressible flow*, volume 2. Gordon and Breach New York.

Le Tallec, P. and Mani, S. (2000). Numerical analysis of a linearised fluid-structure interaction problem. *Numerische Mathematik*, 2(87):317–354.

Le Tallec, P. and Mouro, J. (2001). Fluid structure interaction with large structural displacements. *Computer methods in applied mechanics and engineering*, 190(24-25):3039–3067.

Lee, J. J. (2018). Robust three-field finite element methods for Biot's consolidation model in poroelasticity. *BIT Numerical Mathematics*, 58(2):347–372.

Lee, J. J., Piersanti, E., Mardal, K.-A., and Rognes, M. E. (2019). A mixed finite element method for nearly incompressible multiple-network poroelasticity. *SIAM Journal on Scientific Computing*, 41(2):A722–A747.

Levadoux, D. P. (2022). Analyse numérique de la formulation intégrodifférentielle d'un problème de Maxwell harmonique impliquant un diélectrique traversé de surfaces exfoliées métalliques et impédantes. Technical Report HAL.

Lewis, R. W. and Schrefler, B. A. (1987). *The finite element method in the deformation and consolidation of porous media*. John Wiley and Sons Inc., New York, NY.

Li, T. and Yotov, I. (2022). A mixed elasticity formulation for fluid–poroelastic structure interaction. *ESAIM: Mathematical Modelling and Numerical Analysis*, 56(1):1–40.

Li, X., Han, X., and Pastor, M. (2003). An iterative stabilized fractional step algorithm for finite element analysis in saturated soil dynamics. *Computer Methods in Applied Mechanics and Engineering*, 192(35):3845–3859.

Lions, J. L. and Magenes, E. (1972). *Non-homogeneous boundary value problems and applications*, volume 1. Springer Science & Business Media.

Liu, R., Wheeler, M. F., Dawson, C. N., and Dean, R. H. (2009). On a coupled discontinuous/continuous Galerkin framework and an adaptive penalty scheme for poroelasticity problems. *Computer Methods in Applied Mechanics and Engineering*, 198(41):3499–3510.

Logg, A., Mardal, K.-A., Wells, G. N., et al. (2012). *Automated Solution of Differential Equations by the Finite Element Method*. Springer.

Lopatnikov, S. L. and Cheng, A. H. D. (2004). Macroscopic Lagrangian formulation of poroelasticity with porosity dynamics. *Journal of the Mechanics and Physics of Solids*, 52(12):2801–2839.

Lourenco, W. d. J., Reis, R. F., Ruiz-Baier, R., Rocha, B. M., Dos Santos, R. W., and Lobosco, M. (2022). A poroelastic approach for modelling myocardial oedema in acute myocarditis. *Frontiers in Physiology*, page 1196.

Maria Denaro, F. (2003). On the application of the Helmholtz–Hodge decomposition in projection methods for incompressible flows with general boundary conditions. *International Journal for Numerical Methods in Fluids*, 43(1):43–69.

Markert, B. (2007). A constitutive approach to 3-d nonlinear fluid flow through finite deformable porous continua: With application to a high-porosity polyurethane foam. *Transport in Porous Media*, 70(3):427–450.

Markert, B., Heider, Y., and Ehlers, W. (2009). Comparison of monolithic and splitting solution schemes for dynamic porous media problems. *International Journal for Numerical Methods in Engineering*, pages n/a–n/a.

Massing, A., Larson, M., Logg, A., and Rognes, M. (2015). A nitsche-based cut finite element method for a fluid-structure interaction problem. *Communications in Applied Mathematics and Computational Science*, 10(2):97–120.

Michler, C., Cookson, A. N., Chabiniok, R., Hyde, E., Lee, J., Sinclair, M., Sochi, T., Goyal, A., Vigueras, G., Nordsletten, D. A., and Smith, N. P. (2013). A computationally efficient framework for the simulation of cardiac perfusion using a multi-compartment Darcy porous-media flow model. *International Journal for Numerical Methods in Biomedical Engineering*, 29(2):217–232.

Mikelić, A. (2000). *Homogenization Theory and Applications to Filtration through Porous Media*, volume 1734, pages 127–214. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mikelic, A. and Jäger, W. (2000). On The Interface Boundary Condition of Beavers, Joseph, and Saffman. *SIAM Journal on Applied Mathematics*, 60(4):1111–1127.

Mikelić, A. and Wheeler, M. F. (2013). Convergence of iterative coupling for coupled flow and geomechanics. *Computational Geosciences*, 17(3):455–461.

Mok, D. P. and Wall, W. (2001). Partitioned analysis schemes for the transient interaction of incompressible flows and nonlinear flexible structures. *Trends in computational structural mechanics*, 1.

Murad, M., Guerreiro, J. N., and Loula, A. (2001). Micromechanical computational modeling of secondary consolidation and hereditary creep in soils. *Computer Methods in Applied Mechanics and Engineering*, 190:1985–2016.

Murad, M. A. and Cushman, J. H. (1996). Multiscale flow and deformation in hydrophilic swelling porous media. *International Journal of Engineering Science*, 34(3):313–338.

Nash, M. P. and Hunter, P. J. (2000). Computational mechanics of the heart. *Journal of elasticity and the physical science of solids*, 61(1):113–141.

Nicaise, S. and Venel, J. (2011). A posteriori error estimates for a finite element approximation of transmission problems with sign changing coefficients. *J. Comput. Appl. Math.*, 235:4272–4282.

Nichols, M., Townsend, N., Scarborough, P., and Rayner, M. (2014). Cardiovascular disease in europe 2014: epidemiological update. *European heart journal*, 35(42):2950–2959.

Nordbotten, J. M. (2014). Cell-centered finite volume discretizations for deformable porous media. *International journal for numerical methods in engineering*, 100(6):399–418.

Nordbotten, J. M. (2016). Stable cell-centered finite volume discretization for Biot equations. *SIAM Journal on Numerical Analysis*, 54(2):942–968.

Nunes, H., Schubel, K., Piver, D., Magois, E., Feuillet, S., Uzunhan, Y., Carton, Z., Tazi, A., Levy, P., Brillet, P.-Y., et al. (2015). Nonspecific interstitial pneumonia: survival is influenced by the underlying cause. *European Respiratory Journal*, 45(3):746–755.

Owczarek, S. (2010). A Galerkin method for Biot consolidation model. *Mathematics and mechanics of solids*, 15(1):42–56.

Oyarzúa, R. and Ruiz-Baier, R. (2016). Locking-free finite element methods for poroelasticity. *SIAM Journal on Numerical Analysis*, 54(5):2951–2973.

O'Rourke, M. F. and Safar, M. E. (2005). Relationship between aortic stiffening and microvascular disease in brain and kidney: cause and logic of therapy. *Hypertension*, 46(1):200–204.

Patte, C. (2020). *Personalized pulmonary mechanics: modeling, estimation and application to pulmonary fibrosis*. PhD thesis, Institut polytechnique de Paris.

Patte, C., Genet, M., and Chapelle, D. (2022). A quasi-static poromechanical model of the lungs. *Biomechanics and Modeling in Mechanobiology*, 21(2):527–551.

Pazy, A. (2012). *Semigroups of linear operators and applications to partial differential equations*, volume 44. Springer Science & Business Media.

Phillips, P. J. and Wheeler, M. F. (2008). A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity. *Computational Geosciences*, 12(4):417–435.

Phillips, P. J. and Wheeler, M. F. (2009). Overcoming the problem of locking in linear elasticity and poroelasticity: An heuristic approach. *Computational Geosciences*, 13(1):5–12.

Pironneau, O. and Glowinski, R. (1979). On a mixed finite element approximation of the stokes problem (i). convergence of the approximate solutions. *Numerische Mathematik*, 33:397–424.

Polizzi, B., Bernard, O., and Ribot, M. (2017). A time-space model for the growth of microalgae biofilms for biofuel production. *Journal of Theoretical Biology*, 432:55–79.

Qohar, U. N. A., Zanna Munthe-Kaas, A., Nordbotten, J. M., and Hanson, E. A. (2021). A nonlinear multi-scale model for blood circulation in a realistic vascular system. *Royal Society Open Science*, 8(12):201949.

Quaini, A. and Quarteroni, A. (2007). A semi-implicit approach for fluid-structure interaction based on an algebraic fractional step method. *Mathematical models and methods in applied sciences*, 17(06):957–983.

Rajagopal, K. and Tao, L. (2005). On the propagation of waves through porous solids. *International Journal of Non-Linear Mechanics*, 40(2-3):373–380.

Rajagopal, K. R. (2007). On a hierarchy of approximate models for flows of incompressible fluids through porous solids. *Mathematical Models and Methods in Applied Sciences*, 17(02):215–252.

Rannacher, R. (1992). On chorin's projection method for the incompressible navier-stokes equations. In *The Navier-Stokes Equations II—Theory and Numerical Methods*, pages 167–183. Springer.

Rivière, B. and Yotov, I. (2005). Locally Conservative Coupling of Stokes and Darcy Flows. *SIAM Journal on Numerical Analysis*, 42(5):1959–1977.

Rodrigo, C., Gaspar, F. J., Hu, X., and Zikatanov, L. T. (2016). Stability and monotonicity for some discretizations of the Biot's consolidation model. *Computer Methods in Applied Mechanics and Engineering*, 298:183–204.

Rodrigo, C., Hu, X., Ohm, P., Adler, J. H., Gaspar, F. J., and Zikatanov, L. T. (2018). New stabilized discretizations for poroelasticity and the Stokes' equations. *Computer Methods in Applied Mechanics and Engineering*, 341:467–484.

Rohan, E., Turjanicová, J., and Lukeš, V. (2019). The Biot–Darcy–Brinkman model of flow in deformable double porous media; homogenization and numerical modelling. *Computers & Mathematics with Applications*, 78(9):3044–3066.

Ruiz-Baier, R., Taffetani, M., Westermeyer, H. D., and Yotov, I. (2022). The Biot–Stokes coupling using total pressure: Formulation, analysis and application to interfacial flow in the eye. *Computer Methods in Applied Mechanics and Engineering*, page 30.

Russell, T. F. and Wheeler, M. F. (1983). Finite element and finite difference methods for continuous flows in porous media. In *The Mathematics of Reservoir Simulation*, pages 35–106. SIAM.

Sacco, R., Causin, P., Lelli, C., and Raimondi, M. T. (2017). A poroelastic mixture model of mechanobiological processes in biomass growth: Theory and application to tissue engineering. *Meccanica*, 52(14):3273–3297.

Sacco, R., Guidoboni, G., and Mauri, A. G. (2019). *A Comprehensive Physically Based Approach to Modeling in Bioengineering and Life Sciences*. Academic press.

Saffman, P. G. (1971). On the boundary condition at the surface of a porous medium. *Studies in applied mathematics*, 50(2):93–101.

Saint-Macary, P. (2004). *Analyse mathématique de modèles de diffusion en milieu poreux élastique*. PhD thesis, Thesis Université de Pau et des Pays de l'Adour.

Santos, J. E. (1986). Elastic wave propagation in fluid-saturated porous media. Part I. The existence and uniqueness theorems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 20(1):113–128.

Sayas, F.-J., Brown, T. S., and Hassell, M. E. (2019). *Variational techniques for elliptic partial differential equations*. CRC Press.

Schanz, M. (2009). Poroelastodynamics: Linear Models, Analytical Solutions, and Numerical Methods. *Applied Mechanics Reviews*, 62(3):030803.

Schlömer, N. (2022). pygmsh: A python frontend for gmsh. https://doi.org/10.5281/zenodo.5886856.

Settari, A. and Mourits, F. (1998). A Coupled Reservoir and Geomechanical Simulation System. *SPE Journal*, 3(3):219–226.

Shen, J. (1995). On error estimates of the penalty method for unsteady navier–stokes equations. *SIAM Journal on Numerical Analysis*, 32(2):386–403.

Showalter, R. and Su, N. (2001). Partially saturated flow in a poroelastic medium. *Discrete and Continuous Dynamical Systems Series B*, 1(4):403–420.

Showalter, R. E. (2000). Diffusion in poro-elastic media. *Journal of mathematical analysis and applications*, 251(1):310–340.

Showalter, R. E. (2005). Poroelastic filtration coupled to Stokes flow. In *Control theory of partial differential equations*, pages 243–256. Chapman and Hall/CRC.

Showalter, R. E. (2013). *Monotone operators in Banach space and nonlinear partial differential equations*, volume 49. American Mathematical Soc.

Showalter, R. E. and Stefanelli, U. (2004). Diffusion in poro-plastic media. *Mathematical Methods in the Applied Sciences*, 27(18):2131–2151.

Singer, C. (1925). *The evolution of anatomy: a short history of anatomical and physiological discovery to Harvey: being the substance of the Fitzpatrick lectures delivered at the Royal college of physicians of London in the years 1923 and 1924*. AA Knopf.

Stone, J., Johnstone, D. M., Mitrofanis, J., and O'Rourke, M. (2015). The mechanical cause of age-related dementia (alzheimer's disease): the brain is destroyed by the pulse. *Journal of Alzheimer's Disease*, 44(2):355–373.

Storvik, E., Both, J. W., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2019). On the optimization of the fixed-stress splitting for Biot's equations. *International Journal for Numerical Methods in Engineering*, 120(2):179–194.

Støverud, K. H., Darcis, M., Helmig, R., and Hassanizadeh, S. M. (2011). Modeling concentration distribution and deformation during convection-enhanced drug delivery into brain tissue. *Transport in Porous Media*, 92(1):119–143.

Tavakoli, A. and Ferronato, M. (2013). On existence-uniqueness of the solution in a nonlinear Biot's model. *Appl. Math*, 7(1):333–341.

Temam, R. (1969). Sur l'approximation de la solution des équations de navier-stokes par la méthode des pas fractionnaires (i). *Archive for Rational Mechanics and Analysis*, 32:135–153.

Temam, R. (2001). *Navier-Stokes equations: theory and numerical analysis*, volume 343. American Mathematical Soc.

Terzaghi, K. (1943). *Theoretical soil mechanics*. Wiley, New York.

Terzaghi, K., Peck, R. B., and Mesri, G. (1996). Soil mechanics. *New York: John Wiley & Sons*.

Tully, B. and Ventikos, Y. (2011). Cerebral water transport using multiple-network poroelastic theory: application to normal pressure hydrocephalus. *Journal of Fluid Mechanics*, 667:188–215.

Van Duijn, C., Mikelić, A., and Wick, T. (2019). A monolithic phase-field model of a fluid-driven fracture in a nonlinear poroelastic medium. *Mathematics and Mechanics of Solids*, 24(5):1530–1555.

Van Kan, J. (1986). A second-order accurate pressure-correction scheme for viscous incompressible flow. *SIAM journal on scientific and statistical computing*, 7(3):870–891.

Vardakis, J. C., Chou, D., Tully, B. J., Hung, C. C., Lee, T. H., Tsui, P.-H., and Ventikos, Y. (2016). Investigating cerebral oedema using poroelasticity. *Medical Engineering & Physics*, 38(1):48–57.

Verri, M., Guidoboni, G., Bociu, L., and Sacco, R. (2018). The role of structural viscoelasticity in deformable porous media with incompressible constituents: Applications in biomechanics. *Mathematical Biosciences and Engineering*, 15(4):933–959.

Vuong, A. T., Yoshihara, L., and Wall, W. A. (2015). A general approach for modeling interacting flow through porous media under finite deformations. *Computer Methods in Applied Mechanics and Engineering*, 283:1240–1259.

Wall, W., Wiechert, L., Comerford, A., and Rausch, S. (2010). Towards a comprehensive computational model for the respiratory system. *International Journal for Numerical Methods in Biomedical Engineering*, 26(7):807–827.

Weber, C. (1980). A local compactness theorem for maxwell's equations. *Mathematical Methods in the Applied Sciences*, 2(1):12–25.

Wheeler, M., Xue, G., and Yotov, I. (2014). Coupling multipoint flux mixed finite element methods with continuous Galerkin methods for poroelasticity. *Comput Geosci*, page 19.

Wheeler, M. F. (1973). A priori l_2 error estimates for galerkin approximations to parabolic partial differential equations. *SIAM Journal on Numerical Analysis*, 10(4):723–759.

Wheeler, M. F. and Yotov, I. (2006). A multipoint flux mixed finite element method. *SIAM Journal on Numerical Analysis*, 44(5):2082–2106.

Wilmanski, K. (2005). Tortuosity and objective relative accelerations in the theory of porous materials. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 461(2057):1533–1561.

Wu, D. and Birukov, K. (2019). Endothelial cell mechano-metabolomic coupling to disease states in the lung microvasculature. *Frontiers in Bioengineering and Biotechnology*, 7:172.

Yang, M. and Taber, L. A. (1991). The possible role of poroelasticity in the apparent viscoelastic behavior of passive cardiac muscle. *Journal of biomechanics*, 24(7):587–597.

Yi, S.-Y. (2013). A coupling of nonconforming and mixed finite element methods for Biot's consolidation model. *Numerical Methods for Partial Differential Equations*, 29(5):1749–1777.

Yi, S.-Y. (2017). A study of two modes of locking in poroelasticity. *SIAM Journal on Numerical Analysis*, 55(4):1915–1936.

Zakerzadeh, R. and Zunino, P. (2019). A computational framework for fluid–porous structure interaction with large structural deformation. *Meccanica*, 54(1-2):101–121.

Ženíšek, A. (1984). The existence and uniqueness theorem in Biot's consolidation theory. *Aplikace matematiky*, 29(3):194–211.

Zhang, W., Zhao, Y., Zhang, F., Wang, Q., Li, T., Liu, Z., Wang, J., Qin, Y., Zhang, X., Yan, X., et al. (2020). The use of anti-inflammatory drugs in the treatment of people with severe coronavirus disease 2019 (covid-19): The perspectives of clinical immunologists from China. *Clinical immunology*, 214:108393.

Zienkiewicz, O., Huang, M., Wu, J., and Wu, S. (1993). A new algorithm for the coupled soil-pore fluid problem. *Shock and Vibration*, 1(1):3–14.

Zienkiewicz, O. C., Paul, D. K., and Chan, A. H. C. (1988). Unconditionally stable staggered solution procedure for soil-pore fluid interaction problems. *International Journal for Numerical Methods in Engineering*, 26(5):1039–1055.

Zienkiewicz, O. C. and Shiomi, T. (1984). Dynamic behaviour of saturated porous media; The generalized Biot formulation and its numerical solution. *International Journal for Numerical and Analytical Methods in Geomechanics*, 8(1):71–96.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Cadre mathématique pour la modélisation et la simulation de tissus biologiques perfusés

Mots clés : Poromécanique, Caractère bien posé, Analyse numérique, Limite incompressible, T-coercivité, Schéma à pas fractionnaire

Résumé : De nombreux tissus biologiques peuvent être modélisés comme des milieux poreux, c'est-à-dire des milieux continus composés d'une structure solide irriguée par un fluide. Dans les tissus biologiques, le fluide peut désigner le sang, les flux d'air dans les poumons ou encore le liquide céphalorachidien, fluides qui peuvent tous être considérés comme incompressibles. De plus, pour de telles applications, le milieu poreux en tant que tel est quasi-incompressible. L'objectif de cette thèse est d'analyser un modèle d'équations aux dérivées partielles récent qui décrit le mouvement d'un milieu poreux quasi-incompressible ou incompressible. Ce modèle provient de la linéarisation d'un modèle de poromécanique non linéaire adapté au contexte des tissus mous perfusés, mais il est également fortement relié aux équations de Biot en poroélasticité. Dans ce modèle, les équations du solide et du fluide ont un comportement respectivement hyperbolique et parabolique, et sont couplées par la pression interstitielle associée à la contrainte d'incompressibilité. La première contribution de cette thèse est de démontrer l'existence et l'unicité des solutions fortes ou faibles dans les cas quasi-incompressible et incompressible. La preuve repose sur une combinaison de théorie des semi-groupes, d'estimations d'énergie et fait appel à la notion de T-coercivité. Cette notion, développée originellement pour les problèmes non contraints, est ici étendue aux problèmes de type point-selle avec ou sans pénalisation. Le concept de T-coercivité s'avère également utile pour la conception d'éléments finis stables dans la limite incompressible et pour l'analyse numérique du système. La convergence spatiale et temporelle d'un schéma monolithique est prouvée, avec des estimations d'erreur robustes par rapport à l'incompressibilité, la porosité et la perméabilité. Afin d'accélérer le temps de calcul, un schéma à pas fractionnaires est proposé et analysé. En particulier, des conditions aux limites générales couplant le fluide et le solide sur le bord du domaine sont envisagées et imposées grâce à une méthode de type Robin-Robin. Enfin, la pertinence de ce modèle pour les applications biomédicales est illustrée en comparant des simulations de microvaisseaux sur puce à des données expérimentales.

Title : Mathematical framework for biological tissue perfusion modeling and simulation

Keywords : Poromechanics, Well-posedness, Numerical analysis, Incompressible limit, T-coercivity, Fractional-step method

Abstract : Many biological tissues can be modeled as porous media, namely continuous media composed of a solid skeleton filled by a fluid. In biological tissues, the fluid at stake can be blood, airflows in the lungs or cerebrospinal fluid, all of which can be seen as incompressible fluids. Moreover, in such applications, the porous medium itself can be considered as nearly-incompressible. The goal of this PhD thesis is to analyze a recent partial differential equation model describing the motion of a nearly-incompressible or incompressible porous medium. This model arises from the linearization of a non-linear poromechanics model adapted to soft tissue perfusion, but is also strongly connected to Biot's equations of poroelasticity. In this model, the solid and fluid equations show a hyperbolic – parabolic behavior, and are in addition coupled through the interstitial pressure associated with the incompressibility divergence constraint. The first contribution of this thesis is to show the existence and uniqueness of strong and weak solutions in the nearly-incompressible and incompressible cases. This is achieved by combining semigroup theory, energy estimates and T-coercivity. T-coercivity theory, originally developed for unconstrained problems, is extended here to treat general saddle-point and perturbed saddle-point problems. This concept also appears to be useful for the design of stable finite elements in the incompressible limit and for the numerical analysis of the system. Spatial and temporal convergence analysis are performed for a monolithic scheme, leading to robust error estimates with respect to incompressibility, porosity and permeability. In order to improve computational efficiency, a fractional-step method is proposed and analyzed. In particular, general boundary conditions connecting the fluid and the solid on the boundary are considered and imposed thanks to a Robin-Robin coupling method. Finally, the relevance of the model to biomedical applications is illustrated by comparing microvessels-on-chip simulations with experimental data.

Institut Polytechnique de Paris
91120 Palaiseau, France